Compiled Public Comments on the Request for Information on Measuring and Rewarding Scientific Impact

Guide Notice Number: NOT-OD-26-087

June 18, 2026 – August 19, 2026

Table of Contents:

1. N/A 11

2. Shu Joy Jiang 12

3. Camelia Kantor 14

4. Luigi Ferrucci 16

5. Clifford Saper 18

6. N/A 20

7. N/A 22

8. Anonymous 23

9. CC 25

10. N/A 27

11. Robert Tycko 28

12. Western Michigan University Homer Stryker MD School of Medicine 29

13. Alice Berger 31

14. Joel Pachter 32

15. N/A 33

16. Timothy Robinson 34

17. Colin Reardon 35

18. Francis O Walker, MD 36

19. Rajal G Cohen 37

20. Shiyue Fang 38

21. Dmitry terentyev 39

22. Daniel Herschlag 41

23. Paul O'Connor 42

24. N/A 43

25. Gary Cutter 45

26. yin chen 47

27. N/A 49

28. Oxy-17 LLC 50

29. Andrew Belmont 51

30. Iraida Sharina 54

31. N/A 56

32. N/A 58

33. Mark Chertoff 59
34. Herbert M Sauro 60
35. Olivier Le Saux 63
36. N/A 65
37. Gregory A. Pesko 66
38. David Huang 67
39. Jon A. Krosnick 68
40. Aaron James 70
41. Gordon W. Laurie 71
42. T John Winhusen 72
43. Desmond Heath 74
44. David Schopfer 75
45. Donald Lynam 76
46. Qinglin Yang 78
47. Malaz Boustani 80
48. Zhong Wang 84
49. Stefanie Vogel 87
50. N/A 89
51. E. Robert Burns, Ph.D., Professor Emeritus Retired, College of Medicine, Univ. Arkansas for Medical Sciences: burnsbob@uams.edu 90
52. ROBERT LAIRD JUDSON-TORRES 92
53. N/A 98
54. Edwin S. Monuki 100
55. N/A 102
56. Bycelium 103
57. Ji-Hyun Lee 104
58. N/A 106
59. Edwin S. Monuki 108
60. Benjamin Clarkson 109
61. N/A 110
62. N/A 112
63. Alexander Scheeline 114
64. Chris Batich 117

65. PWXPress 119
66. William Brett Perkison 120
67. Benjamin D. Singer, MD, ATSF 122
68. N/A 127
69. N/A 128
70. James Kochenderfer 130
71. Yousuke Takahama 131
72. Ana-Maria Vranceanu 132
73. Peter Gyarmati 133
74. N/A 134
75. N/A 135
76. N/A 137
77. N/A 139
78. N/A 141
79. N/A 142
80. Gregory Edward Cox 143
81. Katherine Sauder 145
82. Donald Allen Williamson, Ph.D. 146
83. fleming scientific 150
84. Annette Christine Frenk Oquendo 152
85. Rob Holston 153
86. Antony Schwartz 173
87. N/A 174
88. N/A 175
89. N/A 176
90. Manoja kumar mallick 177
91. N/A 178
92. Al George 179
93. N/A 181
94. Hafizullah Safi 184
95. JP Flores 192
96. Peter Deak 196
97. Chunyu Liu 198

98. N/A 199
99. Hossein Ravanbakhsh 200
100. David Rocha 201
101. Neehar Kundurti 202
102. Sean Whelan 207
103. N/A 208
104. N/A 209
105. Fabio Renato Manzolli Leite 210
106. N/A 214
107. Joseph Braun 216
108. Sandia biotech Inc. 217
109. N/A 220
110. Dennis Drayna 221
111. Frederick O. Cope, PhD MS MSc FACN 223
112. N/A 225
113. Jessica Paredes 226
114. N/A 231
115. N/A 233
116. Abdel Abdel-Rahman 236
117. Judith Grisel 237
118. Hoda Badr 238
119. N/A 242
120. hamid khosrowshahi 243
121. N/A 247
122. N/A 248
123. Aimee Payne, MD, PhD 249
124. N/A 251
125. Peter Karp 253
126. Jun Yang 255
127. N/A 257
128. Mikail Rubinov 258
129. Richard Kahn 259
130. Professor 260

131. Lawrence P. Reynolds, PhD 261
132. Danielle Tatum 263
133. Erin Mead-Morse 264
134. Jean Liew 266
135. Erick Kiprotich Yegon, PhD 270
136. N/A 274
137. N/A 277
138. Doug Brugge 278
139. N/A 279
140. Sara Kiesler 281
141. Imon Banerjee 282
142. Sindy KY Tang 286
143. mPower, Inc. 290
144. Lois Sayrs, PhD 300
145. N/A 304
146. Annette Greer 306
147. Alzheimer's Association 307
148. Jingyi Jessica Li 308
149. Essex Management LLC 311
150. Brian Bauer 343
151. N/A 345
152. Chaitanya Jain 347
153. Institute for Progress 349
154. David Piston 351
155. Cleveland Clinic 353
156. N/A 354
157. Ian Hagemann, MD, PhD 355
158. Suresh Palaniyandi 357
159. American Society of Human Genetics 360
160. Kirsten Davison 362
161. Michael Frumkin 368
162. Nigel Paneth 370
163. N/A 372

164. American Psychological Association 374
165. American Dental Association 378
166. N/A 382
167. Aimee Edinger 384
168. N/A 387
169. Good Science Project 389
170. Esmeralda Casas-Silva. Ph.D.; Kanakadurga Addepalli, Ph.D. 391
171. University of New Mexico Health Sciences 395
172. Association of Biomolecular Resource Facilities 401
173. N/A 407
174. Licklider, Inc. 408
175. Endocrine Society 416
176. Cambridge University Press and University of Cambridge Libraries and Archives 421
177. Daniel M Merfeld 422
178. NCI Office of Data Sharing and Trans-NCI DMS Working Group National Cancer Institute 427
179. Federation of American Societies for Experimental Biology (FASEB) 429
180. UNC Center for Health Promotion and Disease Prevention 439
181. Andrew Kant 441
182. Beth Israel Lahey Health 445
183. Weill Cornell Medicine 447
184. Make Data Count 448
185. Digital Science & Research Solutions Inc. 451
186. STM (The International Association 489
187. Medical Research Support 496
188. Xiaoxin Wang 548
189. N/A 550
190. Maryam Al-Mujtaba 553
191. The Massive Analysis and Quality Control (MAQC) Society 557
192. PLOS 560
193. Washington University in St. Louis 566
194. Society for Academic Emergency Medicine (SAEM) 574
195. Washington State University 577
196. Broad Institute 585

197. Emmanuel Okoro 586
198. ASM 587
199. American Medical Association 589
200. Yale University 590
201. Research Solutions, Inc. 591
202. Aleksandra Gruslova & Lindsey Macpherson 593
203. Bradley E. Alger, Ph.D. 596
204. AUTM - Association of University Technology Managers, Inc. 597
205. Kevin Jon Williams, M.D. 603
206. N/A 607
207. Griffin M Weber 609
208. Center for Open Science 610
209. Cherie Marvel 614
210. Open Research Funders Group 615
211. Elsevier 616
212. American Statistical Association 620
213. Americans for Medical Progress 632
214. Pengyin Shan 636
215. UsAgainstAlzheimer's 638
216. National Society for Histotechnology 639
217. Bethany Laursen 640
218. Tufts CTSI 642
219. Chronic Pain Research Alliance - an Initiative of The TMJ Association 644
220. Frontiers Media Inc. 650
221. Alexei Morozov 656
222. The 3Rs Collaborative 657
223. SPARC 663
224. Bayh-Dole Coalition 664
225. Association of Public and Land-grant Universities 665
226. Association of American Publishers 666
227. N/A 670
228. Biophysical Society 672
229. AAMC 674

230. United States Association for the Study of Pain 675
231. Coalition to Illuminate and Address Animal Methods Bias 685
232. People for the Ethical Treatment of Animals 686
233. American Chemical Society 691
234. Information Technology and Innovation Foundation 694
235. American Physiological Society 695
236. N/A 698
237. Zheng Wang 699
238. Federation of Associations in Behavioral and Brain Sciences 708
239. Springer Nature 711
240. The Alliance for Longevity Initiatives (A4LI) 721
241. Youyou Wu (1), Yang Yang (2), Brian Uzzi (3) 725
242. Future of Cancer Care Today 727
243. Massachusetts Biotechnology Council 731
244. Case Western Reserve University School of Medicine 733
245. American Association of Physicists in Medicine (AAPM) 740
246. Inter-university Consortium for Political and Social Research (ICPSR) 759
247. Engineering Biology Research Consortium 768
248. Michael E. Bales, Elizabeth A. Wood, Inmaculada de Melo-Martin, Matt Mikaelian, Kelly McConnell, Julianne Imperato-McGinley, Principal Investigator 772
249. Inter-university Consortium for Political and Social Research (ICPSR) 773
250. Micah Altman and Philip N Cohen 782
251. Wiley 783
252. University of California, San Francisco | Academic Senate Committee on Library and Scholarly Communication and University Library 789
253. James Weatherall 793
254. University of Illinois Urbana-Champaign 796
255. Datavant 800
256. Noam Y Harel 803
257. American Educational Research Association 806
258. Alyson Eggleston 811
259. Peace-Work (peace-work.org) 813
260. American Association of Colleges of Pharmacy 814
261. N/A 818

262. Stanford University 819
263. N/A 822
264. DataSeer 823
265. TK Kozai 825
266. University of North Carolina at Chapel Hill, Renaissance Computing Institute, HEAL Data Stewardship Group 828
267. University of Arizona 832
268. Gerontological Society of America 840
269. The Sijbrandij Foundation 841
270. American College of Rheumatology 844
271. American College of Rheumatology 845
272. American Association for the Advancement of Science 846
273. University of Pittsburgh 854
274. Association of American Universities 858
275. American Urological Association (AUA) 866
276. B. Ian Hutchins 867
277. Physicians Committee for Responsible Medicine 869
278. Woods Hole Center for Oceans and Human Health 870
279. Babak Heydari 871
280. American Medical Informatics Association (AMIA) 881
281. Greater Gulf Coast Translational Science Alliance 885
282. Fred Ledley, M.D. 892
283. Silent Spring Institute 896
284. Mark H Histed 899
285. NIAID Office of Data Science and Emerging Technologies 903
286. BioCrowd, Inc. 907
287. SciClin LLC 911
288. Angela Sy 912

1. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Other

Type of Organization - Other: National laboratory

Role: Institutional Review Oversight Committee Member

Date Submitted: 6/18/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: Overarching Recommendation
NIH should avoid reliance on any single metric (such as publication count, journal impact factor, or citation totals). Instead, a balanced scorecard approach combining scientific rigor, collaboration, impact, mentorship, openness, innovation, and research culture would better reflect the collaborative, rigorous, and mission-driven nature of modern biomedical science. Metrics should be normalized for career stage, discipline, project type, and institutional context to ensure fairness and minimize unintended incentives.

2. Shu Joy Jiang

I am responding: On behalf of myself

Name: Shu Joy Jiang

Name of Organization: Washginton university in st. louis

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/18/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: Current incentive structures reward publishing about tools more than building and sharing them. A researcher who builds validated, deployable software used by 50 institutions has more real-world impact than one who publishes 10 papers about methods nobody implements.
Specific indicators:

-Number of institutions actively using shared tools
-Download and deployment metrics
-Validation studies by independent groups
-Infrastructure built on shared foundations

3. Training and Mentorship: Students who go on to build companies, get regulatory clearances, or lead policy changes are invisible in current mentorship metrics. We count publications from trainees. We should count:

-Startups founded by trainees
-Regulatory submissions by trainees
-Policy positions held by trainees
-Industry roles that translate research

4. Collaboration: Multi-site federated research is completely invisible in current metrics. Leading a 10-site consortium is infinitely harder than single-site research — yet authorship conventions and grant metrics don't capture this.
Specific indicators:

-Number of institutions in collaborative networks
-Cross-disciplinary team composition
-Data sharing agreements established
-Infrastructure coordination role vs. passive participation

5. Entrepreneurship and Translation: Current metrics completely miss translational impact. FDA regulatory clearance should be a measurable indicator equivalent to a high-impact publication. A Breakthrough Device Designation represents peer review by the most rigorous regulatory body in the world — yet it appears nowhere in academic promotion criteria. Similarly, industry acquisition of research tools, licensing agreements, and startup formation from academic research should be formally tracked and weighted in grant renewal and promotion decisions.Specific measurable indicators:

Regulatory submissions and clearances (FDA, Health Canada, EMA)
-Industry partnerships with dollar value
-Startup formation and funding raised
-Technology licensing agreements
-Time from research to clinical deployment

6. Foundational Scientific Exploration: N/A

7. Public Impact: Media coverage, policy citations, and clinical guideline inclusion are real measures of public impact that current NIH metrics ignore entirely. A research finding covered by NYT reaching millions is more impactful than a paper cited 50 times within academia.
Specific indicators:

-Media coverage reach
-Clinical guideline citations
-Policy document citations
-Patient advocacy engagement
-Health economic impact estimates

8. Other Areas: N/A

3. Camelia Kantor

I am responding: On behalf of myself

Name: Camelia Kantor

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/18/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: With the future of science being increasingly AI-assisted, the differentiator will likely not be technical expertise gained but mastery of human specific capabilities.
- Human skills development index to be created as a strong measure of success (e.g., a lab producing a technically excellent but dysfunctional scientist from a society engagement, emotional intelligence, resilience, acceptance of critique, and empathy perspective).

Network creation index - trainee-initiated/led engagements (collaborations, cross-disciplines, institutions, etc.)

Team science readiness score (Do trainees manage smaller projects with interdisciplinary teams? Do they communicate across disciplines?)

4. Collaboration: Team science contribution to receive high credit. More specifically, recognition of the "translators" between disciplines, sectors, institutions, methodologies who bring significant value but tend to be less cited or only in support roles, not senior PI/co-PIs roles. Those who make "convergence" possible. Examples include: research development professionals, synthesis leaders, center/institute directors, strategic initiatives leaders, etc.

Rewards for institutions and projects that integrate multiple disciplines, encouraging diversity of expertise (previously disconnected but complementary disciplines) and novel combination of methods, cross-field citations.

Recognition and rewards for investigators who create or leverage shared resources.

Evaluation of collective intelligence.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Reward intellectual novelty - ideas that have not/have rarely been connected before.

Institution-level evaluation of risk-taking portfolio (fund more those who have some high risk/high impact transformative research in their portfolio)

Projects that present lessons learned from failures that help refine theories, address certain gaps, and disprove hypotheses (well-documented failure that advances understanding, not any kind of failure).

7. Public Impact: Reward visible changes in nearby communities (geography matters) before solving problems elsewhere (e.g., actual local schools' nutrition not just public health recommendations).

8. Other Areas: N/A

4. Luigi Ferrucci

I am responding: On behalf of myself

Name: Luigi Ferrucci

Name of Organization: IRP/NIA

Type of Organization: Not Applicable

Role: Government Official

Date Submitted: 6/18/2026

1. Rigor and Reproducibility: The reliability of scientific findings is as important as their novelty. Indicators of rigor might include:

Data and code sharing.
Development and dissemination of protocols and methods.
Replication and validation studies.
Transparent reporting practices.
Correction of errors when identified.

2. Data, Software, and Model Sharing: Some of the most influential scientific contributions involve the creation of resources that support decades of discovery. Examples include longitudinal cohorts, biobanks, databases, software platforms, and shared methodological tools.

Potential indicators include:

Utilization of shared resources by the scientific community.
Scientific outputs enabled by those resources.
Long-term sustainability and accessibility.

This is particularly important in fields where scientific progress depends on sustained investments over many years. There should be way or rewarding those who produce data that are then used extensively by the scientific community. These key investigators have often difficulty to generate may paper because they focus on data generation.

3. Training and Mentorship: Possible indicators include:

Trainee development and career progression.
Evidence of effective mentoring practices.
Contributions to training programs.
Development of diverse and interdisciplinary scientific talent.

Assessment should emphasize quality and effectiveness rather than simple counts of trainees.

4. Collaboration: Possible indicators include:

Trainee development and career progression.
Evidence of effective mentoring practices.
Contributions to training programs.
Development of diverse and interdisciplinary scientific talent.

Assessment should emphasize quality and effectiveness rather than simple counts of trainees.

5. Entrepreneurship and Translation: Examples include:

Contributions to clinical guidelines.
Influence on public health policy.
Development of diagnostics, therapeutics, or technologies.
Demonstrated improvements in health outcomes.

6. Foundational Scientific Exploration: This aspect of science is critical for the NIH mission. Basic science medical and biophysics discovery produce the found for translation science that has the potentials for translational applications. There two problem that this unique to NIH: 1. Is expensive and there is high risk or failure. 2. Even when successful, the translation to practical applications may occur may years (decades sometime) later. For this reason, this type of research is not done by the private sector.

The practical assessment of what will be important in the future and what will not is complex and requires personal judgment, and should be assessed over time.

7. Public Impact: Very difficult to measure, and requires long time.
Some possible suggestions:

Adoption of findings into clinical guidelines
Demonstrated improvements in patient outcomes
Contributions to public health programs
Development of widely used datasets, methods, or tools
Regulatory approvals influenced by the research
Evidence of reducing disparities or improving access
Educational and public engagement activities with measurable reach

8. Other Areas: The objective should not be to replace traditional scholarly metrics but to develop a more comprehensive framework that evaluates both the generation of scientific knowledge and the contributions that enable scientific progress. Such a framework should recognize discovery, rigor, collaboration, mentorship, infrastructure development, and translation of knowledge into practice.

5. Clifford Saper

I am responding: On behalf of myself

Name: Clifford Saper

Name of Organization: Beth Israel Deaconess Medical Center/ Harvard Medical School

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/18/2026

1. Rigor and Reproducibility: While rigor in scientific design is important, rigidity of thinking about that design can still lead to erroneous results. The beauty of science is that it is self-cleaning. Even rigorous work that is barking up the wrong tree is eventually corrected, when new ideas and new approaches become available. The true test of the reproducibility and value of the work is best reflected in its long term citation. Many newsy papers are a flash in the pan, while work that is still cited 5, 10 or 20 years later after it has been tested, re-tested, and challenged, has the greatest value. So, I would propose that the 5 year citation rate be a standard for estimating the value of work across time.

2. Data, Software, and Model Sharing: NIH says it values sharing, but does little or nothing to foster it. Perhaps the NIH should have a program that would support the time and effort it takes to share resources: to put them in a form that makes them available, provide a forum for a standard way to communicate that, and help pay for the costs of distribution.

3. Training and Mentorship: NIH takes way too long to support training. Students and postdocs can have to submit, resubmit, and sometimes submit again to obtain support. Their graduate school or postdoc research time is generally 3 or 4 years, so a support grant that takes 3 cycles (read: three years!) to achieve is almost worthless. There should be a special short form for application and short cycle for funding for training.

4. Collaboration: Perhaps this could be recognized by the products, not the grant proposals (which are easily gamed). Look at the publications by investigators, and their collaborations, to assess the extent to which this is happening, and perhaps reward for that (for example, extend the payline by 5 percentile points for projects with investigators who have a history of collaboration).

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Basic research is alive and well in academia, but it is under attack from the current NIH priority statements. This is particularly true for organismal problems, such as brain function (which frequently requires an intact brain for example to measure cognition, anxiety, attention, wake-sleep, etc). The recent NIH priority statements advertise the priority of almost everything else above this critical work, which is the only to take air-fairy concepts from reduced preparations (cell lines, organoids, organs on a chip), which lack validation for whole animals, and test them for their applicability to the real world.

7. Public Impact: Look at new diagnostics and treatments, and follow the idea back in time to its basic science roots. Society has to be educated that the origin of almost every single advance in medicine and biotech in the last 50 years has been due to NIH support for basic science.

8. Other Areas: N/A

6. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/18/2026

1. Rigor and Reproducibility: While rigor and reproducibility are essential pillars of scientific progress, the current peer-review system has significant limitations in ensuring research integrity. Reviewers are generally unpaid volunteers who must evaluate a large volume of manuscripts and grant proposals under substantial time constraints. As a result, reviews often focus on scientific novelty and technical merit rather than a thorough assessment of data integrity, reproducibility, or potential misconduct.

Furthermore, allegations of research misconduct are frequently difficult to investigate and may be dismissed as errors, oversight, or poor research practices rather than being subjected to rigorous scrutiny. This can allow problematic findings to remain in the scientific literature and continue influencing future research and funding decisions.

To strengthen rigor and reproducibility, additional review mechanisms should be established that specifically focus on research integrity. Dedicated reviewers or independent integrity panels could evaluate data transparency, methodological rigor, statistical validity, image and data manipulation, and reproducibility considerations for both grant applications and manuscripts. Creating incentives, infrastructure, and accountability mechanisms for replication studies would further improve the reliability of the scientific record and enhance public trust in biomedical research.

2. Data, Software, and Model Sharing: While data, software, and model sharing are important for transparency and reproducibility, sharing alone does not ensure research quality or integrity. Researchers can share datasets that contain errors, biases, or even fabricated results. The larger challenge is creating incentives that reward rigorous, reproducible, and trustworthy science rather than primarily rewarding publication counts, novelty, citation metrics, or funding success. Policies that recognize replication studies, data validation efforts, negative results, and long-term reproducibility may have a greater impact on scientific reliability than data-sharing requirements

3. Training and Mentorship: NIH has invested substantially in training programs and protected research time, but additional measures are needed to evaluate the long-term effectiveness of these investments. Current metrics often emphasize grant acquisition, publication counts, and career advancement rather than the quality, rigor, reproducibility, and scientific impact of the resulting research. In some cases, investigators with established funding records continue to receive substantial support despite limited assessment of whether previous awards produced robust, reproducible, and influential findings. NIH should consider incorporating stronger outcome-based evaluations of prior funding, including research

quality, data transparency, reproducibility, mentorship outcomes, and long-term scientific impact. Such measures would help ensure that funding decisions reward excellence and stewardship rather than primarily prior funding success

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: NIH should expand support for early-career investigators and high-risk, high-reward research through dedicated funding mechanisms that are less influenced by traditional peer-review biases. While NIH recognizes the importance of bold and innovative science, the current review system often favors established investigators, well-known institutions, and projects with extensive preliminary data, creating barriers for transformative ideas and researchers from smaller universities.

Many high-risk proposals are judged by panels that are inherently risk-averse, despite programs being specifically intended to support innovative research. In addition, investigators at smaller or less research-intensive institutions may face disadvantages due to institutional reputation, available resources, or reviewer perceptions rather than the scientific merit of the proposal itself.

NIH should explore review models that reduce institutional and investigator bias, including greater use of partially anonymized review processes, dedicated panels trained to evaluate high-risk research, and funding mechanisms that place greater emphasis on innovation and potential impact rather than prior funding history. Expanding opportunities for early-career investigators and researchers outside major research-intensive institutions would broaden participation, diversify scientific perspectives, and increase the likelihood of paradigm-shifting discoveries.

7. Public Impact: N/A

8. Other Areas: N/A

7. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: In the clinical space, make training awards more fairly applicable across types of scientists so that trainee success can be more clearly understood. For example, NIH trainee award salary caps are too low to matter for MD applicants but the new rule where trainees must be funded at their current pay rate disincentivizes PhD postdocs (who always make under the pay cap) from applying, and thus it may appear MD trainees are applying for, and being awarded more early career awards. New rules around pay rate only affect those scientists who make under the NIH cap- who are also the scientists more like to be PhDs, to come from smaller institutions, or come from states not as heavily funded by NIH and thus understanding impact is skewed. Additionally, delays from the NIH in making awards or erroneously canceling awards should be factored into trainee success rates and other measures of impact, as these months of delay significantly impact trainee productivity and career progression.

4. Collaboration: N/A

5. Entrepreneurship and Translation: Take into account when the NIH delays awards or shuts down programs as it quantifies the successful grant cycles, such as the STTR and SBIR programs

6. Foundational Scientific Exploration: N/A

7. Public Impact: NIH should make their award process fully transparent (e.g., what exactly happens between council and funding and who is involved? Who exactly decides which proposals that are "competitive but not discussed" move forward and which do not, and by what exact criteria are proposals placed in this new category? This should be 100% public information!) and take political appointees completely out of award decisions (e.g., do not implement the current proposed executive order to allow non-scientists to make scientific funding decisions). Only then can the public and the scientific community actually trust that the awards being funded are truly based on merit.

8. Other Areas: N/A

8. Anonymous

I am responding: On behalf of myself

Name: Anonymous

Name of Organization: Tier 1 University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/20/2026

1. Rigor and Reproducibility: More then replicating randomly identified efforts, the emphasis should be on practices that support rigor in the first place. Incentives should be present for solid design - with sufficient power, sufficient control conditions, and milestones to decision making. NIH currently doesn't offer any incentive for reproducibility of findings (the emphasis is on innovation). Providing RFAs that support replication perhaps would incentivize replication efforts.

2. Data, Software, and Model Sharing: These activities require resources. The primary obstacle to utilizing or contributing to such efforts is lack of resources. As such if NIH is interested in incentivizing such activity it needs to offer financial supplements that make the hiring of staff with skillsets to support data/software/model sharing. It is impossible for each lab to maintain such expertise, which then leads to inequality whereby large centers are the only ones who can do this work. Alternatively NIH can create a team that will provide that service to researchers, or reduce the barriers (usually of high overhead burden).

3. Training and Mentorship: The emphasis should not be on numbers (of trainees or workshop attendees) but on products. What do these trainees produce? Do they produce papers, do they produce a collaboration, do they produce a new diagnostic effort.

4. Collaboration: Unfortunately the recent proposal to cap #of RPGs per investigator (https://rfi.grants.nih.gov/?s=6a1f200c6e0a7544e30b3f12) and also the OMB provision that limits international collaborations (§200.220, https://www.federalregister.gov/documents/2026/05/29/2026-10817/regulation-for-federal-financial-assistance) is in direct opposition to collaboration. So perhaps a first step would be to oppose these efforts. Team based science requires both multiple RPGs per investigator and international investigators.

Rewarding MPI efforts would be the right way forward but this is in direct conflict with the proposals above. Either we do team science and hold multiple grants (so that we can work together and combine expertise) or we work in isolation so that more people can get RPGs. This conflict is deeply problematic and prohibits a recommendation to this question.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: The R21 mechanism is supposed to support high-risk initiatives but it is a total of $250,000 over 2 years. This covers about the salary+benefits of 1 individual (staff

member) for 2 years. If the desire is to support high-risk initiatives then there should be RFAs with significantly more funding dedicated to them!

7. Public Impact: Then this should be built into RFAs as a required product and outcome.

8. Other Areas: N/A

9. CC

I am responding: On behalf of myself

Name: CC

Name of Organization: Prefer not to Share

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 6/21/2026

1. Rigor and Reproducibility: Staff scientists and postdoctoral fellows are mature scientists who play a seminal role in advancing science in many NIH funded labs. They are well suited to independently report reproducibility as they are funded by non profit agencies and taxpayers via grants, not by the PI or university. Leveraging that existing wealth of granular expertise without additional cost to the PI or the university, while serving the greater cause of publicly funded science, could be practically implemented at virtually no cost via a web portal similar to FDAs famous side effect reporting site. AI agents could summarize or pinpoint recurring areas of concern to a human examiner. Inadvertently, such a forward looking practice also frees these talented colleagues from the fear of negative consequences, similar to culture surveys granting agencies have put in place to deter or identify habitual harassers. Most scientists are scrupulously honest and hopefully would support their postdocs and staff scientists ability to report.

2. Data, Software, and Model Sharing: Reviewers could be asked to rate an additional component on how impactful a method or tool or big data paper has been for those kinds of grants by assessing how widely the tool or data is being used. This metric feeds also links the quality of big data and its reproducibility to general applicability for large grants. This has worked quite well for the Human Genome project and epigenetics consortia.

3. Training and Mentorship: Definitely need a hotline to report harassers that is independent of the Professors chain of command and the university or institute training office--as internal processes such as these are loth to investigate and have vested interest in the outcomes. A preferable method is an independent agency or appointed council who can neutrally evaluate claims made in accordance with laws or policies. Currently trainees are advised to move on and not shake the tree --allowing the very small fraction of serial harassers and fraudsters to keep climbing the tree. For taxpayer funded work, integrity should extend to every component of the product and process.

4. Collaboration: This is much harder as some collaborations are truly impactful while some are merely form letters to get important names on a paper.
Something like the Manhattan project for a given intractable disease is sorely needed to focus efforts towards specific 5-10 year needs with measurable outcomes, preferably under the leadership of visionary broad minded scientist with a reputation for excellence in science and global impact.

5. Entrepreneurship and Translation: Same principle as above would apply

6. Foundational Scientific Exploration: You need leadership that can identify bold and creative scientists vs. those who are merely sloppy scientists but excellent snake oil salespeople. Non trivial endeavor requiring astute leadership that can parse that difference.

7. Public Impact: Having patient representatives on Grant panels can help. Non profits already do this.

8. Other Areas: N/A

10. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/22/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: Add one section in the biosketch to allow one to document data and software sharing by the group and include the amount of shared data and the usage of the shared data/tool by the community, such as the size of the dataset and citation of the shared tool and dataset. Include the outcome of data sharing in grant application, particularly for competing renew.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

11. Robert Tycko

I am responding: On behalf of myself

Name: Robert Tycko

Name of Organization:

Type of Organization: Government Agency

Role: Investigator/Researcher

Date Submitted: 6/22/2026

1. Rigor and Reproducibility: Scientific findings that have real impact within a given research field are always examined critically and tested by other scientists within the relevant field. This process may not happen instantaneously, but it happens inevitably. To make it happen more quickly, the NIH could consider funding projects with explicit goals of testing the reproducibility of important results from other labs. The NIH should also encourage investigators to include full details of their experimental protocols, calculations, and data analyses in all publications. Review panels could be asked to examine whether these details have been included.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: Entrepreneurship by scientists is already strongly encouraged by universities, the venture capital community, and state governments. The desire for personal wealth is also a strong motivation for some individual scientists. Additional incentives from NIH seem unnecessary.

6. Foundational Scientific Exploration: There is definitely too much emphasis on publications in "high-impact" journals. The NIH should explicitly discourage review panels and program officers from using the "impact factor" of journals or other simple-minded metrics in their decisions. Instead, decisions should be based on careful reading of publications, proposals, and other information by true experts. This is a laborious process, but it is necessary.

7. Public Impact: N/A

8. Other Areas: N/A

12. Western Michigan University Homer Stryker MD School of Medicine

I am responding: On behalf of an organization

Name: Robert G. Sawyer

Name of Organization: Western Michigan University Homer Stryker MD School of Medicine

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Foundational/basic research: Issue a strong message to study sections that applications that seek funding to reproduce previously generated data (even on a smaller scale) either by the original investigator or others as the first step to an extension of research is not just acceptable but desired.

Clinical research: Actively fund research that reproduces clinical trial results either in a key subgroup or uses the experimental arm intervention as the control in a subsequent trial.

Health services research: Require all funded studies in large health services data sets to perform the same analysis in a similar but separate data set, either domestic or international.

2. Data, Software, and Model Sharing: Foundational/basic research: Construct a central, national repository for data generated by NIH-funded research (and perhaps non-NIH-funded researchers) that is accessible to other legitimate researchers. This repository should be searchable and include detailed methods and results.

Clinical research: Work immediately to define standards to make patient data acceptably anonymous to encourage clinical data set sharing between institutions. A central repository would be ideal, but this first step will allow easier sharing without resorting to one-by-one data use agreements.

Health services research: Fund projects that seek to devel tools that allow for easier bridging of health services data sets, for example, between COSMOS and HCUP/NIS.

3. Training and Mentorship: No matter how you cut it, the most important indicators need to be intermediate to long term (5+ years) after training stops. Measuring output (publications, etc.) during training is slightly helpful, but very dependent on the mentor. The best measure would be likelihood of trainees from a specific program staying in academics at the 5 and 10-year post-training mark. To have a trainee be very productive during training but not pursue an academic career is a fairly wasteful result, since an alternate trainee that was less productive during training but had a prolific academic career afterwards would be a far better outcome.

4. Collaboration: There needs to be some recognition in funding mechanisms that PIs in multi-PI efforts all contribute a great deal, and not that their salary support is just a division of the traditional 20% salary

support for a PI. For example, if an award historically might have a single PI at 20% effort, it does not mean that research in the same area with four PIs as part of a multi-PI project should result in 5% effort for each PI. One way to work toward this is to have a smaller number of awards but increase support for legitimate work done by each PI in a multi-PI grant.

5. Entrepreneurship and Translation: Take away the required business partnership in SBIR/STTR applications IMMEDIATELY. There are thousands of excellent ideas in the device space that could easily be prototyped in an academic/university setting with the goal being to sell the IP to a commercial entity. Currently, the hardest part in device development is not developing the device, but having to establish a company, find a company as a partner, or spin a company out of a university before applying for NIH or NSF funding!

6. Foundational Scientific Exploration: Be very, very careful when constructing review panels for funding applications that are meant to lead to new "bold and innovative research directions." If the reviewers all represent mainline, historically successful institutions, the panel will fund mainline, historically accepted ideas. Representation from non-traditional and newer organizations would help a great deal.

7. Public Impact: Fund more prospective, multicenter trials with reasonable reimbursement. The New England Journal of Medicine, Lancet, and JAMA are filled with high-quality trials from Canada, Europe, Australia, and Asia and very few from the US and even fewer NIH-funded trials. This is especially true for studies that do not include novel therapeutic agents. Trial data from North America are, far and away, the most influential in terms of clinical practice and public trust.

8. Other Areas: N/A

13. Alice Berger

I am responding: On behalf of myself

Name: Alice Berger

Name of Organization: Fred Hutchinson Cancer Center

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: In my view, there is a natural selector for rigor and reproducibility in our scientific system, and explicit reproducibility studies are not needed. For truly impactful research (think of CRISPR), the field quickly moves to replicate and extend the finding. The citation index of the individual paper that reported the advance therefore provides a reasonable indicator over time of how impactful and reproducible the work was. Alternative methods for quantifying the degree to which the finding was externally validated or extended could be considered, but I think citation index is reasonable.

2. Data, Software, and Model Sharing: the pros of data, software, and model sharing is that we can ensure usefulness of these products after their production. The con is that this takes time and money. If these data sources, software, or models, are not broadly useful, then the resources needed to share these products are not worth the cost. I think a great success is the GEO database, which makes it reasonably easy for investigators to share data. Continued investment in GEO is recommended.

3. Training and Mentorship: Training and mentorship is very important, but there seems to be a great emphasis on "training" in biomedical science than in other career paths such as tech. Sometimes I wonder why we treat 35 year old post-docs as "trainees" rather than full individual scientists. I think this is worth discussing. At what point does training end and doing science begin?

4. Collaboration: A collaboration metric could be developed to quantify how much of a person's research portfolio is collaborative. This would allow tracking and analysis to understand how collaboration impacts scientific productivity.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

14. Joel Pachter

I am responding: On behalf of myself

Name: Joel Pachter

Name of Organization: UConn Health

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: The best indicator of rigor and reproducibility in research, is when other laboratories independently confirm your findings and successfully apply your methodologies to their programs. Publication in the highest impact journals doesn't necessarily convey rigor or assure reproducibility. In fact, too often, high impact journals - in their rush to publish articles they deem will serve as the highest click bait - are obliged to offer retractions. Research that stands the test of time, and significantly moves one or more fields forward, is the best reflection of rigor and reproducibility.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

15. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: I actually think the problem is you are often funding the same intervention over and over that has no real-world value. Focus on real world value. Who cares if it is reproducible if it has no real world value.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: Right now it feels like it is K award or bust for an ESI. You also give most of your ESI grants to the same people, doing

4. Collaboration: Have team science based grants. That will solve your problem. Right now, you make Multiple PI proposals more burdensome, instead of facilitating or rewarding them.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: You need to fund more innovative digital health (extended reality, AI, etc.) That is the future. Most of what you fund is too dated. Stop funding dated research. That will require getting new people in the study section who value real innovation.

7. Public Impact: N/A

8. Other Areas: Stop funding the same JTAI and text messaging studies. Stop just funding elite researchers who have no real connection to real-world problems or needs. Stop funding useless DEI research. Start funding stuff with more real world application. Your study sections are too biased for certain research approaches that carry very little real-world impact. Start funding for the future: extended reality, AI, digital health. Most people outside of academia would be shocked by the silly stuff that NIH funds and confused by the more practical things it doesn't fund. Also, stop funding research that is just about giving people free stuff. That is not practical in the real world, yet you keep on funding it.

16. Timothy Robinson

I am responding: On behalf of myself

Name: Timothy Robinson

Name of Organization: Yale University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: early phase studies in humans (phase 0/1) could provide less artifactual study findings that arise due to nuances of in vitro artificial systems.

2. Data, Software, and Model Sharing: Novel methods should be judged on whether other people actually use them.

3. Training and Mentorship: There is little to no reward these days for being a good mentor or being a good teacher - in fact quite the opposite. Training grants that provide some mentor support would be a great idea.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: In oncology, there are too many preclinical studies being performed with models that do not reflect human physiology and immune systems. These can be helpful, but there needs to be a dramatic shift towards emphasizing early phase 1 / phase 0 trials that are directly exploring actual human physiology and cancer. This can be done through non-invasive, low-risk imaging and other correlative approaches. These have tremendous potential to directly learn about REAL human cancers. Some animal models have been studied for decades and then when they are attempted to be translated into humans - they don't work. Preclinical science is critically important, but early phase 0/1 trials are EQUALLY critical for understanding tumor biology - not just for trying a new drug.

8. Other Areas: N/A

17. Colin Reardon

I am responding: On behalf of myself

Name: Colin Reardon

Name of Organization:

Type of Organization: Not Applicable

Role: Member of the Public

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Although I am an academic scientist at a public university, I write today as a concerned private citizen. Rigor and Reproducibility has always been part of science in this country. It has fueled incredible growth, prosperity and provided new advances in health care. These continued efforts and a "gold standard science" agenda seem to suggest that we haven't followed a mantra of ensuring our results are correct, and that we in the scientific community do not seem to care if our conclusions are correct. This would seem to be opposite to the stated goals of reassuring the public. There is indeed a disconnect in our country when those who deny the value of vaccines, or suggest that new technologies that have evolved over 20 years of careful peer reviewed science are not safe, but that colloidal silver or Ivermectin will cure many conditions. This administration and the leadership at HHS is destroying our standing in the world, and handing over our future to geopolitical rivals.

2. Data, Software, and Model Sharing: Increasing the amount of meta data as part of data sets would be welcome. too often one finds samples labeled in a simple manner that do not allow reconstruction of analysis. The efforts to data have been good though. improvements could look to simplified viewing of data such as in cellxgene for scseq to more rapidly determine if the data set is useful.

3. Training and Mentorship: N/A

4. Collaboration: The limitations on the number of applications per cycle have curtailed our ability to collaborate. the administration seems intent to cherry pick some small number of investigators and punish the rest of the community. You are having the exact opposite effect of your stated goal.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Again the limits on submission for these applications means that I will no longer apply for R21 or other types of these applications. you are incentivizing the community to submit only safe applcaiton for standard competitions.

7. Public Impact: It is truly amazing how this country can lead the work in number of therapies developed, operation warp speed, and to see this question asked. if you think the public distrusts experts, the administration could start by not demonizing public servants that dedicate their lives to the American people.

8. Other Areas: N/A

18. Francis O Walker, MD

I am responding: On behalf of myself

Name: Francis O Walker, MD

Name of Organization: Wake Forest School of Medicine

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: 1) Reproducibility: At the time of funding, studies should be divided into those that can be replicated cost-effectively and those in which rigor can be cost-effectively assessed. For example, many biomarker studies can be easily replicated, as can drug trials in animal models. A replication RFP should be put out for a percentage of such studies, which would simultaneously fund another lab to replicate the study, ideally before the results of the initial study is funded; obviously confidentiality and shared resources would be required. The knowledge that a study is likely to be undergoing a replication effort would help ensure rigor as well as accuracy, and the process would be instructive for reviewers.
2) Rigor: Using the above criteria, put out an RFP for key aspects of a study amenable to testing of rigor. A biomarker could be tested for reliability, and/or metabolites from an animals blood stream or urine might be able to validate drug/placebo dosing. Studies producing a new antibody, compound, or proposing a new technique might be amenable to having one or two aspects of the study validated in a second laboratory. Again, confidentiality and shared resources would need to be required. The knowledge that some aspects of a study's rigor were undergoing validation would help ensure reproducibility of findings.

2. Data, Software, and Model Sharing: no suggestions

3. Training and Mentorship: no comment

4. Collaboration: no comment

5. Entrepreneurship and Translation: no comment

6. Foundational Scientific Exploration: no comment

7. Public Impact: This needs some exploratory work. If one looks at citations, it reflects what academicians are reading and using. One needs to survey practicing physicians to see what new technology has significantly changed their practice and the care of their patients. Experience designing and interpreting such surveys will take a few years, but in the long run will probably be the most informative way to sort this out.

8. Other Areas: N/A

19. Rajal G Cohen

I am responding: On behalf of myself

Name: Rajal G Cohen

Name of Organization: University of Idaho

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: There should be grants just for exact replications of other researchers' work. The ceaseless emphasis on novelty works against this being a priority. Money spent here might seem to slow down innovation initially but could pay off in the long run.

2. Data, Software, and Model Sharing: Could there be some concrete benefit to a researcher whose data is used (by a non-affiliated researcher) from a repository? Maybe free publication in an open access journal, or a boost in their score for their next NIH submission?

3. Training and Mentorship: Training grants could require some percentage of studies by mentees be replications of work from other labs.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

20. Shiyue Fang

I am responding: On behalf of myself

Name: Shiyue Fang

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: Is it possible to use AI to do these evaluations? In principle, if human evaluation data are generated for a significant number of cases (need to ensure the data is fair and no biase), they can be used to train AI to do the work. It should be better than relying solely on pub #, citation #, and grant amount.

21. Dmitry terentyev

I am responding: On behalf of myself

Name: Dmitry terentyev

Name of Organization: The Ohio State University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: In my opinion, one of the biggest problem with reproducibility and rigor is that follow-up studies are not encouraged and not funded. The race for innovation makes it impossible for the researcher to spend time an invest financially to test whether study published earlier can withstand a challenge with refined/novel experimental approaches/tools. Furthermore, more meticulous experimentalists are getting punished in the environment with shortened grant cycles and overall attitude favoring fresh, exiting, and borderline scandalous stories vs boring tedious data collection. There is a massive internal conflict between fast and rigorous, between exciting and meticulous, between deep diving into the problem and jumping between research fields. So, if my research withstands scrutiny and repetition for almost 30 years, I do not think it's a good idea to punish me for being "boring".
Indicators: High quality publications in journals with rigorous review process is the strongest indicator of researcher's scientific "value". Sure, they might not be cited a lot, but review process must be rigorous. For example, it is hard to publish both in Circulation (IF40) and Journal of Physiology London (IF4). The review process in Journal of Physiology is more rigorous than in Circulation whose impact factor is an order higher.
Publications cannot be discounted as a measure of researcher's productivity. I see more and more scientists who do not publish at all and spend all their time writing grants. In recent years these grant-writers became significantly better funded than researchers that regularly publish their findings. If you publish something - it is not novel-exiting anymore. So you cannot finish this line of research because you will not get funds already distributed to people who intend to hold their finding unpublished for decades.

2. Data, Software, and Model Sharing: It is questionable how much data sharing can help validation of the results. The parallel independent studies of the same problem can have much better outcome for correct solution for a specific problem. There is no doubt that collaboration is good. However, grant reviewers often see too many Co-Is as a major weakness. Furthermore, collaborators need to get paid, which is not feasible with current acceptable grant budget limits.

3. Training and Mentorship: N/A

4. Collaboration: Increase in PPG grant funding could be a good solution

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: High-risk high reward research still have to be balanced with improving our understanding of fundamentals.

7. Public Impact: N/A

8. Other Areas: N/A

Description: Several opinions

22. Daniel Herschlag

I am responding: On behalf of myself

Name: Daniel Herschlag

Name of Organization: State Farm

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: The emphasis is wrong. We are wasting much time and effort trying to reproduce 'old' results instead of ensuring that results lead to models and that THOSE models get tested --that's the most effective way to test the validity of prior results while moving science forward.
And the best ways to do this are to remove the reward structure for publishing in Science/Cell/Nature and to fund (much) more research so that scientists can explore their ideas of interest and are not incentivized to come up with 'big' results that are often (typically) marketing exercises that do not move a field forward as much as oft-derided 'incremental' advances do.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: Hard or impossible to have one assessment for this --but limiting lab sizes (say <=12) is generally a good idea.

4. Collaboration: You should fund it (at NIGMS) instead of expecting MIRA awardees to collaborate with no funds to work with --this is unrealistic and a disincentive. Removing program project grants several years ago while stating that collaborations are important was, at best, a mixed message. Change happens through funding, which there's not enough of and too much is targeted rather than the fundamental or exploratory, which will fuel science decades into the future.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: They don't have to bold to be of value. They do have to be well grounded and creative.

7. Public Impact: It is good when we know more.

8. Other Areas: N/A

23. Paul O'Connor

I am responding: On behalf of myself

Name: Paul O'Connor

Name of Organization: Medical college of georgia

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Peer review does not and cannot examine the quality of data. Institutions need to take more responsibility in verifying researches work.
While reproducing everything is cost prohibitive the main findings where possible should be reproduced. Perhaps basic scientists could pre-register with their institution critical studies for a manuscript (I.e main figure). The institution could then verify best practices were followed and the quality of raw data. Validate no data points were excluded/based on animals entering/raw data /blinded analysis etc. These types of figures would then hold extra weight at publication.

Even replication by other labs is prone to confirmation bias.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: A bar exam for scientists testing principles of experimental design rigor and reproducibility. These concepts are often lost in graduate school and many now are legitimately unaware their research practices are below par - bending the rules makes publication and career advancement easier and (my feeling) is that this has crept into (normalized) in lab training. Fooling ourselves is the easiest thing to do but very few labs emphasize how to avoid in practice anymore. Teaching how to avoid bias needs to be explicit at all graduate schools and is generalizable and easily testable no matter the topic. If they don't have already, NIH should develop a training module similar to others (university of Glasgow) and test knowledge of people applying for funding. There will still be people bending the rules and producing less rigorous science but they could not claim they are unaware and more likely to be called out by others with training.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: There is nothing inherently wrong with animal research. Most drugs that work in humans do the same thing in animals. The problem underlying translation (IMO) is 1. It never worked (even in animals) due to bias/fraud 2. The animal model does not fit the disease.

7. Public Impact: N/A

8. Other Areas: N/A

24. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Government Agency

Role: Other

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: Each grant currently must provide resource and data management sharing plans, but adherence to these is not reported in a way that can be tracked over time. Incentivization can include "scoring" the extent of sharing to common NIH-supported repositories that require specific formatting to facilitate secondary use, with that score being associated lifelong to an investigator. This could serve as an internal metric that can influence funding decisions.

Team science can be encouraged, with reviewers recognizing synergistic studies that would encourage data and resource sharing as a part of the review process.

3. Training and Mentorship: Reports of training is currently under-reported in progress reports. Similar to training grant applications, a table of training opportunities and successes could be integrated into grant administrative review, with scores rather than Yes/No as an output.

4. Collaboration: Equal contribution publications are currently not tracked or equally recognized in promotion and tenure considerations. Finding a way to highlight or quantify this would be a start in the right direction.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: The majority of grants and papers revolve around "proving" the proposed hypothesis. Explicit discussion of the scientific merit of supporting or refuting a given hypothesis can be an enforced expectation as part of experimental design and alternative outcomes in grant applications and publications.

7. Public Impact: Current public information about funded grants is found in RePORTER, which only states what is EXPECTED to result from a given grant application. An additional section in RePORTER could be what was successfully accomplished from this grant that could be readily searched and highlighted in the media.

Additionally, RePORTER is difficult to absorb as a member of the general public. Adding an AI link to RePORTER that provided an overall summary of the studies (and/or resulting publications) for a given search, with outcomes that support or refute a particular subject, would be a helpful resource.

8. Other Areas: N/A

25. Gary Cutter

I am responding: On behalf of myself

Name: Gary Cutter

Name of Organization: UAB

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: When one is fostering trust, my first comment is that I was reluctant to put my name down and just skip this -The culture of fear has now landed and even responding here evokes concern. That said, I would offer the reproducibility and replicability in research are part of the scientific method, but mandate it within a grant rather than via independent replication is a tainted approach. It doesn't focus on others being able to replicate the work, but is asking the researcher to validate themselves. Certainly there are times when this is appropriate within studies, but so many NIH studies are underpower (in reality) because of funding limitations, the power to even detect lack of reproducibility or replication is interpreted as success. No one can really argue this is bad as a principle, but as a mandate it makes little sense - it is a goal; it should occur; but one size fits all is often the solution by NIH which creates misapplication of resources that could be directed at more realistic effect sizes.

2. Data, Software, and Model Sharing: The simplicity of the idea is not matched by the complexity of data and data management. There are many decisions that are made to perform an analysis, but the demands and ability to achieve appropriate documentation is diminished by a lack of funding. It is getting increasingly worse as more and more researchers use RedCap for their data management. It is a wonderful tool, but creating a dataset that can be used to elicit replication of someone's results is simply not paid for by NIH. FDA addressed this problem to a degree by mandate CDISC file creation. An excellent concept, but the cost of which can be borne by industry is simply not affordable within academic settings. There may be ways to make it affordable, but just making a CDISC file in industry can cost over 250k.

One potential for high value research projects is to provide funding to faciliate this as a last step when it is determined that a dataset is high value.

Another aspect of deterrence sharing is that when investigators are contacted to provide data, IRBs have to weigh in and approach even though the data are already in the public domain. One potential solution would be to exempt all data that has been made public from local IRB oversight.

3. Training and Mentorship: The simultaneous push to limit the number of RPGs per investigator isgoing to diminish this mentorship rather than enhance it. The funding formula would allow for more time for a PI and less for CO-Is, but as restrictions are placed on the 10% of grants that exceed 2 RPGs and replaces them with other investigators, the support for mentoring will be diluted. Mandating mentors on grants

ends up assigning or "rewarding" grants that include these linkages, but what is the probability that true mentoring exists - meetings might occur, but there is much more to mentoring that mandating the illusion of a mentor.

4. Collaboration: Collaboration and collaborators are generally easy to recognize. Rather than try to create assessments, using H-indices, multi-authored papers with varying position in the authorship as well as appearance on grants and contracts can easily be quantified. Using existing data of course favors older researchers because they tend to be better known and accessible than younger individuals. This interconnectedness can be measured, but won't this disadvantage younger investigators? Further does this enhance recognition for teams, but at the expense of potentially the quality of the isolated work horse.

5. Entrepreneurship and Translation: Science moves at varying paces depending on the disease or topic - Many institutions put too much burden for commercialization on the investigator, thus incentivizing this aspect of translation is a balance between the risk aversion of the institution, the way rewards are shared and an understanding that many developments are way ahead of their time. We wouldn't want to focus on short term gains or translation at the expense of the value of the unknown, but new or innovative.

6. Foundational Scientific Exploration: NIH has been called the institute of logistics which inhibits foundational scientific exploration trading ideas for implementation details. This is going to have to be large efforts at the reviewer level. Too much of reviews are focused not on the long term, but is the biosketch the right version, did they address the data monitoring and all the other mandates that allow for concrete criticisms, but maybe miss the downstream implications.

7. Public Impact: This is a two edge sword in this hyper partisan environment. Old methods of distribution of knowledge are fragmented and picked apart as to a point that enriching public impact is going to take enormous efforts at educating the public. They need to understand several key factors of research - failure is an option and it often more important than succeeding as it leads to better understanding. The media hype of major advances when the impacts are incremental and often small needs more visibility. NIH is so large and houses so many thousands of research results, it will be hard to make the general public realize the balance of success and failute.

8. Other Areas: N/A

26. yin chen

I am responding: On behalf of myself

Name: yin chen

Name of Organization: University of Arizona

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: NIH could create a dedicated grant mechanism, or periodic RFAs, to support independent replication and verification of key biomedical findings identified by NIH program officers, advisory groups, review panels, or the original investigators themselves as highly influential, clinically important, or central to future NIH investment. This would be especially valuable for findings from human clinical and translational studies, which are often difficult to replicate because they depend on unique patient cohorts, limited biospecimens, specialized clinical data, or complex analytical pipelines.

The original investigators could participate in or even initiate such replication applications by providing protocols, reagents, datasets, biospecimen access, or scientific context, while NIH-supported independent examiners, validation cores, or third-party replication teams would provide objective verification, similar in concept to independent monitoring or adjudication in clinical trials. Such a mechanism would not be intended to criticize original investigators, but to strengthen confidence in high-impact findings before they are used to guide clinical trials, therapeutic development, diagnostic decisions, or major follow-up investments.

2. Data, Software, and Model Sharing: NIH could strengthen incentives and infrastructure for data, software, and model-system sharing by designating a set of NIH-approved, searchable repositories for raw data, analysis code, software tools, protocols, and model-system information. Many tools and repositories already exist, but they remain fragmented, inconsistently used, and difficult to search across. NIH could build or designate a PubMed-like discovery framework for research outputs beyond publications, in which datasets, code, computational pipelines, animal or cell models, reagents, and protocols receive persistent identifiers and are linked to grants, publications, and investigators.

3. Training and Mentorship: No Comment

4. Collaboration: NIH could leverage the recent removal of the prior-approval requirement for applications requesting $500,000 or more in direct costs to encourage more ambitious team-based science. Large interdisciplinary projects often require multiple investigators, shared infrastructure, and complementary expertise, but current metrics still emphasize individual productivity. NIH could use this policy change to reward well-justified collaborative team science.

5. Entrepreneurship and Translation: no comment

6. Foundational Scientific Exploration: NIH may create additional mechanisms to reward groundbreaking foundational science, while distinguishing bold, evidence-informed exploration from unsupported speculation. The current “no preliminary data required” concept can be confusing. High-risk research needs to be innovative and paradigm-shifting, but still built up on prior observations, conceptual plausibility, or limited supporting data. Because review of groundbreaking ideas depends heavily on human judgment, NIH could consider partial anonymization, broader reviewer pools, and inclusion of international experts when appropriate. These approaches could help NIH support unconventional but achievable science more fairly and rigorously.

7. Public Impact: no comment

8. Other Areas: NIH may also consider grant review reform as part of measuring and rewarding scientific impact. For example, NIH could consider a mechanism for applicants to respond to factual errors or clearly unsupported critiques before final scoring.

27. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Areas ripe for replication and reproducibility efforts can best be determined by active researchers. Congress rightfully assigns funding to broad areas of biomedical research and application, but funding decisions about specific areas must be made by experts in these areas. NIH's peer review system is well suited for the task.

2. Data, Software, and Model Sharing: Large data sets allow investigation of complex processes, but the use of data collected by other teams cannot excel in rigor and reproducibility. Small scale studies with full control of relevant variables are needed. Funding must be provided to both approaches.

3. Training and Mentorship: Usually, the better the training, the better the results. But trainees are not all equal, and more qualified trainees can achieve more by receiving better training. If the goal is to train the best biomedical personnel, success can be measured by measuring the later achievements of the trainees.

4. Collaboration: Some questions require collaborative research groups, but other questions are better addressed through small groups. Success should not be measured by the extent of contribution to team-based science.

5. Entrepreneurship and Translation: Entrepreneurship is valuable, but basic research is essential. There is plenty of snake oil in the biomedical marketplace. The current balance between basic research and translation in the NIH and NSF is appropriate. This is not an area that needs change.

6. Foundational Scientific Exploration: Yes, innovation is a cornerstone of biomedical research. The current peer review system at NIH includes two main components: Significance/innovation and Rigor/feasibility. Scores for the first component are considered fundamental and can only be made worse by scores for the second component. Thus, high-risk high-reward initiatives are already being emphasized in the review of grant applications.

7. Public Impact: Because clinical practice involves much more than the direct application of scientific knowledge, it is difficult to evaluate the true public impact of biomedical research even many years down the road. Rigorous clinical trials are extremely helpful, even if applicable only to a few situations.

8. Other Areas: No comments

28. Oxy-17 LLC

I am responding: On behalf of an organization

Name: Pradeep M. Gupte

Name of Organization: Oxy-17 LLC

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Independent reproducible data generation is an excellent approach. Getting others to confirm one's findings during this phase will be most useful and democratic.

2. Data, Software, and Model Sharing: Sharing of data generated saves duplication of effort and dollars spent for the same data.
Sharing also exposes the PI to others. A very important step in sharing knowledge for the greater good.

3. Training and Mentorship: This should also be done for SBIR/STTR funded projects. This way scientists/entrepreneurs can be developed in today's fast paced AI world. The world is now accessible to everyone and that has increased the pace and reach.

4. Collaboration: A team based approach to investigation will yield faster and authentic results. This will enable meeting the objectives of NIH and use the taxpayer funds effectively.

5. Entrepreneurship and Translation: With AI this has to be the area to focus and fund. This technology is making every global potential participants. With internet and video calls etc the number of researchers who can and will participate has exploded. We must take that into account or loose the initiative in new ideas to others who are just waiting . WE MUST CHANGE OUR PRIORITIES .

6. Foundational Scientific Exploration: Universities will always lead this early theoretically based investigations but data sharing will reduce the waste of time and money.
New ideas will mushroom faster and with gl9bal access this has the potential to transform the way we livea and enjoy our environment.

7. Public Impact: There has to be someway to relate or quantify dollars spent and the outcomes of each research funded. Positive or negative data. Etc.

8. Other Areas: Applying AI to glean and develop models will further enhance the pace of using data in new ways to get new and fast interpretations of the same. There has to be the outcome as the tool to judge effectiveness of dollars being spent.

29. Andrew Belmont

I am responding: On behalf of myself

Name: Andrew Belmont

Name of Organization: University of Illinois, Urbana-Champaign

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Generally, important discoveries have always been expected to be followed by confirmation by others. In basic science, if results can not be confirmed, then ultimately over time the field will question the discovery and challenge what can not be repeated or what is inconsistent with the older discoveries.

The challenge is more for where there is large investment in translational research or commercial applications and certainly medical trials.

But I am worried about over-reacting to the point that funding that supports new discoveries takes a major hit in order to fund disproportionally replicate studies.

The important point is to support many avenues of research, rather than the field converging on just one view point. What happened in the Alzheimer's field is a good example where funding went excessively to one particular theory of the disease. I would say the problem here was not the lack of replication of the one dominant paper that led to this, but rather the subsequent assumption that all funding should be steered towards one model of the disease.

2. Data, Software, and Model Sharing: I believe this train has left the station. Journals increasingly are asking for deposition of these data sets as a requirement for publication. In genomics, people already recognize the contributions of scientists generating the large data sets and those that focus on analyzing these data sets.

What might be helpful is ensuring more stable long term support for the data repositories. For example, in my laboratory we have placed a great deal of data on the 4D Nucleome Data portal. Right now there is free hosting by Google of much of the data and the data repository is still there. But we are now depositing past data also in GEO in case this data repository goes away due to lack of funding.

Having funding to maintain the life time of the data repositories is a major challenge that should be addressed.

3. Training and Mentorship: I don't think NIH needs efforts for "promoting the growth and stability of the biomedical research workforce". The goal of NIH should be maximizing research output as a function of tax payer dollars- not creating a secure job for scientists. The two are related but not exactly. Job force

stability should be promoted to the degree that it maximizes output but it should not an end in itself.

1. In NIGMS they have pushed the idea of sustaining the maximum number of scientists rather than maximizing the scientific output. They have argued that these are the same based on studies that have tracked output per research dollar.

However, I would ague that their analysis is fundamentally flawed and may be wrong. The argument has been that the number of publications or the total number of citations per research dollar actually declines at high levels of funding. However, the number of publications or total number of citations I would argue is a poor metric of scientific impact.

As Erin O'Shea once said during a NIH study section, "We are not farmers, that should be maximizing yield per acre" - ie number of publications. We should be maximizing new, highly impactful scientific discoveries, which does not scale with number of papers.

More specifically if you compare top University departments with average University departments the number of citations per Professor does not vary that much. Also the h-index of faculty doesn't vary that much. What does vary is the number of HIGHLY CITED papers. This has been shown in publications. Whereas the number of publications or h-index may vary by small amounts (ie ~30%) from the top department to top 30 department, the number of papers cited 100 times or 500 times or 1000 times varies several fold.

Thus there is an important analysis that needs to be done before NIH moves to much in this direction. If you repeat the study about impact per research dollar, how would the results change if you use as a metric the number of highly impactful papers rather than number of papers?

2. It needs efforts for stabilizing overall funding from year to year. The switch to multi-year payouts for new grants from the budget year in which these grants are approved conceptually is a good idea as it will significantly buffer the impact of small changes in funding from year to year.

However, its sudden implementation is on the shorter term absolutely decimating American biomedical science as we make the multi-year transition to the new budget model. Many of us will not survive this transition.

4. Collaboration: Encourage journals to require more paper specific descriptions of each author's contributions to the final research paper. Right now many journals use generic categories- ie "Conceptualization, Funding, Methodology, etc" which provides essentially no real information about each person's contribution.

Much more meaningful would be to describe each author's contribution to each Figure or experiment described in the paper.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: I don't believe NIH actually believes this- at least NIGMS. If they did they would reverse directions in which they have limited investigators to a single grant and even a

single grant application.

The message I am getting from NIGMS to younger investigators is that they can start with a a NIGMS grant as a new investigator, but then to grow their research program in the future they will need to mature by applying for funding through other Institutes for translational work.

It used to be that young investigators with creative and innovative ideas could rapidly grow their research programs by securing a second grant on a different or related direction that would allow them to build rapidly on new advances. Now with the idea of MIRA grants, etc, they are limited in the growth of their research program to incremental growth every 5 years with their MIRA renewal- but no faster.

This is terrible as it doesn't recognize the innovation does not occur steadily but in bursts. The current NIGMS model prevents exploitation of new advances which frequently occurs with new investigators who are more creative and have newer expertise, etc. In the past young investigators could create entirely new fields of research over 5-10 years of establishing their own laboratories. But the NIGMS new funding mechanism prevents that for investigators focused on basic science;.

The hidden message seems to be that basic research is no longer valued relative to translational, disease-oriented research.

7. Public Impact: N/A

8. Other Areas: N/A

30. Iraida Sharina

I am responding: On behalf of myself

Name: Iraida Sharina

Name of Organization: UTHouston McGovern MS

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Absolutely agree: No science exist without proper statistical analysis and reproducibility.

2. Data, Software, and Model Sharing: It is also necessary to introduce reasonable rules for data reuse—similar to citation or authorship—to create a clear incentive for research groups to share their hard-earned data. Furthermore, commercial data usage must be addressed. Paying a dividend to the original source for using critical information would be fair; if the music industry can implement such a model, the scientific community can too

3. Training and Mentorship: Turnover rates for faculty and, especially, research assistants serve as strong indicators of workplace stability. Currently, retaining supporting research personnel is exceptionally difficult because their salaries rely on extramural grants and are inherently temporary. This instability is highly disruptive to preserving a research group's technical expertise and legacy.

4. Collaboration: Scientific collaboration is vital, but it becomes highly complex when funding, labor distribution, and credit are involved. The NIH could significantly enhance collaborative efforts between research groups by establishing clear guidelines for participation, funding distribution, and authorship within specialized collaborative grants.

5. Entrepreneurship and Translation: I agree that patenting critical data is a powerful incentive for research. However, a major challenge is predicting which data will ultimately lead to a successful end product. In some cases, it might be more effective to reward innovation retroactively—identifying and honoring the pioneering groups that contributed most to a specific technological or medical advancement.

6. Foundational Scientific Exploration: I absolutely agree. However, this must be done in a balanced way, as there is a growing tendency to favor short-term, profitable, applied projects at the expense of fundamental, basic research.

7. Public Impact: The United States has a highly successful history of funding public-interest medical research, as demonstrated by landmark advancements in cancer research and treatment. This targeted funding model should be maintained and scaled to address other critical public health and epidemiological challenges, including cardiovascular disease, healthy aging, maternal health, and neonatal care.

8. Other Areas: It is absolutely vital to insulate science from political influence. Political interference can rapidly dismantle decades of established scientific thought and legacy, ultimately dealing a devastating blow to the training and development of the next generation of scientists.

31. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Research laboratories are increasingly struggling to maintain the baseline funding required to sustain their programs. This challenge is particularly acute for Select Agent (SA) laboratories and others that must comply with additional layers of biosafety and security requirements, which are both costly and time-intensive. As the regulatory tier increases, so too does the complexity and burden of compliance.

For these laboratories, securing sufficient and stable funding to continue their work—already subject to rigorous oversight—has become progressively more difficult. Yet these facilities are essential for generating rigorous, high-quality, and reproducible data, particularly in the study of novel pathogens and potential pathogens. They possess the specialized infrastructure, technical capabilities, and expertise required for this work, and their core missions often directly support public health, economic resilience, clinical practice, public trust, and broader societal benefit.

Despite their importance, pursuing an academic career in this area is becoming increasingly unattractive. Some PIs also rely on “soft money” support, making it difficult to sustain stable academic positions. In addition, recruiting and retaining qualified trainees and staff and convincing them to build a career in a scientific area with so many restrictions is an ongoing challenge. Without meaningful intervention, academic SA laboratories will decline significantly—or even disappear—within the next 10–20 years, resulting in a substantial loss of critical biosecurity and public health infrastructure in the United States.

A viable solution would be the establishment of stable, long-term base funding mechanisms for these laboratories to maintain their specialized research programs. Such support would not only ensure continuity but also facilitate collaboration with other research groups, industry partners, and public health agencies, enabling rapid, coordinated responses to emerging threats—for example, the identification of a novel toxin serotype or the emergence of a pandemic-capable virus.

Importantly, sustained funding would also strengthen research rigor and reproducibility. It would enable independent validation of critical findings within the field and support collaborative efforts that ensure the use of appropriate and methodologically sound approaches. In doing so, it would reduce the risk of disseminating incomplete or constrained datasets that arise from regulatory limitations, and instead promote robust and reliable science.

In today’s highly competitive funding landscape, there is little capacity or incentive for investigators to prioritize replication, validation, or methodological rigor. Efforts aimed specifically at ensuring reproducibility are rarely funded and carry limited professional reward. Establishing funding mechanisms tied to already demonstrated, rigorous, and responsible research programs would both enable these essential activities and provide appropriate recognition for them.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

32. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: establish a special discussion panel to identify the important results from the PIs that hold multiple grants, either concurrently or in series. The panel officer can then contact the PI for detailed methodology, testing models etc.. The officer can then contract internal or external groups that possess similar expertise to repeat the experiment, and report to the panel the results. If the results are consistently reproducible, the group will be labeled and be subjected to a different criteria of grant award, such as awarding at score of 20% instead of regular 10% score. IF the results are confirmed not reproducible, then a penalty 5% score can be applied to that PI instead of regular 10%.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

33. Mark Chertoff

I am responding: On behalf of myself

Name: Mark Chertoff

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Maybe experiments need to be designed to show both within-and between-subject variability, and hence reproducibility. At least within a laboratory. Perhaps research that extends a concept/idea should require a piece of replication from a previous study? May be asking too much?

2. Data, Software, and Model Sharing: Repository of data sets. I think that this exists for some type of research, e.g genetic. Maybe for electrophysiology, and include a rigorous description of calibration measures.

3. Training and Mentorship: Maybe listing contributions to the paper and academic position. This may already exists?

4. Collaboration: Require a published paper with the major PIs before they can get funded for a collaborative study. This shows that they can work together.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Maintain a grant mechanism for this!!

7. Public Impact: N/A

8. Other Areas: N/A

34. Herbert M Sauro

I am responding: On behalf of myself

Name: Herbert M Sauro

Name of Organization: Univesity of Washington

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: My area of work is supporting the reproducibility of computational models in systems biology through publishing best practices and providing technological solutions in the form of grass-roots driven standards for model exchange and representation. We are also pursuing a similar strategy to develop a grassroots standard for describing computational experiments in a reproducible manner. These standards are designed to allow persistence of both models and reproducibility of experiments over a 20-year time frame.
A comment on terminology: In computational modeling we define repeatability as the ability of a user to repeat their own work and reproducibility as the ability of a third-party to recreate what was done. I assume the technical terms you are using are from the National Academies report. The panel that wrote that report did not have any biologists or molecular biologists, and so the language they use doesn't always match what the biological community, or in fact other disciplines, use in practice.

In order to encourage reproducibility, which is the most important of the two, the best mechanism is to provide incentives. One such incentive is that it has been recently shown that authors who published reproducible models (using our approaches) get cited more often. The simple reason is that reproducible models can be more easily reused. Another more forceful approach would be for the NIH to demand that the computational models be deposited at a central location (such as NLM) and would have to be reproducible in order to get, say, a NLM DOI.

As for rigor, that is served by peer review. However, because of the extra activities scientists have to do today, peer review is given less attention than it should. In addition, journals now require very fast turnaround, which is not consistent with rigorous review. When I was a postdoc and a graduate student, my supervisor would spend up to a week reviewing a single paper. I suspect that rarely happens today.

2. Data, Software, and Model Sharing: Model Sharing: In systems biology, the sharing of models is largely a solved problem, provided researchers use the established open standards such as SBML. It is estimated that over 50% of published models now use this standard, which has resulted in the establishment of large model repositories such as Biomodels. In addition, all genome-scale models use SBML and large repositories of such models now exist. The use of SBML has had a significant impact on both reproducibility and access to published models. In addition, any model in SBML can be more easily reused in future work, preserving the investment made by the NIH and other agencies. The human-readable form of SBML, called Antimony, is also well-suited for AI use, and the most well-established AIs

understand Antimony very well. SBML has one other significant advantage: it stores the biological information alongside the quantitative aspects of the model. Models represented as pure mathematical equations lose the vast majority of the biological information from which they were derived; SBML does not. In practice, SBML compilers exist that convert the SBML into a target execution language such as C/C++ , Python etc.

There are researchers who continue to use executable formats such as MATLAB or other programming languages to publish their models. These models are largely inaccessible, cannot easily be reproduced, and cannot be reused without considerable effort.

While there is an established model-sharing standard for subcellular models, very little effort has been devoted to developing other open standards for larger-scale systems, specifically multicellular, organ, whole-body, and even population models. Part of the problem is that the NIH has never had no funding program to support such activities.

Interestingly, during the pandemic, because infectious models were based on executable formats such as C, Java, etc, none of the models could be easily reduced or even understood by others. Later into the pandemic, some groups successfully used SBML to store infectious models. The episode illustrated the problem in publishing models as executable artifacts rather than in terms of the underlying biological systems.

Software Sharing: Software sharing has been largely solved with sites such as GitHub. Personally, I don't think there is an issue today with sharing software other than the researcher making the effort to make it happen.

Data Sharing: This is by far the hardest to problem to solve because of the huge variety of data types and the fact that new data types keep getting developed because there are new ways and new things to measure. I don't have a good solution to this.

3. Training and Mentorship: N/A

4. Collaboration: Collaboration is best done at the grassroots level, especially between graduate students and postdocs from different research groups. Such collaboration can be fostered by small grants to fund travel and accommodation, the NSF does this very effectively. The NIH instead expects us to go through a full grant proposal to request relatively large amounts of funding that are not necessarily required. Smaller funds ($20K-$20K) at more frequent intervals is far better for encouraging collaboration. Assessment can be done by the group writing *short* reports on what they accomplished, this could be structured in a way that would be easily digestible by LLMs. Over time a record of contributions could be created.

5. Entrepreneurship and Translation: As researchers we are always exposed to translation opportunities. The SBIR and STTR awards are excellent ways to help stimulate initial translation from the lab to the bed side. I hope that the NIH will continue these opportunities.

6. Foundational Scientific Exploration: Foundational research is the bedrock of all innovation and discovery, providing the start of a long pipeline from lab to treatments and technology. It has been one

of the great success stories of the US scientific enterprise of the last 80 years, and for whatever reason, it has not been replicated as successfully in other countries. The key to foundational research is that it should be undirected, meaning we should avoid placing biases on the direction of research because this will result in missed opportunities. The study of other organisms has resulted in a wealth of knowledge that has been translated to human biology and disease. A recent example of this is CRISPR, which started off as a curiosity-driven study of research into unassuming repeat sequences in bacterial DNA. Who would have known it could eventually lead to the potential to cure inherited disease? The development of AlphaFold relied on over 60 years of protein research to access the necessary training data. Foundational research plays the long game; discoveries 20 years ago may only now have an impact on the general population. This is why it is critical to maintain funding for foundational work, a break now will be felt 10 or 20 years into the future.

7. Public Impact: I would encourage researchers at universities and other institutions to make an effort to communicate with the general population on the work they do. Not all work of course will lead to immediate health improvements. However, the public is also receptive to research that is interesting and exciting from a scientific point of view if communicated well. Obviously research that is closer to the bed side will be of more interest but from my experience the general population is also interested in foundational work related to human physiology or how discoveries in other organisms can highlight similar process in humans. The general population also needs to be aware that research takes time, a single grant to study an autoimmune disease won't result in a cure in 3 years, but 100 grants over 15 years may.

Finally, universities in particular need to change their assessment criteria for promotion. When I applied for promotion, outreach was not a factor that was considered. This needs to change. In addition, any outreach mentioned in grant proposals is generally either ignored by panels or considered an weakness. That also needs to change. The more we communicate with the public the better.

8. Other Areas: Lighten the bureaucratic load on researchers. Most of this however comes from the federal and state governments so perhaps little can be done. Given the growth in paper work I estimate that within 100 years almost 100% of our time will be spend on bureaucracy.

35. Olivier Le Saux

I am responding: On behalf of myself

Name: Olivier Le Saux

Name of Organization: University of Hawaii

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Here is my 2 cents. I my view there is more to this than how science is done but more systemic change also need to happen.

1- Data should systematically include larger number than n=3. I still commonly see n=3 and more recently n=2 with p values in a prominent journal !

- Peer-review should be strengthened to include rigor and reproducibility, model validation, etc...
- Faculty evaluation should evolve (perhaps return to the tenure system) to reduce pressure to publish AT ALL COST for continued employment, promotion and for extramural funding.

2- Analysis from cultures or animals shouldn't be divided into 3 or more samples but come multiple independent animals or cultures. Preferable not done at the same time/day/lab.

3- Blind analysis or double blind experiments and analyses whenever possible even in the most mondain basic science experiment.

4- Duplication and validation (of models) in another lab. Whenever appropriate, validate model with human data.

5- Anonymity for submitting authors and PI whether for publication or grants should be considered: see below in collaboration. I understand this would get in the way of the evaluation of the environment and feasibility in grant reviews. But reputation is a major bias in peer-reviews.

I have personally done 3 and 4 with various collaborators whereby we duplicated (whenever possible) each other's experiments during collaboration and obtained human data and/or samples.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: This notion should be expanded and engrained in graduate education.

4. Collaboration: As stated above, collaboration can be used as an essential tool for rigor and reproducibility but the mindset should change with respect to competition and data theft (this comes from experience and I am not alone here). I have published data (from my lab only) that I partially

question the reproducibility of, years later. Not that it wasn't rigorous at the time but subsequent results and exchanges at conferences make me wonder whether the model or the experimental approach was correct. Collaborating on this research would have been beneficial. Recently, an article was published that contradicts some of my published results. Here, I have confident in my data (multiple line of independent data) as it cames from 2 labs, animal models + validating human samples/analyses. This competing lab - despite prominent names on the publication - has a history of sub-par rigor in their publication.

Heavy weights in various fields should NOT receive an easy pass simply because of reputation or a long list of publications.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: perhaps some special NIH NOFO for data reproducibility should be developed and offered.

7. Public Impact: Better education and outreach of the public of what scientific research AND data entails. We, scientists, are sorely lacking on communication.

I found that the public (i.e. friends, social acquaintances) have simply no idea of what I do as a faculty and researcher and how I do it. I have given many talks in patients meetings and I can tell how hard it is from lay people to understand what 15+ years of education and training has taught us. Few lay person understand critical thinking, or that science is not a belief system.

A simple example far from biology: hardly anyone understands that there IS gravity in space, why? Simply because since Apollo, everyone has seen astronauts floating in space and thereby assume there is NO gravity. Appearance can be deceiving.

8. Other Areas: N/A

36. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: Publishing in peer reviewed journals is critically important for assessing rigor. It is important to have peers review the work. Presenting at national meetings with audience feedback is also helpful. Supporting “science of science” research through CTSAs is also important for knowing what works.

2. Data, Software, and Model Sharing: Central websites where data can be shared would be helpful. Data safety rules are key.

3. Training and Mentorship: Mentoring should be part of academic life. Instead of mentoring grants, how about a small percentage of effort added to grants for incorporating junior faculty as mentees and continuing the pipeline of research.

4. Collaboration: Effort for researchers on grants is key to encouraging collaboration. There should be support for clinician researchers so the health impact isn’t missed.

5. Entrepreneurship and Translation: Competitions with funding could help.

6. Foundational Scientific Exploration: N/A

7. Public Impact: More support for health services research and research on health quality is needed. This type of research can make significant impact. In addition, more support for clinician researchers is needed to make sure there is truly clinical impact.

8. Other Areas: N/A

37. Gregory A. Pesko

I am responding: On behalf of myself

Name: Gregory A. Pesko

Name of Organization: Harvard Medical School/Brigham and Women's Hospital

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: I strongly urge you to read Stephen J. Gould's "The Mismeasure of Man". It may keep you from making the same tragic mistakes that the people who developed the IQ test - and the people who misused it - made. There are two fallacies in your effort: the first is that something as complex and multifaceted as research success is susceptible to accurate qualification. This is called the fallacy of reification. The second is the fallacy of ranking, in which imperfect or inaccurate quantities are ordered and used as the basis for making decisions.

Do not fall into this trap. It is damaging to the United States's scientific enterprise and will lead to important issues of impact and quality being made by bureaucrats instead of by peers.

2. Data, Software, and Model Sharing: Same comment. Stop thinking you can measure what cannot be measured accurately.

3. Training and Mentorship: Same comment. Stop thinking you can measure what cannot be measured accurately.

4. Collaboration: Same comment. Stop thinking you can measure what cannot be measured accurately. Instead of trying to assess this, look for ways to incentivize it, such as promoting collaborations with foreign institutions instead of impeding them. Quality science is an international enterprise.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: These approaches are already embedded in the academic enterprise. The problem is that ham-handed attempts to judge research based on its concordance with the goals of the part in power will never be able to evaluate this kind of science properly. It doesn't have a political aim or even a practical end. It is driven by the curiosity of the individual investigator. It must be protected from political interference and from the mistaken notion that science must serve a practical purpose.

7. Public Impact: Same comment as the first section. You can't properly assess the broader impact quantitatively. It's like Potter Stuart's definition of pornography: you know it when you see it. The best approach is to have these things evaluated qualitatively by other scientists.

8. Other Areas: N/A

38. David Huang

I am responding: On behalf of myself

Name: David Huang

Name of Organization: Oregon Health & Science University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/23/2026

1. Rigor and Reproducibility: To alleviate the pressure for investigators to publish only statistically significant results, NIH should host an electronic publication that allow grant awardees to publish all results, whether statistically significant or not, that were specific aims of funded grants.
For research results that are changing clinical/research practice and having significant impact, NIH should specifically fund studies to replicate the results if they had not be replicated already.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: NIH should develop indices to track translational impact of investigators and discoveries in terms of licensed patents, commercialized products, shared software, and resulting diagnostic and therapeutic procedures and drug prescriptions.

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

39. Jon A. Krosnick

I am responding: On behalf of myself

Name: Jon A. Krosnick

Name of Organization: Stanford University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: I am writing to propose two topics for future research efforts in this arena.

First, the basic problems of reproducibility and replicability of empirical findings are well documented. What is less understood are the causes of such problems. Any efforts to reduce such problems must begin with a rigorous understanding of the psychological, organizational, and cultural factors that cause the problems. And yet, remarkably little scholarship has been devoted to developing hypothesis about those causes and to rigorous testing of those hypotheses.

My colleagues and I proposed a model of the causal processes at work:

https://pubmed.ncbi.nlm.nih.gov/31565236/

Designing interventions to minimize problems of reproducibility and replicability will be significantly advantaged by tests of this model.

Secondly, I'd like to identify an important likely reason for failures to replicate: inattention to the process of sampling participants in studies of human subjects. For nearly a century, survey researchers have known that drawing a random sample of participants from a well-defined population allows for generalization from the sample to the population, even when response rates are considerably below 100%. But a body of accumulating evidence shows that recruiting participants for a study in a non-random way yields results that are routinely notably less accurate and that vary considerably from sample to sample.

Yet during these many years, the field of psychology, the field of medicine, and others have generated thousands of research studies with non-random, haphazard samples of participants. Because a huge body of literature shows moderation of effects by moderator variables (including characteristics of participants), there is a strong theoretical basis to expect that results of studies will vary from study to study if the participants are not randomly sampled from a single population of interest. Such variation can be responsible for the appearance of lack of replication, when in fact, the problem is lack of generalizability of findings across population subgroups.

This problem can be solved by implementing random sampling from well-defined populations for more

studies, even if initial, exploratory investigations employ haphazard samples. It is therefore worthwhile to document how replication rates increase when participant sampling is made more systematic.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

40. Aaron James

I am responding: On behalf of myself

Name: Aaron James

Name of Organization: Johns Hopkins University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: I would suggest identifying those researchers who consistently retract published work. This should be considered at council meeting or some other place in secondary review.

2. Data, Software, and Model Sharing: Incentivizing such activity should take the form of specific grants to support these activities.

3. Training and Mentorship: I believe that growth and stability come from stable funding for biomedical research. More than that, expansion of funding to improve funding rates, re-institution of paylines, easing restrictions and fees on H1B visas, and increasing the modular R01 budget would go far to ensuring stability of the biomedical research workforce

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Agree. R21 funding amounts should be doubled in size and extended to a 3 year period. The current rates do not allow for much science to be conducted.

7. Public Impact: N/A

8. Other Areas: N/A

41. Gordon W. Laurie

I am responding: On behalf of myself

Name: Gordon W. Laurie

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: Rigor and Reproducibility are themes inherent in the practice of science. Full transparency of data has been well adopted in many journals, but not in all moderate to lower tier ones. A uniform standard could be required/encouraged of all journals. NIH should be wary of introducing investigator burden.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: Themes of Rigor and Reproducibility are seeded in strong training and mentorship. Joint mentorship of trainees can be very healthy. Also helpful is a third mentor from a different department who meets regularly to be supportive of trainee well-being.

4. Collaboration: Collaboration develops organically to accelerate a discovery or when discovery screening. Team science should not be forced. Through assessment, NIH should be wary of introducing investigator burden.

5. Entrepreneurship and Translation: I would encourage all predoctoral training grants to include a company internship requirement of 1 - 3 months. Also, there should be more cross-fertilization between business schools and the sciences. SBIR funding is outstanding for early stage company growth. VC investment beyond this stage can be very challenging.

6. Foundational Scientific Exploration: As a human enterprise, the burden and risk adverse nature of study section review can suppress enthusiasm for truly innovative proposals. The NIGM R35 mechanism is admirable.

7. Public Impact: N/A

8. Other Areas: N/A

42. T John Winhusen

I am responding: On behalf of myself

Name: T John Winhusen

Name of Organization: University of Cincinnati

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: Multi-site clinical trials are among the most rigorous forms of biomedical research and already incorporate many features intended to enhance reproducibility, including prospective registration in ClinicalTrials.gov, pre-specified primary and secondary outcomes, statistical analysis plans, independent safety oversight, protocol standardization, monitoring, and adherence to CONSORT reporting standards. NIH should consider successful completion and dissemination of registered clinical trials as indicators of rigorous and reproducible science.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: Multi-site trials provide unique training environments that expose junior investigators to protocol development, regulatory oversight, participant recruitment, data management, and collaborative science. These contributions should be recognized as valuable mentoring activities.

Suggested indicators:

Junior investigators mentored within networks;
Authorship opportunities for trainees;
Career advancement of mentees;
Participation of early-stage investigators in trial leadership activities.

4. Collaboration: Multi-site clinical trials often require years of coordination among investigators, coordinators, statisticians, laboratories, and clinical sites. Leadership of such studies frequently results in fewer publications over a given period compared with smaller studies, creating disadvantages within promotion and funding environments that emphasize publication counts. NIH should develop metrics that explicitly recognize contributions to team science and leadership of large collaborative studies.

Suggested indicators:

PI or MPI roles in multi-site trials;
Number of participating sites coordinated;
Contributions to protocol development;
Steering committee leadership;
Mentorship within clinical trial networks.

Potential incentive:

Explicit recognition of trial leadership and network contributions in biosketches and grant review.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: The impact of clinical trials is often not adequately reflected by traditional citation metrics. Once trial findings are incorporated into systematic reviews, meta-analyses, clinical guidelines, or evidence-based recommendations, subsequent publications frequently cite these secondary sources rather than the original trials. NIH should consider inclusion in systematic reviews, meta-analyses, practice guidelines, and policy documents as indicators of scientific influence and public health impact.

Biomedical research evaluation should reward the generation of reliable evidence that informs clinical practice and public health, even when that work produces fewer publications and accrues citations more slowly than smaller, less resource-intensive studies.

Suggested indicators:

Inclusion in systematic reviews;
Inclusion in meta-analyses;
Citations in clinical guidelines;
Influence on healthcare delivery or policy;
Evidence of changes in clinical practice.

8. Other Areas: Multi-site clinical trials represent one of the most rigorous and impactful forms of biomedical research but are inadequately recognized by traditional measures of scientific productivity. These studies require substantial coordination, infrastructure development, regulatory oversight, and long-term commitment, often resulting in fewer publications over a given time period compared with smaller studies. Clinical trials already incorporate robust mechanisms that support reproducibility, including prospective registration, pre-specified analyses, independent monitoring, and standardized reporting. Furthermore, the scientific influence of clinical trials may be underestimated by citation-based metrics because individual trials are frequently incorporated into systematic reviews, meta-analyses, and clinical guidelines that subsequently become the preferred sources for citation. NIH should consider metrics that recognize contributions to multi-site trials, inclusion in evidence syntheses and guidelines, successful completion of registered studies, and leadership roles in team science initiatives.

Description: Multi-site clinical trials already embody many of the practices NIH seeks to incentivize

43. Desmond Heath

I am responding: On behalf of myself

Name: Desmond Heath

Name of Organization: None

Type of Organization: Academic Institution

Role: Healthcare Provider

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: This communication fits your opening statement that NIH "welcomes all comments." It describes research outside the specific interests you listed.
It hardly relies on "measurable indicators" except that this research aims at identifying regularities in the behavioral expression of the constraint that two physical properties of matter have on each other. These physical properties of matter are expressed throughout nature and are utilized by or brains in the construction of mind. The "potential benefits" are in the construction of Laws of Mind that might set off a Sane Society Revolution just as Newton's Laws of Motion set off the Industrial Revolution.

I am happy to see that this initiative by NIMH hews to Bacon's Razor in that it focuses research on matters that might benefit human kind and relieve suffering.

If the construction of Laws of Mind interests you I will give you a full description of my research.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: I hope this Laws of Mind project qualifies as "basic research" that is of some risk but it is of m"high reward."m

7. Public Impact: Set off a Sane Society Rally evolution .

8. Other Areas: N/A

44. David Schopfer

I am responding: On behalf of myself

Name: David Schopfer

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: Rigor and reproductibility are already very well achieved in most federally funded research.

2. Data, Software, and Model Sharing: Validation of research should require that researchers publish their methods in detail (appendix or elsewhere), beyond the very cursory descriptions currently provided in manuscripts. Especially for research that involves datasets, common data models, machine learning, and AI, the specific computer code for every single decision, analysis, and results should be required to be directly published if federal funds are received.

As a researcher, numerous times I try to understand what others did to acheive their results, and the published information is too vague and nonspecific to make any conclusions, so of course I cannot replicate it even with the same dataset. That is unacceptable. It's not proprietary information. It should be shared.

3. Training and Mentorship: The new, existing policies and programs are going to directly undercut growing the biomedical research workforce. The administration does not understand or seem to care that their actions are directly harming the very thing they claim to care about. I know numerous young investigators and potential investigators who are completely avoiding this career path due to the current policies in place. No amount of training awards or programs will change that. Young investigators will not choose altruism, lower income, and significant job instability over higher paying and stable job options.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

45. Donald Lynam

I am responding: On behalf of myself

Name: Donald Lynam

Name of Organization: Purdue University

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: I think that historic transparency should be considered in the reivew process. Has the investigator routinely engaged in Open Science practices in some preceding period (5 years)? Do they preregister their studies--a priori hypotheses; sample size; decisions rules for outliers, covariates, and missing data; outcomes; statistical approaches, etc. Do they make their data publicly available? If data are deemed too sensitive, do they make synthetic data sets available? Do they make their analytic code publicly avialble? Do they make publications open access (either by publishing in OA journals or sharing on preprint servers? This needs to be based on past behavior not future promises.

NIH could submit past publications (2, 3, 5 years) to a z-curve analysis to quantify false discovery rates as a means of detecting questionable research practices. I suspect AI will be capable of doing this shortly (if it cannot do so already).

2. Data, Software, and Model Sharing: why can't NIH demand it?

As above, make historic data sharing and code sharing an important part of the review process. I do not think there is ever a reason to not share code. If an investigator claims the data cannot be fully deidentified, then they can create synthetic data sets that at least allow checking of results.

3. Training and Mentorship: Again, I suggest a research history approach.
How many students have the mentored? How many students have successfully completed their degreees. How many students have first-authored publications? How many first authored publicatons do they have? Where have they been placed after receiving their degrees?

Has the investigator been recognized for their mentoring within their department, college, university, or society?

I do not believe mentoring plans for students will behelpful. They are simply promissory notes that are difficult to enfore and too easily gamed.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Historically, how many failures have they published? Null results are very informative and need to be published. The only legitimate reason for not doing so is a failed or grossly underpowered study. The first can happen but the second should never happen. Why is an investigator conducting a study where there is only outcome thatthey can publish.

7. Public Impact: N/A

8. Other Areas: N/A

46. Qinglin Yang

I am responding: On behalf of myself

Name: Qinglin Yang

Name of Organization: LSUHSC-New Orleans

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: When a Principal Investigator's (PI) career survival depends almost entirely on winning the next grant, hyper-competition takes over, which naturally erodes rigor and reproducibility. Shifting PIs away from the "funding-first" survival loop could significantly improve research quality. fund a scientist's overall laboratory rather than specific, high-risk project proposals. Extend standard R01 grant periods from 4–5 years to 7–8 years to give labs breathing room to troubleshoot and replicate data without panic. Reward PIs for publishing well-controlled negative findings or validation studies that save the scientific community time and resources. We will also need to curb the trend of bloated papers in the name of reporting a complete story.

2. Data, Software, and Model Sharing: While well-intentioned, the escalating mandates for research rigor and reproducibility have inadvertently created a punishing "compliance tax" that disproportionately crushes small laboratories and alienates Principal Investigators (PIs) from actual science. Unlike well-funded mega-labs equipped with dedicated administrative staff, small labs lack the manpower and financial margins to seamlessly manage intensive requirements like complex Data Management and Sharing (DMS) plans, cell line authentications, and strict double-blind protocols. This operational imbalance forces small-lab PIs to transform from innovative thinkers into full-time bureaucrats, spending the majority of their time navigating paperwork and compliance portals rather than mentoring trainees or analyzing data at the bench. Consequently, this regulatory creep consolidates funding into massive "grant factories" that can easily absorb the administrative workload, systemically starving out the smaller, agile laboratories that historically drive diverse, niche scientific discoveries.

3. Training and Mentorship: Training and mentorship are the most direct ways to improve research reproducibility because they embed rigorous habits into scientists at the start of their careers. While administrative policies force compliance from the top down, effective mentorship builds a culture of precision from the bottom up.

4. Collaboration: To fairly recognize team-based science, institutions must replace traditional first/last authorship metrics with granular, portfolio-based evaluations. Embracing standardized systems like the Contributor Roles Taxonomy (CRediT)—already used by the NIH Intramural Research Program to resolve disputes—explicitly credits essential, specialized roles like data curation, software development, and validation. Furthermore, updating tenure criteria to include narrative CVs and collaborative portfolios rewards scientists for collective impact and high-value digital artifacts, such as open-source code and reusable public datasets. Redesigning these institutional promotion and grant structures ultimately

replaces isolated, high-volume publishing with an ecosystem that actively incentivizes collaborative transparency.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: To unlock truly innovative scientific exploration, funding agencies should transit from hyper-vigilant compliance monitoring to a model of trusted stewardship that liberates PIs from administrative overload. This shift requires expanding flexible "people, not projects" awards, and implementing streamlined, two-stage grant applications that review a short conceptual abstract before demanding extensive paperwork. Furthermore, adopting "just-in-time" compliance for institutional protocols and funding centralized administrative co-ops would effectively remove the bureaucratic burden from the researcher's desk. By stripping away this operational red tape, funding bodies can transform PIs from full-time managers back into active scientists, giving them the cognitive freedom to pursue high-risk, high-reward breakthroughs and directly mentor the next generation of innovators.

7. Public Impact: N/A

8. Other Areas: N/A

47. Malaz Boustani

I am responding: On behalf of myself

Name: Malaz Boustani

Name of Organization: Indiana University School of Medicine

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: NIH should modernize rigor and reproducibility metrics by moving beyond publication counts, citation rates, journal impact factors, and grant totals. These traditional measures reward scientific production more than health creation. A paper, grant, dataset, or tool should not be assumed to have high value simply because it was published, cited, or funded. Its value should be judged by whether it produces reliable knowledge that improves health, reduces suffering, strengthens care delivery, reduces disparities, or improves decision-making for patients, caregivers, clinicians, communities, and policymakers. NIH should create a minimally viable universal set of reproducibility measures, then allow localization by scientific field, disease area, study design, and local context. The universal measures could include: whether the finding was independently replicated; whether methods, protocols, code, measures, and data are reusable; whether null and negative findings were reported; whether unintended consequences were measured; whether pre-specified termination or adaptation criteria were used; and whether the work reduced or increased scientific, clinical, or participant burden. Reproducibility should not mean rigid duplication of the same protocol in every setting. In complex adaptive health and research systems, reproducibility should include the ability to preserve essential scientific functions while allowing local adaptation. NIH should therefore distinguish between fidelity to core scientific mechanisms and appropriate localization of delivery form. NIH should also reward humility. Investigators should receive credit for identifying failed assumptions, stopping low-value work, publishing negative findings, and adapting based on evidence. These practices are signs of rigor, not weakness.

2. Data, Software, and Model Sharing: NIH should reward data, software, and model sharing when sharing creates usable scientific and public value, not merely when files are deposited into repositories. The key question should be: Can others actually reuse, validate, implement, or improve the resource? Measurable indicators could include: number and diversity of independent reuse cases; successful external validation; interoperability with other platforms; clarity of documentation; time from data collection to availability; privacy and security protections; user burden; and whether the shared resource accelerated discovery, replication, implementation, or public health decision-making.
NIH should avoid creating overly centralized data infrastructures unless the scientific problem truly requires centralization. Large centralized platforms can become slow, expensive, and disconnected from local scientific reality. NIH should instead support decentralized networks of smaller agile teams connected by common minimally viable data standards, open tools, and shared learning infrastructure. The principle should be: standardize the minimum, localize the rest. This would allow NIH to create a

common language for reuse while preserving the flexibility needed across basic science, clinical trials, implementation research, community-engaged research, and data science.

3. Training and Mentorship: NIH should measure the impact of training and mentorship by whether trainees develop the capacity to conduct rigorous, useful, reproducible, ethical, and adaptive science. Current measures often overemphasize publications, grants, titles, and institutional prestige. These are incomplete indicators of whether a scientist can improve health. Training metrics should include: ability to ask important and falsifiable questions; use of rigorous and transparent methods; capacity to work in teams; skill in data sharing and reuse; ability to report negative and null findings; competence in community and stakeholder engagement without advocacy capture; and ability to move discoveries toward implementation, diffusion, or real-world decision value. NIH should also reward mentors who help trainees develop scientific humility, learning velocity, and responsible risk-taking. A successful mentor should not be judged only by how many papers or grants their trainees produce, but by whether trainees become independent, ethical, collaborative, and mission-driven scientists. NIH should consider training programs in agile science and its derivatives. These methods help scientists develop solutions that can be adapted, implemented, sustained, and diffused in real-world systems.

4. Collaboration: NIH is right to reward collaboration and team science. However, the deeper problem is not only that team members receive insufficient credit. The deeper problem is that the measures used to judge team success are often misaligned with the NIH mission. NIH should reward teams for producing health impact, reproducible knowledge, reusable infrastructure, and learning that others can apply.
Team science should be measured by contribution to shared mission, not only by authorship position or grant role. Useful indicators could include: clarity of team roles; quality of interdisciplinary integration; shared data and tools; speed from idea to test to learning; evidence that collaboration improved rigor or usefulness; stakeholder-assessed value; and whether the team built infrastructure that others can reuse.
NIH should also minimize unnecessary centralization. Collaboration does not always require large projects, large consortia, or large coordinating centers. Large centralized structures can become slow, expensive, risk-averse, and administratively protected. NIH should support decentralized networks of smaller agile teams that can test, learn, adapt, and share discoveries rapidly. In complex adaptive systems, intelligence is distributed. NIH should fund collaboration models that capture distributed intelligence rather than suppress it through excessive central control.

5. Entrepreneurship and Translation: NIH should broaden its definition of translation. Translation is not only movement from bench to product or patent to company. Translation should include implementation into practice, diffusion across networks, reduction of burden, and measurable improvement in health, care delivery, trust, safety, access, or affordability.
NIH should reward entrepreneurship when it creates public value, not merely when it creates a company, patent, license, or investment event. Useful indicators could include: time from discovery to real-world test; evidence of user demand; adoption by patients, clinicians, communities, or health systems; sustained use; affordability; equity impact; interoperability; implementation readiness; and evidence that the solution solves a real problem without creating unacceptable unintended consequences.
NIH should also recognize that much current innovation occurs outside NIH-funded studies because NIH processes are often too slow and risk-averse. NIH should preserve scientific standards while reducing unnecessary administrative friction. It should create rapid-cycle mechanisms for minimally viable experiments, early implementation testing, and responsible risk-taking.

Agile Innovation, Agile Implementation, and Agile Diffusion offer one framework. They ask whether a solution addresses a real demand, fits the local context, can be tested quickly, can be adapted while preserving essential functions, and can be sustained or diffused across real-world networks.

6. Foundational Scientific Exploration: NIH should continue to protect foundational science and high-risk, high-reward research. However, foundational science should not be evaluated only by traditional prestige metrics. Basic research can create enormous value through mechanisms, methods, tools, models, datasets, conceptual breakthroughs, and future platforms for discovery. For foundational science, impact measures should be localized to the nature of the work. Useful indicators may include: validated mechanisms; reproducible methods; high-quality shared tools; new models; robust datasets; theoretical clarity; falsifiable predictions; reuse by independent investigators; and contribution to future scientific pathways. NIH should reward responsible scientific risk. Responsible risk is not the same as reckless speculation. It includes a clearly defined problem, plausible mechanism, transparent methods, pre-specified decision criteria, proportional safeguards, and willingness to report failure. NIH should also avoid forcing foundational science into immediate advocacy or premature claims of public impact. Basic science often creates long-term value that cannot be fully predicted at the time of discovery. NIH should protect curiosity-driven work while still expecting rigor, transparency, reproducibility, and humility.

7. Public Impact: NIH should move from a prestige-based measurement system to a mission-based measurement system. The NIH mission is to generate knowledge that enhances health, lengthens life, and reduces illness and disability. Therefore, public impact should be measured by how much health and trustworthy knowledge science creates, not merely by how much science it produces.
Public impact indicators could include: improvement in patient, caregiver, clinician, community, or population outcomes; reduction in disparities; improvement in safety, quality, access, trust, or affordability; real-world implementation; sustained use; policy or practice relevance; reduction in burden; and contribution to public trust.
Patients, caregivers, clinicians, and communities should help judge usefulness. Scientists are essential judges of rigor, but they are not the only judges of value. Patients can judge whether outcomes matter. Caregivers can judge whether interventions are usable. Clinicians can judge workflow fit. Communities can judge trust. Health systems can judge sustainability.
At the same time, NIH should keep science distinct from advocacy. Science should inform policy and practice, but it should not be confused with advocacy. Advocacy starts with a desired position and seeks support for it. Science starts with uncertainty and seeks a more accurate understanding of reality. NIH should reward clear separation between hypothesis and desired outcome, transparent reporting of null and negative findings, replication across independent contexts, and communication that informs without overstating certainty.
Public trust will not be restored by louder claims of impact. It will be restored by humble, transparent, rigorous science that is willing to learn in public.

8. Other Areas: NIH should reward agile bureaucracy, not rigid bureaucracy. Science needs bureaucracy: human subjects protection, fiscal accountability, data stewardship, privacy protection, conflict-of-interest management, peer review, and public reporting are essential. The problem is not bureaucracy itself. The problem is rigid bureaucracy.
Rigid bureaucracy asks whether every box was checked. Agile bureaucracy asks whether the process created trustworthy, ethical, useful, and timely science. Rigid bureaucracy adds layers. Agile bureaucracy learns what can be simplified. Rigid bureaucracy assumes one process fits every risk level. Agile

bureaucracy matches oversight to risk.

NIH should measure and reward institutions that reduce unnecessary administrative burden while preserving integrity. Indicators could include: time from award to study launch; time from IRB submission to approval, stratified by study risk; time from data collection to usable feedback; administrative cost per unit of scientific or health impact; investigator time spent on science versus administration; participant and clinician burden; redundant approvals eliminated; and evidence that oversight prevented harm without delaying low-risk science unnecessarily. NIH should not impose a new metric system nationally without testing it. NIH should run agile experiments across institutes, mechanisms, and review panels. These could include: progress reports focused on learning velocity and burden reduction; renewal criteria based on real-world usefulness; rapid-cycle R21 mechanisms for minimally viable experiments; mechanisms that fund decentralized agile teams; and institutional rewards for reducing the time and burden required to launch rigorous low-risk studies. NIH should consider creating an Agile Impact Score with a universal core and local adaptation. Core domains could include health impact, implementation impact, learning impact, collaboration impact, speed impact, burden reduction, replicability and reuse, and localization impact.

The overall principle is simple: NIH should measure success not only by how much science we produce, but by how much health we create.

48. Zhong Wang

I am responding: On behalf of myself

Name: Zhong Wang

Name of Organization: Washington State University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: Current scientific evaluation systems primarily reward outputs such as publications, citations, and grant funding. While these remain important indicators of productivity, many of the most valuable contributions to long-term scientific progress function as scientific assets: reproducible findings, shared datasets, software tools, validated methods, trained scientists, collaborative networks, and research infrastructure. Future evaluation systems should recognize both outputs and assets. Outputs measure scientific production, whereas assets measure the creation of enduring scientific capacity that continues generating value beyond a single publication or grant cycle.

Reproducibility and replication should be recognized as distinct scientific contributions rather than ancillary activities. NIH could establish dedicated funding mechanisms specifically for independent replication studies, multi-site validation projects, and methodological reproducibility assessments. Potential indicators include successful independent replication of key findings, adoption of standardized protocols across laboratories, publication of validation datasets, and demonstrated reproducibility of analytical pipelines.

To avoid creating incentives that favor only novelty, NIH could encourage portfolio diversification at both investigator and laboratory levels. For example, laboratories could be encouraged to allocate a portion of their research effort toward validation of influential findings within their field. Review criteria could explicitly recognize replication studies as contributions to scientific infrastructure rather than judging them solely on novelty. This approach would strengthen confidence in biomedical findings while reducing the systemic pressure toward publication of only positive and novel results.

2. Data, Software, and Model Sharing: Data, software, and computational models should be evaluated as scientific products in their own right. Rather than rewarding simple deposition of datasets or code, NIH should prioritize indicators of meaningful reuse. Examples include external dataset citations, incorporation into community benchmarks, reuse by independent laboratories, software downloads accompanied by documented scientific applications, and integration into training or educational resources.

NIH could support dedicated "scientific asset" grants focused on maintaining, curating, and expanding high-value datasets, software platforms, and computational models. Such mechanisms would recognize that the long-term value of many research investments depends on their accessibility and usability

beyond the originating laboratory. Evaluation systems should reward quality, interoperability, documentation, and demonstrated reuse rather than mere availability.

3. Training and Mentorship: Training should be evaluated based on development of scientific independence rather than trainee counts alone. Potential indicators include trainee retention in scientific careers, progression into independent leadership roles, interdisciplinary skill acquisition, and contributions to open science and collaborative research. NIH could support longitudinal tracking of trainee outcomes and recognize laboratories that consistently develop successful scientists across diverse career paths, including academia, industry, government, and healthcare. The goal of mentorship should be viewed not as producing assistants, but as producing future contributors and leaders within the scientific ecosystem.

4. Collaboration: Modern biomedical research increasingly relies on interdisciplinary teams whose contributions are not adequately captured by authorship position alone. NIH should encourage more structured contribution reporting and recognize activities such as development of shared infrastructure, coordination of multi-site studies, integration of expertise across disciplines, and leadership of collaborative networks. Evaluation systems should reward substantive collaborative contributions rather than simple co-authorship counts. Team science should be recognized as a scientific asset that increases the collective capacity of the research enterprise.

5. Entrepreneurship and Translation: Translation should be assessed based on real-world adoption rather than only patents or startup formation. Potential indicators include incorporation into clinical workflows, regulatory milestones, technology licensing, commercial deployment, implementation in healthcare systems, and measurable improvements in patient care. NIH could encourage partnerships between academia, industry, and healthcare organizations while recognizing translational activities as legitimate scientific contributions. The goal is not merely commercialization but successful movement of discoveries into practical use.

6. Foundational Scientific Exploration: High-risk, high-reward science requires evaluation frameworks that tolerate uncertainty and failure. Many transformative advances arise from projects that initially lack obvious translational outcomes or immediate publication opportunities. NIH could support dedicated mechanisms for foundational studies, method development, theoretical frameworks, and exploratory investigations. Evaluation should focus not only on immediate outcomes but also on whether a project generated new questions, methods, conceptual frameworks, or research directions. Scientific failure should not automatically be interpreted as lack of scientific value when rigorous approaches are used to address important questions.

7. Public Impact: Public impact should be assessed across multiple timescales. Near-term indicators may include contributions to clinical practice, health outcomes, public health initiatives, educational resources, and economic activity. Longer-term indicators should include public trust, accessibility of scientific information, community engagement, and development of scientific literacy. Because many societal benefits emerge years or decades after discovery, evaluation systems should recognize both direct impacts and contributions that strengthen the broader scientific ecosystem. Public benefit is often produced through cumulative scientific assets rather than single publications or individual grant cycles.

8. Other Areas: NIH should complement traditional output-based metrics with asset-based metrics. Publications, citations, and grants measure productivity. Reproducible findings, datasets, software,

methods, trainees, collaborative networks, and shared infrastructure measure the creation of enduring scientific capacity. Long-term scientific impact arises from both outputs and assets, and future incentive systems should recognize the complementary roles of each.

49. Stefanie Vogel

I am responding: On behalf of myself

Name: Stefanie Vogel

Name of Organization: University of Maryland Baltimore

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: I recently read a speech by Dr. Battacharya in which he espoused the repeating of published experiments to ensure reliability. I fully concur that reproducing published data is important to ensure that the data are correct when reproduced in a different laboratory. The problem is that NIH grants are based to a large extent on innovation, and if you are proposing to repeat someone else's findings (even in a different way), you will not be considered innovative.

2. Data, Software, and Model Sharing: While all of our primary data from publications are uploaded into OSF, how is NIH going to incentivize and reward these activities unless they have a method of checking them for completeness? As far as I know, there is no mechanism for this to catalyze collaborations or accelerate scientific impact.

With the exception of high-value data sets, and promoting data reuse for other studies, is there a mechanism to check if storing all this primary data is actually benefitting anyone?

3. Training and Mentorship: Evaluate the number of trainees who advance into scientific/science-related professions within a department.
Individuals who oversee or participate in training programs (e.g., T32) should be used to quantify impact.
Evaluate the proportion of graduate students/post-docs who obtain F awards within a department.

It is really not clear to me how NIH "protects time through existing policies and programs." I direct a T32 and the NIH does not cover any of my salary, for example.

4. Collaboration: You should be able to create an "interactome" to illustrate the number of interactions between a specific PI and other scientists.

5. Entrepreneurship and Translation: The STTR and SBIR programs should be greatly expanded to support additional translational studies.
If a R21 or R01 has a translational component that is deemed highly innovative by study section or Program, perhaps this could be used to move them into a more fundable range.

6. Foundational Scientific Exploration: This is what the R21 mechanism was supposed to be -- high risk/high reward without the need for preliminary data; however, this is not what R21s have morphed into.

7. Public Impact: N/A

8. Other Areas: The bottom line is that unless more grants are funded, academic science will be lost for a generation. The new rules proposed by OMB are onerous and will destroy many individual careers as well as institutions.

50. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: This is critical - currently institutions are not rewarding faculty for team science unless they are PI, we are in the era that all science is team science.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: This needs to be elevated. As well as how foundational work was built upon to achieve impact.

8. Other Areas: Academic medical centers need to find another model for funding scientists or all will leave for industry.

51. E. Robert Burns, Ph.D., Professor Emeritus Retired, College of Medicine, Univ. Arkansas for Medical Sciences: burnsbob@uams.edu

I am responding: On behalf of myself

Name: E. Robert Burns, Ph.D., Professor Emeritus Retired, College of Medicine, Univ. Arkansas for Medical Sciences: burnsbob@uams.edu

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: CIRCADIAN RHYTHMS CHARACTERIZE EVERY NORMAL PHYSIOLOGICAL EVENT IN EVERY LIVING THING -SEE 2017 NOBEL PRIZE FOR PHYSIOLOGY OR MEDICINE. NOTHING IN BIOMEDICAL RESEARCH, THEREFORE, IS HOMEOSTATIC. CHRONOBIOLOGICAL INVESTIGATIONS REPEAT THE SAME EXACT EXPERIMENT IN DIFFERENT GROUPS OF ANIMALS OR HUMANS EVERY 3-4 HOURS OVER ONE OR MORE 24 HOUR PERIODS. WITH THIS FREQUENCY OF SAMPLING IT IS NOT UNCOMMON, WHEN ADDING THE DIMENSTION OF TIME TO RESEARCH EFFORTS, TO FIND BODY CLOCK TIMES WHEN THE SAME TREATMENT STIMULATES OR INHIBITS OR HAS NO EFFECT ON AN ACTIVITY COMPARED TO TIME-MATCHED TREATED CONTROLS. PUBLICATIONS USING ONLY USING SINGLE TIME-POINT SAMPLING OVER A DAY – THE CURRENT USUAL PRACTICE CAN REACH VERY DIFFERENT CONCLUSIONS EXPLORING THE SAME STUDY, BUT NOT CONTROLLING FOR THE IMPACT THAT CIRCADIAN RHYTHMICITY HAS ON BOTH THE EXPERIMENTAL DESIGN AND THE DATA INTERPRETATION. THERE IS NO FICKLENESS OR NON-REPRODUCIBILITY WHEN CHRONOBIOLOGY IS PRACTICED, JUST REMARKABLE REPRODUCIBILITY OF RESULTS, EVEN BETWEEN DIFFERENT LABS.

FOR EXAMPLE, THE PERCENT MORTALITY IN MICE CAN RANGE FROM 10% TO 70% FOR ANY DRUG TESTED ACCORDING TO BEST PRACTICES OF CHRONOBIOLOGICAL INVESTIGATION. THE SAME CONCEPT IS TRUE FOR THE LONG-TERM CURE RATE OF TUMOR-BEARING MICE TREATED WITH THE SAME CHEMOTHERAPY DRUGS AND DOSAGES, AND ONLY CHANGING THE BODY CLOCK TIME OF TREATMENT. THE CONCEPT OF CHRONOCHEMOTHERAPY HAS PRODUCED EXCELLENT CLINICAL RESULTS WITH HUMAN CANCER.

NO BIOMEDICAL RESEARCH SHOULD BE DESIGNED AND CONDUCTED WITHOUT AN UNDERSTANDING OF THE ROLE THAT CIRCADIAN RHYTHMICITY PLAYS IN THE EVENT BEING INVESTIGATED. USE BIOLOGICAL TIME AS AN ALLY, NOT A CONFOUNDING FACTOR.

Burns, E.R. 2000 Biological Time and In Vivo Research: A “Field Guide” to Pitfalls. Anatomical Record (New Anatomist) 261:141-152.

Levi, F. 2001 “Circadian Chronotherapy for Human Cancer”, Lancet Oncology 2:307-317.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/LeviLancet.pdf

Description: TWO SUPPORTIVE PUBLICATIONS: Burns, E.R. 2000 Biological Time and In Vivo Research: A "Field Guide" to Pitfalls. Anatomical Record (New Anatomist) 261:141-152. Levi, F. 2001 "Circadian Chronotherapy for Human Cancer", Lancet Oncology 2:307-317.

Circadian chronotherapy for human cancers

Francis Lévi

Cell physiology is regulated by a 24-hour clock, consisting of interconnected molecular loops, involving at least nine genes. The cellular clock is coordinated by the suprachiasmatic nucleus, a hypothalamic pacemaker which also helps the organism to adjust to environmental cycles. This circadian organisation brings about predictable changes in the body's tolerance and tumour responsiveness to anticancer agents, and possibly also for cancer promotion or growth. The clinical relevance of the chronotherapy principle, ie treatment regimens based upon circadian rhythms, has been demonstrated in randomised, multicentre trials. Chronotherapeutic schedules have been used to document the safety and activity of oxaliplatin against metastatic colorectal cancer and have formed the basis for a new approach to the medicosurgical management of this disease, which achieved unprecedented long-term survival. The chronotherapy concept offers further promise for improving current cancer-treatment options, as well as for optimising the development of new anticancer or supportive agents.

Lancet Oncol 2001; **2:** 307–15

Anticancer treatments are generally given at doses near their maximum tolerated dose (MTD), to achieve best efficacy. However, cancer chemotherapy and radiation therapy are toxic to both malignant and normal cells. This lack of specificity leads to the narrow therapeutic index of cytotoxic agents; it also hampers the demonstration of *in vivo* anticancer activity for most biologically targeted molecules currently in development.

Normal cell physiology is characterised by predictable changes over a 24-hour timescale coordinated by the suprachiasmatic nucleus (Figure 1). These circadian rhythms can be used to reduce cytotoxic treatment insults provided that temporal adjustments are made to the administration schedule (chronotherapy). Drugs that target the regulation of cell-cycle events or angiogenesis may also be more effective if administered at specific times, so as to achieve fully *in vivo* the pharmacodynamic effects they display *in vitro*. Preclinical and clinical evidence supports investigations of the chronotherapy hypothesis in cancer patients.

Indeed, the tolerability and the efficacy of chemotherapeutic drugs, as well as of radiotherapy, and cytokines vary by 50% or more as a function of dosing time in mice or rats.[1] The main circadian rhythms of the body, for instance those that modulate secretion of cortisol[2] or melatonin, show nearly normal patterns in groups of patients with advanced or metastatic cancer (Figure 2).[2] This review of the main experimental and clinical prerequisites will be followed by a discussion of the present status and perspectives of chronotherapy in cancer patients.

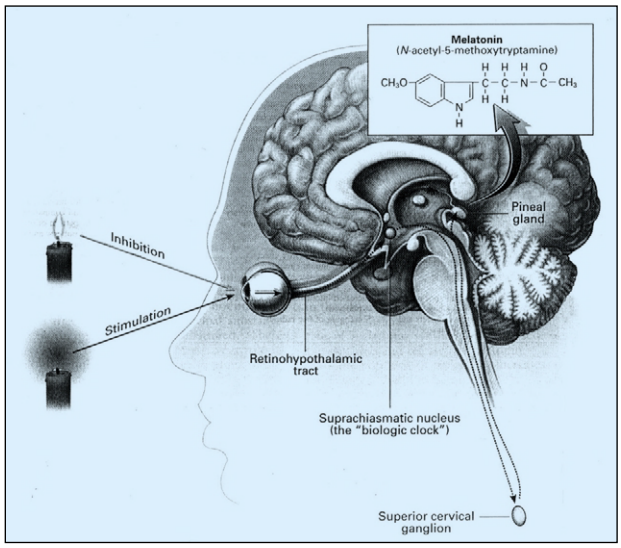


Figure 1. The suprachiasmatic nucleus controls circadian rythyms in response to hormonal signals.

From circadian rhythms to chronotherapy

Circadian rhythms have surprisingly similar properties in cyanobacteria, in plants, and in mammals. They are not merely the reflection of the regular alternation of light and darkness, but persist even in constant environmental conditions, and are therefore endogenous. Interacting molecular loops involving nine genes generate circadian rhythmicity in mammalian cells, and this cellular circadian clock modulates the transcriptional and post-transcriptional processes of several other genes, thereby creating 24-hour changes in cell physiology.[3–5]

The cellular circadian clocks are coordinated by a central

FL is head of the Chronotherapy Unit in the Medical Oncology service at Paul Brousse Hospital, Villejuif, France, Director at the Institut National de la Santé et de la Recherche Medicale, and Chairman of the Chronotherapy Group of the European Organisation for Research and Treatment of Cancer.

Correspondence: Dr Francis Lévi, INSERM E0118 Chronothérapeutiques des Cancers, Institut du Cancer et d'Immunogénétique, Université Paris, Hôpital Paul Brousse, 14 Avenue PV Couturier, 94800 Villejuif, France. Tel: +33 145593855. Fax: +33 145593602. Email: chronbio@club-internet.fr

For personal use. Only reproduce with permission from The Lancet Publishing Group.

one, located at the floor of the hypothalamus, the suprachiasmatic nucleus (SCN), through mechanisms which are yet to be clarified.[6] The SCN generates the circadian rest–activity cycle and coordinates the 'peripheral' oscillators. It also controls the adaptation of the whole circadian time structure to various environmental cycles, known as synchronisers. The regular alternation of light and darkness over a 24-hour period is the main synchroniser of the circadian system (Figure 3). Melatonin, which is mainly secreted by the pineal gland during darkness, contributes to the calibration of the endogenous period to precisely 24 hours.[7] As a result of this synchronisation, mammals with normal circadian function show rhythms in their cellular metabolism and proliferation, with predictable amplitudes and times of peaks and troughs. These rhythms modulate anticancer-drug pharmacology and, ultimately, the tolerability and efficacy of cancer treatments.[1]

Experimental models

Time of administration of anticancer agents influences the extent of toxicity in mice or rats synchronised with an alternation of 12 hours of light and 12 hours of darkness. In most cases, survival varies by a factor of two to eight as a function of dosing time.[1] Although drug pharmacokinetics can vary with time of administration, cellular rhythms appear to be the main determinants of anticancer drug chronopharmacology, because they can modulate the generation or the catabolism of intracellular cytotoxic substances, their interactions with the molecular targets leading to cell dysfunction or death, and the repair of cytotoxic damage (Table 1).[8–14]

A drug usually achieves the best antitumour activity when it is administered at the particular point in the circadian cycle when it is best tolerated; this is true of antimetabolites, alkylators, intercalants, and antimitotic agents (Figure 3). A survival improvement has been achieved in most studies, because drug doses could be safely and selectively increased by 30–50% at the time of best tolerability.[13,15] Nevertheless, in some tumour models, researchers were able to identify a circadian rhythm in anticancer activity, which was coincident with one of drug tolerance, despite the fact that chemotherapy doses were below the MTD.[16–18] The biological and cytokinetic effects that follow the administration of a first drug may influence both the interval and the best time to give a second agent, to achieve optimum therapeutic index in a combination therapy.[14] Fortunately, in many cases, the optimum therapeutic index from combination therapy results from the delivery of each drug near its respective circadian time of best tolerability.[1, 14] Thus, in the experimental model, the delivery of standard doses of chemotherapy at the least toxic time improves tolerability, while anticancer efficacy remains similar or improves, compared with administration every 12 hours. However, in several tumour-bearing animal models, the administration of the MTD at the least toxic circadian time is necessary to improve survival. The application of these strategies to cancer patients would allow the 'best' chronotherapy schedule to enhance patients' quality of life, through a reduction in the incidence of toxic effects, while achieving at least similar antitumour activity, or to prolong survival or to increase cure rate, while producing a tolerability similar to that of standard treatment schedules.

Table 1. Mechanisms of circadian changes in anticancer drug toxicity in rodents

Rhythmic variable	Drug affected
Pharmacokinetics	
(plasma, urine, or cellular)	Mitoxantrone, methotrexate, cisplatin, carboplatin,oxaliplatin, irinotecan
Cellular functions	
Dehydropyrimidine dehydrogenase Orotate phosphoribosyltransferase Uridine phosphorylase Thymidine kinase Thymidylate synthase	Fluoropyrimidines (fluorouracil, floxuridine)
Topoisomerase I	Camptothecins (irinotecan)
Reduced glutathione	Platinum complexes (cisplatin, carboplatin, oxaliplatin) alkylating or intercalating drugs
O^6 alkyguanine methyltransferase	Alkylating agents (nitrosoureas, such as cystemustine)
BCL2, BAX	Taxanes (docetaxel)
Cell cycle regulation	Nearly all agents

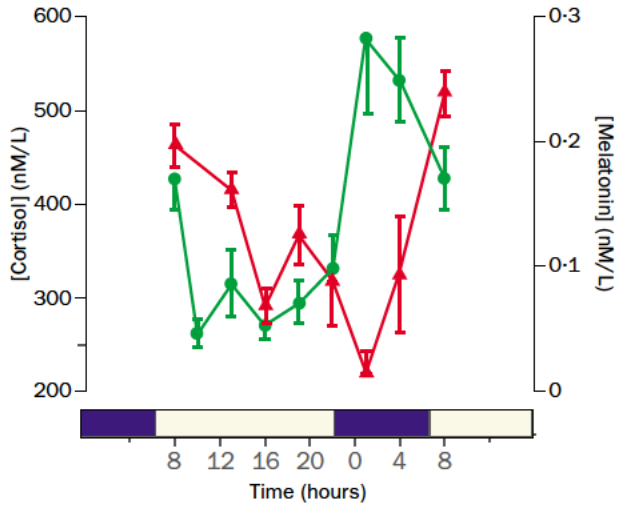


Figure 2. Circadian rhythms in plasma concentrations of cortisol (red) and melatonin (green) in a group of 18 patients with metastatic colorectal cancer. All the patients had received chemotherapy for metastatic disease before the study. Blood samples were obtained every 3–6 h for 2 days in each patient. Light–dark synchronization is shown on the x axis.[2]

Chronomodulated infusions

Rhythms in target tissues

Cells that are engaged in DNA synthesis (S-phase) are generally more susceptible to antimetabolites or intercalating agents. The proportion of S-phase cells in bone marrow, gut, skin, and oral mucosa varies by 50% or more over each 24-hour period in healthy human beings[19] (also reviewed in 1). Similar changes have been found for the

For personal use. Only reproduce with permission from The Lancet Publishing Group.

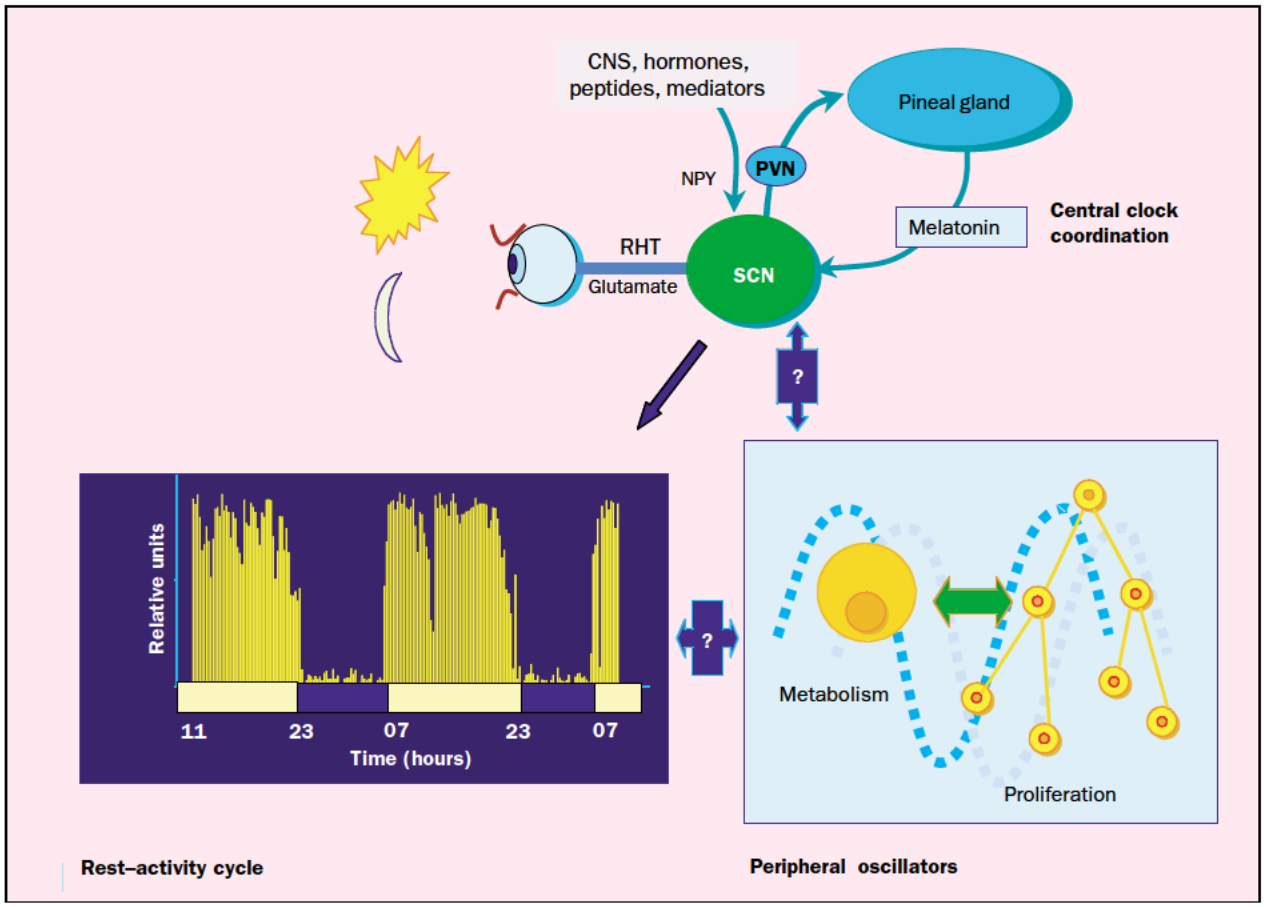


Figure 3. Schematic view of the human circadian system. The SCN is a biological clock located at the floor of the hypothalamus. It displays an approximate 24-hour cycle in the expression of several genes and biochemical functions. Its period (cycle duration) is calibrated by the alternation of light (directly) and darkness (through melatonin secretion by the pineal gland). The SCN generates the rest-activity cycle (left) and coordinates many circadian rhythms in the body, and possibly those that modulate cellular metabolism and proliferation (right). RHT, retinohypothalamic tract; PVN, paraventricular nucleus; NPY, neuropeptide Y.

expression of P53, cyclin E, cyclin A, and cyclin B1 in human oral mucosa.[20] The circadian amplitude of cyclin E was nearly twice as large as that of the other variables, which supports circadian regulation of the G1-S checkpoint. Furthermore, the expression of *PER* and that of *BMAL-1*, two of the main circadian genes, also show 24-hour rhythms in this tissue, indicating that oral mucosal cells are equipped with a molecular circadian clock.[21] The mechanisms through which the cellular circadian clock regulates cell-cycle checkpoints are unknown.

In all these tissues, the mean values for the proportion of S-phase cells are lower between 0000 h and 0400 h, while higher mean values occur between 0800 h and 2000 h.[1] The activity of dehydropyrimidine dehydrogenase (DPD), the rate-limiting enzyme of fluorouracil catabolism, increases by nearly 40% between 2200 h and 0000 h, in the circulating mononuclear cells of both healthy individuals and cancer patients (Figure 4).[22–24] These mechanisms of anticancer-drug chronopharmacology display a similar phase relation to the rest–activity cycle in mice and in human beings, despite the fact that the former are active at night and the latter during the day. During constant-rate infusion of fluorouracil, a circadian rhythm can be detected in the plasma concentrations, both in mice and in cancer patients. The peak concentration of fluorouracil occurs in the early part of the resting period in both species, even if the drug is infused continuously over a week or less.[25–28] On the other hand, patients with very advanced cancer, who are in poor

Table 2. Endogenicity of the rest activity circadian cycle

Property	Comments
Ubiquitous	In flies, rodents, human beings
Endogenous	Persists in constant environmental conditions
Molecular basis	Rhythm suppressed in case of mutation or deletion of: *per* gene in *Drosophila* *clock*, *cry1* or *cry*$_2$ genes in mouse Mammalian genes involved: *Per1*, *Per2*, *Per3*, *Clock*, *Bmal1*, *Cry1*, *Cry2*, *Tim*, *CKI*$_\varepsilon$
Central coordination	Rhythm suppressed if SCN destroyed, restored if in mammals
SCN transplanted (rodents)	
Monitoring (human beings)	Wrist-worn actimetry watch for a minimum of 3 days (sampling of wrist movements with a frequency of 1–10 per min)
Pathology (human beings)	Rhythm alteration in psychiatric or malignant diseases Sleep disorders

For personal use. Only reproduce with permission from The Lancet Publishing Group.

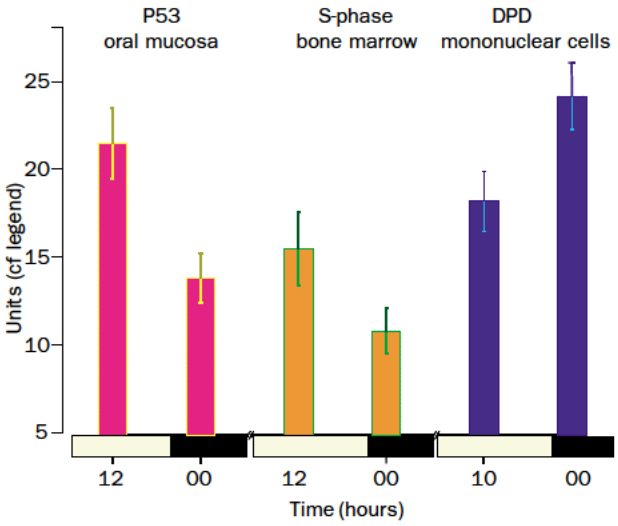


Figure 4. Circadian changes in cellular parameters relevant for chemotherapy tolerability in humans. Oral mucosa P53 in healthy subjects;[20] bone marrow S-phase cells,[19] and mononuclear cell dehydropyrimidine dehydrogenase activity in cancer patients.[23]

general condition, have received prior or concurrent treatments, or are on a low-dose, protracted infusion schedule may show altered DPD or fluorouracil rhythmicity.[22, 29]

The rest-activity circadian cycle

The regular alternation of rest and activity over a 24-hour cycle is one of the most obvious and extensively studied circadian rhythms (Table 2). [4–7, 30] The usual onset of the rest period occurs at the transition from darkness to light in mice or rats and between 2100 h and 2400 h in human beings. These synchroniser clock hours have been used as references for the average optimum times for anticancer drug exposure, because of the physiological coupling between the circadian rest–activity cycle and several chronopharmacology mechanisms, seen in all species. This strategy does not take into account the variability between patients in circadian-rhythm function. Indeed, the demonstration of a circadian rhythm in the hormone production of a single person requires sampling every 10–20 min for 24 h.[31] In other words, the demonstration of individual circadian rhythms in cancer patients is limited to the variables that are amenable to non-invasive monitoring, such as the rest–activity cycle which can be recorded by wrist actimetry. Individual rhythms in mechanistically related variables can be estimated only on data from samples obtained every 2–8 hours for 24–48 hours. With this approach, individual rhythms can be classified as being 'normal' or 'altered'. Rhythm alterations may consist of a decrease in amplitude, modification in peak time or cycle duration, or circadian rhythm suppression.[32]

Programmable multichannel pumps

Two randomised clinical trials, each involving 30 patients with advanced ovarian cancer, have shown that doxorubicin and theprubicin (5–10 min infusion) were better tolerated near 0600 h and cisplatin (1–4 h infusion) between 1600 h and 2000 h than 12 hours apart.[33–34] However, further testing of this treatment method was hampered by the difficulties involved in administering drugs at odd times of the day or night.

Since infusional chemotherapy generally results in better tolerability and antitumour activity than bolus administration,[35] semi-intermittent infusions at a constant rate over 8 or 12 hours have also been investigated, but with limited success.[36,37] The administration of quasi-sinusoidal infusions, with several consecutive steps involving constant-rate infusion at different flow rates over 24 hours, or chronomodulated sinusoidal infusion, a pattern similar to the simplest rhythm model, have become feasible without admission to hospital, through the development of time-programmable pumps.[38–40] This system has mainly been used to test the clinical relevance of chronotherapy. Multichannel devices have facilitated outpatient chronomodulated chemotherapy, with selected peak delivery times for up to four agents over a period of several days. Stability and reservoir compatibility were shown for periods of chronomodulated infusions of floxuridine, interferon-α, fluorouracil, leucovorin, oxaliplatin, and irinotecan at ambient temperature and under normal lighting conditions. Peak delivery was scheduled at 0400 h for fluorouracil and leucovorin, at 1600 h for oxaliplatin, and at 0500 h for irinotecan, based on previous experimental and clinical findings (Figure 5).

Drug pharmacokinetics during chronomodulated infusions

The time course of plasma drug concentrations parallels that of chronomodulated delivery for up to 14 days for fluorouracil,[26, 29] which has a half-life of 10–20 min.[41] This was also true for unbound platinum, during chronomodulated infusion of oxaliplatin, an agent with an initial half-life of 5–30 min and a terminal half-life of one to several days.[42] However, this terminal half-life accounted for plasma accumulation, which resulted in persistent exposure at unappropriate circadian times. Chronomodulated delivery of fluorouracil and leucovorin allowed better control of the plasma concentrations of these drugs if peak flow rates were scheduled at 0400 h than if they were scheduled at 1300 h or 1900 h, or constant-rate infusion. The three latter schedules also had greater toxicity than the first.[23]

These examples show that 24-hour changes in drug plasma concentrations match reasonably well with the chronomodulated delivery waveform, if peak delivery is scheduled at an appropriate time in the circadian cycle. The terminal half-life of the drug accounts for possible drug accumulation in the plasma, and may alter drug exposure patterns, either by shifting peak concentration time or by producing significant concentrations at delivery nadir. There is a need for pharmacokinetic studies on chronomodulated infusions, in order to optimize the alignment of chronotherapy delivery with the targeted circadian exposure pattern.

Chronotherapy for human cancers

The clinical relevance of the chronotherapy principle has been tested in over 1500 patients with gastrointestinal

For personal use. Only reproduce with permission from The Lancet Publishing Group.

disease enrolled in phase I, II, and III clinical trials. Chronomodulated infusions of fluorouracil, generally given with leucovorin, have formed the basis of most chronotherapy schedules developed so far,[24, 43] as a result of the well-established schedule dependency of fluoropyrimidine pharmacology and its toxicity profiles.[41]

MTD of chronomodulated infusions

Patients with gastrointestinal cancers receiving chronomodulated infusions of fluorouracil (alone or with leucovorin), oxaliplatin, or irinotecan at conventional doses showed a low frequency of severe toxic side-effects, in phase I/II trials of 25 to 35 patients each.[24, 38-40, 43] This apparent improvement in tolerability was used to increase the dose and/or to shorten the interval between courses, so as to achieve equivalent toxicity with conventional schedules. Chronomodulated delivery therefore enabled the dose intensity (cumulative dose over a given timespan) and/or the recommended dose of these drugs to be increased up to 100% (Table 3). The antitumour activity of these regimens was assessed near MTD in most trials. This strategy was used early in the course of oxaliplatin development.[24, 39, 42, 44-46] For fluorouracil plus leucovorin, a 4-day chronomodulated infusion of both drugs was administered with a fixed daily dose of 150 mg/m² of leucovorin, and the fluorouracil dose ranged from 900 to 1100 mg/m² daily (nearly twice the conventional MTD). This chronomodulated schedule was given every 2 weeks to 100 patients with previously untreated metastatic colorectal cancer in a phase II trial. The median fluorouracil dose intensity was 1800 mg/m² per week and the objective response rate, as assessed by computed tomography reviewed by an independent panel, was 41% (31.5–50.5 95% CI range). Median survival was 17 months, with 3-year survival of 18·6%.[47] This schedule achieved a 45% objective response rate in a separate Canadian study.[48] These results compare favourably with conventional or intermittent non-chronomodulated treatment schedules for this drug combination,[37] which have produced objective response rates of 10–30% and median survival of 10–14 months in multicentre trials.

Combination chronotherapy schedules

The good tolerability of chronomodulated fluorouracil with or without leucovorin allowed this regimen to be combined with other agents, including oxaliplatin, carboplatin or cisplatin, irinotecan, mitomycin, paclitaxel, mitoxantrone, vinorelbine, or floxuridine into the hepatic artery, or radiation therapy, in order to devise safe and effective outpatient combination regimens against several malignant diseases, without compromising on drug doses.

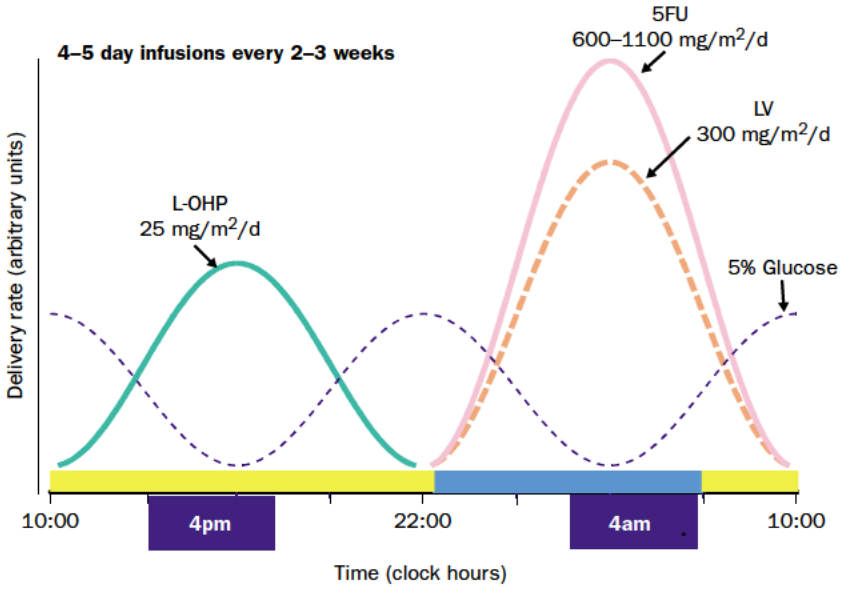


Figure 5. Example of a current chronotherapy schedules delivered to outpatients using a multichannel infusion pump: chronomodulated infusion of fluorouracil (5FU), leucovorin (LV), and oxaliplatin (L-OHP).

The role of oxaliplatin addition to chronomodulated fluorouracil and leucovorin was investigated in a multicentre, open, randomized, phase II/III study.[46] Fifteen centres recruited 200 patients with previously untreated, measurable metastases from colorectal cancer; they were randomly assigned chronomodulated fluorouracil and leucovorin (700 and 300 mg/m² daily, respectively; peak delivery rate at 0400 h) with or without oxaliplatin. Oxaliplatin (125 mg/m²) was given as a 6-hour flat infusion from 1000 hours to 1600 hours on the first day of each course, every 3 weeks. This infusion schedule was devised to remain close to the least toxic time for oxaliplatin. Response, the main judgment criterion, was assessed by means of extramural review of computed tomography scans. Severe gastrointestinal toxic effects (grade 3–4) occurred in 5% of patients given chronomodulated fluorouracil plus leucovorin and 43% of those also receiving oxaliplatin. An objective response was obtained in 16% (95% CI: 9–24%) of patients on fluorouracil plus leucovorin, and 53% (42–63) of those receiving additional oxaliplatin ($p < 0.001$). Median progression-free survival time was 6.1 months (4.1–7.4) and 8.7 months (7.4–9.2), respectively ($p = 0.048$). Median survival times were similar in the two treatment groups (19.9 and 19.4 months, respectively). A possible explanation

Table 3. Maximum tolerated dose of chronomodulated chemotherapy

Drug	Schedule	% Increase in MTD	Reference
Fluorouracil	5d every 21d	40%	65
Fluorouracil–leucovorin	5d every 21d	80%	43
Oxaliplatin	5d every 21d	33%	66
Floxuridine	14d every 20d	45%	38

For personal use. Only reproduce with permission from The Lancet Publishing Group.

was that 57% of the patients from the fluorouracil plus leucovorin group received second-line therapy with the three-drug regimen, which was given as an intensified chronomodulated schedule in most cases.

Validation of chronotherapy in randomised multicentre studies

The importance of chronomodulated delivery rate in cancer therapy was investigated in two consecutive international, multicentre, phase III studies, which compared flat and chronomodulated infusion of fluorouracil plus leucovorin and oxaliplatin in 278 patients with previously untreated metastatic colorectal cancer.[24] In the second study, which involved 186 patients,[44] severe stomatitis was seen in 76% of the patients receiving fixed-rate infusion, and 14% of those on chronotherapy. Cumulative peripheral sensory neuropathy with functional impairment was reported in 31% patients on constant delivery and in 16% patients on chronotherapy (illustrative figures can be found on the website at http://oncology.thelancet.com). The better mucosal status in the chronotherapy group allowed a 40% increase in median dose, a 22% increase in median dose intensity of fluorouracil, and a 30% prolongation of the oxaliplatin treatment span. Objective response rate was 51% for chronotherapy and 29% for constant-rate delivery ($p = 0.003$). According to this multicentre randomised trial, therefore, the most active chronomodulated schedule was also the least toxic one. Median survival was 16 months in both modalities, possibly because 24% of the patients crossed over from the flat schedule to chronotherapy.[44]

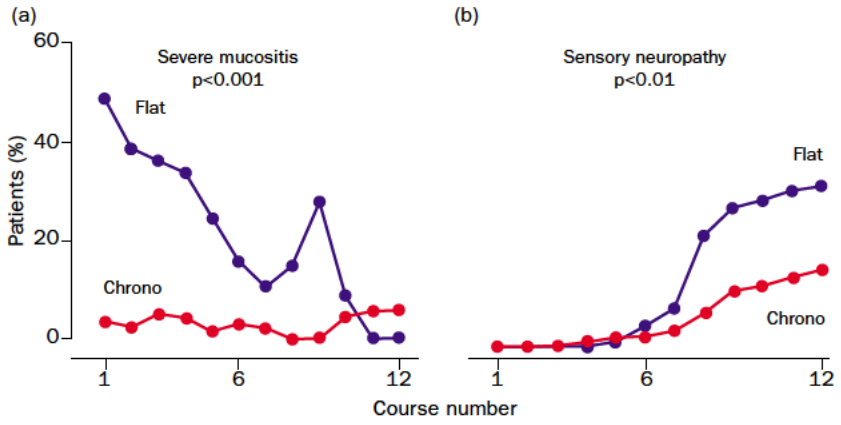


Figure 6. Comparative mucosal (a) and neurologic (b) toxicities of 5-day infusions of fluorouracil, leucovorin and oxaliplatin during treatment, as a function of drug delivery schedule.[44]

Dose intensity and antitumour activity

The good tolerability of a 5-day infusion of chronomodulated fluorouracil plus leucovorin with oxaliplatin allowed an increase of 20% in the 5FU daily dose and a shortening of the interval between courses of 1 week, in two consecutive trials involving a total of 103 patients with previously untreated metastatic colorectal cancer.[45, 49] This resulted in an increase of median dose intensity of 32% for fluorouracil and by 18% for oxaliplatin, compared with the previous 5-day chronomodulated schedule. In this multicentre trial, objective response rate was 66% (95% CI: 57–75) and median survival was 19.4 months (15.2–23.6). A summary of the data obtained by our multicentre study group, from patients receiving first-line chemotherapy for metastatic disease, illustrates the respective roles of chronomodulation and fluorouracil dose intensity on objective response rate.[24]

The achievement of a 50% or greater objective response rate with the three-drug chronomodulated regimen allowed surgical resection of metastases in a substantial proportion of patients.[50] The clinical relevance of this strategy was first demonstrated in a series of 151 patients with initially unresectable colorectal metastases confined to the liver, 80% of whom underwent chronomodulated chemotherapy with fluorouracil, leucovorin and oxaliplatin. Liver surgery with curative intent was attempted in 51% of the 151 patients and was successful in 38% of the total. A complete histological response was documented in four patients. The median overall survival for the 151 patients with liver-only disease was 24 months (95% CI: 19–28) with 28% surviving at 5 years (20–35). The estimated 5-year survival rates were 50 % (38–61) for the 77 operated patients and 3% for the 74 unoperated patients. Therefore, combination of effective and safe chemotherapy with surgery appears to alter the natural history of primary unresectable colorectal cancer metastases.[50] Indeed, surgery of metastases after chemotherapy was the major independent prognostic factor of survival in patients with previously non-resectable metastatic disease. In our multicentre experience, which is not limited to patients with liver-only disease, nearly twice as many patients underwent successful metastases surgery after first-line chronotherapy, as compared with those who received constant-rate infusion (25% versus 13%),[51] and this rate was further increased by intensification of chronotherapy (illustrative figures can be found on the website at http://oncology.thelancet.com).[24]

Circadian rhythm as a prognostic indicator of survival

Several clinical and biological variables influence the natural history of human malignant disease and have been identified as independent prognostic indicators of survival. The combination of several of these factors accounted for median survival times ranging from 3.8 to 21.3 months in patients with metastatic colorectal cancer who did not receive specific therapy.[52] Circadian rhythm alterations were found in patients with breast, ovarian, and colorectal cancer, and were usually associated with increased tumour burden or poor general condition (WHO performance status of 2 or greater).[32] However, two recent studies have shown that circadian rhythm alterations can be encountered in patients of good performance status with colorectal or breast cancer, and constitute an independent prognostic indicator of survival.[53,54]

The relevance of the rest–activity and cortisol circadian

For personal use. Only reproduce with permission from The Lancet Publishing Group.

rhythms to quality of life and survival was prospectively investigated in 200 patients with metastatic colorectal cancer.[53] This assessment took place before the administration of a first course of chronotherapy with fluorouracil, leucovorin and oxaliplatin. Previous chemotherapy or radiotherapy had been given to 59% of the patients. An abnormal cortisol rhythm was defined as a decrease of less than 40% in plasma concentration between 0800 and 1600 h. An abnormal rest–activity cycle was characterised by an autocorrelation coefficient at 24 hours of less than 0.28, an indication of poor reproductibility in the rest–activity pattern from one day to the next. About 30% of the patients had circadian rhythm alterations, by these criteria. Although cortisol-rhythm estimate was unrelated to quality of life or survival, the patients with a marked rest–activity cycle had a better quality of life and longer survival than those with poor rhythmicity. Survival at 4 years was twice as high in the patients with a marked rest–activity cycle, as in those with damped or altered rhythms. Multivariate analysis indicated that the prognostic value of circadian rhythmicity for rest–activity was independent of other well-known prognostic factors such as performance status, number of metastatic organs, or degree of liver replacement by tumour.[53]

The relation between the salivary cortisol cycle and survival was explored in 104 patients with metastatic breast cancer, previously given chemotherapy, radiotherapy, or both; 69% were receiving hormonal therapy at study entry.[54] While 37% of the patients had cortisol concentrations that peaked at 0800 h and declined thereafter, 49% of patients had peaks that occurred later in the day, and 14% had no apparent rhythm. Patients from the two latter groups were considered to have altered rhythms. The 4-year survival was nearly twice as high in the patients with normal cortisol patterns as in those with altered patterns. Furthermore, cortisol diurnal slope remained a highly statistically significant factor in multivariate analysis, indicating that it was an independent predictor of survival (illustrative figures can be found on the website at http://oncology.thelancet.com).

These data indicate that patients with metastatic colorectal or breast cancer may display circadian rhythm abnormalities, which either reflect aggressive disease or contribute to disease aggressiveness, via mechanisms that need to be explored further. Both studies suggest that circadian clock alterations may contribute to the acceleration of cancer growth. Another study has shown that hormonal circadian rhythms were altered in women at high risk of breast cancer, compared with women at low risk.[31, 55] The circadian rhythm abnormalities that result from frequent transmeridian flights could contribute to an increased risk of developing cancer in flight attendants, a hypothesis supported by epidemiological data[56, 57] as well as by results from experimental studies.[58, 59]

Conclusions

Rhythms in cell function and proliferation circadian rhythms account for predictable circadian changes in cancer-treatment tolerability and efficacy. The extrapolation, from mice to human beings, of the least toxic times to administer chemotherapy has been validated in patients with metastatic colorectal cancer in phase III clinical trials. The chronotherapy concept has played an important part in the recognition of the activity of a new drug, oxaliplatin, against colorectal cancer, and has given rise to a new medicosurgical strategy with curative potential in patients with metastatic disease. Thus, the median survival of patients with colorectal cancer metastases who receive chronotherapy has consistently ranged from 16 to 21 months, the longest time reported for this disease in multicentre trials.

Most patients who participated in the control groups of chronotherapy trials received chronotherapy later in the course of their disease, once the main endpoint (ie tumour response) had been reached. The selection of survival as the main endpoint may thus be unrealistic for metastatic disease, since patients should not be denied the benefit of proven active therapeutic modalities. Nevertheless, current multicentre trials by the Chronotherapy Group of the European Organization for Research and Treatment of Cancer (EORTC) are investigating the use of chronotherapy in prolonging the survival of patients with colorectal or pancreatic cancer (EORTC 05962 and 05963). Nevertheless, meta-analyses may be required to demonstrate survival differences between chronotherapy and conventional treatment schedules in patients with metastatic disease, because survival was only weakly correlated with response rate.[60] However, the findings do suggest the need for further investigations of chronotherapy for prolonging survival in the adjuvant setting. Indeed, survival was largely improved with evening rather than morning administration of maintenance chemotherapy in children with acute lymphoblastic leukaemia.[61, 62]

The role of this treatment concept for improving the tolerability of chemotherapeutic drugs such as irinotecan or vinorelbine (EORTC 05971) is being explored in colorectal and breast cancer, respectively. Infusional chrono-chemoradiation is also being carried out in clinical phase II evaluation in patients with primary biliary cancer (EORTC 05991 and Radiation Therapy Oncology Group – RTOG), and the role of radiation timing on mucosal toxicity is being studied in patients with head and neck tumours by the National Cancer Institute of Canada. These trials involve up to 30 centres from ten countries and mostly follow up on interesting data gathered by single institutions.

The results obtained in colorectal cancer further warrant the assessment of chronochemotherapy in patients with breast, lung, or other malignant disease, and in the adjuvant setting. Yet circadian rhythms also modulate cell-cycle proteins, growth factors, coagulation factors, immune functions, and the expression of many genes.

Search strategy and selection criteria

Published data for this review were identified by searches of NCBI PubMed over the past 5 years using 'circadian rhythms' and 'cancer' as keywords and without any restrictions on language of publication. Also, the references of relevant articles from the authors' own database were used.

For personal use. Only reproduce with permission from The Lancet Publishing Group.

Chronomodulated delivery may therefore help to improve the therapeutic index of cytokines and regulators of cell growth or angiogenesis.[63, 64] The genetic origin of circadian rhythmicity may make chronotherapy potentially relevant for gene therapy as well. Finally, the association of circadian system dysfunctions with accelerated cancer growth is puzzling and may indicate a need for specific supportive care, a measure that could improve quality of life, and possibly survival, in these patients.

References

1 Lévi F. Chronopharmacology of anticancer agents. In: Redfern PH, Lemmer B (Eds). *Handbook of Experimental Pharmacology: Physiology and Pharmacology of Biological Rhythms.* Berlin: Springer-Verlag 1997; 299–331.

2 Mormont MC, Hecquet B, Bogdan A, *et al.* Non-invasive estimation of the circadian rhythm in serum cortisol in patients with ovarian or colorectal cancer. *Int J Cancer* 1998; **78:** 421–24.

3 Balsalobre A, Damiola F, Schibler U. A serum shock induces circadian gene expression in mammalian tissue culture cells. *Cell* 1998, **93:** 929–37.

4 Dunlap JC. Molecular bases for circadian clocks. *Cell* 1999; **96:** 271–90.

5 Shearman LP, Sriram S, Weaver DR, *et al.* Interacting molecular loops in the mammalian circadian clock. *Science* 2000; **288:** 1013–19.

6 Moore RY. Entrainment pathways and the functional organization of the circadian system. *Prog Brain Res* 1996; **111:** 103–19.

7 Cassone VM. Melatonin's role in vertebrate circadian rhythms. *Chronobiol Int* 1998; **15:** 457–73.

8 Boughattas N, Lévi F, Fournier C, *et al.* Circadian rhythm in toxicities and tissue uptake of 1,2-diamminocyclohexane (trans-1) oxalatoplatinum (II) in mice. *Cancer Res* 1989; **49:** 3362–68.

9 Ohdo S, Makinosumi T, Ishizaki T, *et al.* Cell cycle-dependent chronotoxicity of irinotecan hydrochloride in mice. *J. Pharmacol Exp Ther* 1997; **283:** 1383–88.

10 Boughattas N, Lévi F, Fournier C, *et al.* Stable circadian mechanisms of toxicity of two platinum analogs (cisplatin and carboplatin) despite repeated dosages in mice. *J Pharm Exp Ther* 1990; **255:** 672–79.

11 Zhang R, Lu Z, Liu T, *et al.* Relationship between circadian-dependent toxicity of 5-fluorodeoxyuridine and circadian rhythms of pyrimidine enzymes: possible relevance to fluoropyrimidine chemotherapy. *Cancer Res* 1993; **53:** 2816–22.

12 Li XM, Metzger G, Filipski E, *et al.* Pharmacologic modulation of reduced glutathione (GSH) circadian rhythms by buthionine sulfoximine (BSO) : relationship with cisplatin (CDDP) toxicity in mice. *Toxicol Appl Pharmacol* 1997; **143:** 281–90.

13 Tampellini M, Filipski E, Liu X-H, *et al.* Circadian rhythm in docetaxel tolerability and efficacy in mice. *Cancer Res.*, 1998, **58:** 3896–04.

14 Focan C. Circadian rhythms and cancer chemotherapy. *Pharmacol Ther* 1995; **67:** 1–52.

15 Filipski E, Amat S, Lemaigre G, *et al.* Relationship between circadian rhythm of vinorelbine toxicity and efficacy in P388-bearing mice. *J Pharm Exp Ther* 1999; **289:** 231–35.

16 Sothern R, Lévi F, Haus E, *et al.* Control of a murine plasmacytoma with doxorubicin-cisplatin: dependence on circadian stage of treatment. *J Natl Cancer Inst* 1989; **81:** 135–45.

17 D'attino RM, Filipski E, Granda TG, *et al.* Irinotecan (CPT-11) and oxaliplatin (l-OHP) synergistic activity at specific circadian times in tumor-bearing mice. *Proc Am Assoc Cancer Res* 2000; **91:** (abstr) 1268.

18 Granda TG, Filipski E, D'Attino RM, *et al.* Experimental chronotherapy of mouse mammary adenocarcinoma MA13/C with docetaxel and doxorubicin as single agents and in combination. *Cancer Res* 2001; **61:** 1996–2001.

19 Smaaland R, Laerum OD, Lote K, *et al.* DNA synthesis in human bonne marrow is circadian stage dependent. *Blood* 1991; **77:** 2603–11.

20 Bjarnason GA, Jordan RCK, Sothern RB. Circadian variation in the expression of cell-cycle proteins in human oral epithelium. *Am J Pathol* 1999; **154:** 613–22.

21 Bjarnason GA, Jordan RCK, Wood PA, *et al.* Circadian expression of clock genes in human oral epithelium and skin: association with cell cycle progression. *Am J Pathol* 2000 (in press).

22 Harris B, Song R, Soong S, Diasio RB. Relationship between dihydropyrimidine dehydrogenase activity and plasma 5-fluorouracil levels: evidence for circadian variation of plasma drug levels in cancer patients receiving 5-fluorouracil by protracted continuous infusion. *Cancer Res* 1990; **50:** 197–201.

23 Langouët AM, Metzger G, Comisso M, *et al.* Plasma concentration of 5-fluorouracil and mononuclear cell dihydropyrimidine dehydrogenase activity in patients treated with different chronomodulated schedules. *Biol Rhythm Res* 1995; **26:** 409 (abstr) 135.

24 Lévi F, Giacchetti S. Chronomodulation of chemotherapy against metastatic colorectal cancer. In: Bleiberg H, Kemeny N, Rougier PH, Wilke HJ (Eds). *Colorectal cancer: a clinical guide to therapy.* London: Martin Dunitz Publishers 2001 (in press).

25 Petit E, Milano G, Lévi F, *et al.* Circadian varying plasma concentration of 5-FU during 5-day continuous venous infusion at constant rate in cancer patients. *Cancer Res* 1988; **48:** 1676–79.

26 Metzger G, Massari C, Etienne MC, *et al.* Spontaneous or imposed circadian changes in plasma concentrations of 5-fluorouracil coadministered with folinic acid and oxaliplatin: relationship with mucosal toxicity in cancer patients. *Clin Pharm Ther* 1994; **56:** 190–201.

27 Bressolle F, Joulia JM, Pinguet F, *et al.* Circadian rhythm of 5-fluorouracil population pharmacokinetics in patients with metastatic colorectal cancer. *Cancer Chemother Pharmacol* 1999; **44:** 295–302.

28 Codacci-Pisanelli G, van der Wilt CL, Pinedo HM, *et al.* Antitumor activity, toxicity and inhibition of thymidylate synthase of prolonged administration of 5-fluorouracil in mice. *Eur J Cancer* 1995; **31A:** 1517–25.

29 Takimoto CH, Yee LK, Venzon DJ, *et al.* High inter- and intrapatient variation in 5-fluorouracil plasma concentrations during a prolonged drug infusion. *Clin Cancer Res* 1999; **5:** 1347–52.

30 King DP, Zaho Y, Sangoram AM, *et al.* Positional cloning of the mouse circadian clock gene. *Cell* 1997; **89:** 641–53.

31 Halberg F, Cornelissen G, Sothern RB, *et al.* International geographic studies of oncological interest on chronobiological variables. In: Kaiser HE (Ed). *Neoplasms – comparative pathology of growth in animals, plants and man.* Baltimore: Williams & Wilkins 1981; 553–96.

32 Mormont MC, Lévi F. ircadian-system alterations during cancer processes: a review. *Int J Cancer* 1997; **70:** 241–47.

33 Hrushesky W. Circadian timing of cancer chemotherapy. *Science* 1985; **228:** 73–75.

34 Lévi F, Benavides M, Chevelle C, *et al.* Therapy of advanced ovarian cancer with 4'-0-tetrahydropyranyl doxorubicin (THP) and cisplatin: a phase II trial with an evaluation of circadian timing and dose intensity. *J Clin Oncol* 1990; **8:** 705–14.

35 Meta-analysis Group in Cancer. Efficacy of intravenous continuous infusion of fluorouracil compared with bolus administration in advanced colorectal cancer. *J Clin Oncol* 1998; **16:** 301–08.

36 Adler S, Lang S, Langenmayer I, *et al.* Chronotherapy with 5-fluorouracil and folinic acid in advanced colorectal carcinoma. Results of a chronopharmacologic phase I trial. *Cancer* 1994; **73:** 2905–12.

37 Falcone A, Allegrini G, Antonuzzo A, *et al.* Infusions of fluorouracil and leucovorin: effects of the timing and semi-intermittency of drug delivery. *Oncology* 1999; **57:** 195–201.

38 Roemeling RV, Hrushesky WM. Circadian patterning of continuous floxuridine infusion reduces toxicity and allows higher dose intensity in patients with widespread cancer. *J Clin Oncol* 1989; **7:** 1710–19.

39 Lévi F, Misset JL, Brienza S, *et al.* A chronopharmacologic phase II clinical trial with 5-fluorouracil, folinic acid and oxaliplatin using an ambulatory multichannel programmable pump. High antitumor effectiveness against metastatic colorectal cancer. *Cancer* 1992; **69:** 893–900.

40 Giacchetti S, Zidani R, Goldwasser F, *et al.* Chronomodulated CPT-11: a pilot study. *Proc Am Soc Clin Oncol* 1999; **18:** 259a (abstr).

41 Diasio RB, Harris BE. Clinical pharmacology of 5-fluorouracil. *Clin Pharmacokinet* 1989; **16:** 215–37.

42 Lévi F, Metzger G, Massari C, Milano G. Oxaliplatin. Pharmacokinetics and chronopharmacological aspects. *Clinical Pharmacokinet* 2000; **36:** 1–21.

43 Garufi C, Lévi F, Aschelter AM, *et al.* A Phase I trial of five day chronomodulated infusion of 5-fluorouracil and l-folinic acid in

For personal use. Only reproduce with permission from The Lancet Publishing Group.

patients with metastatic colorectal cancer. *Eur J Cancer* 1997; **33:** 1566–71

44 Lévi F, Zidani R, Misset JL. for the International Organization for Cancer Chronotherapy: randomized multicentre trial of chronotherapy with oxaliplatin, fluorouracil, and folinic acid in metastatic colorectal cancer. *Lancet* 1997; 350: 681–86.

45 Lévi F, Zidani R, Brienza S, *et al.* for the International Organization for Cancer Chronotherapy: A multicentre evaluation of intensified ambulatory chronomodulated chemotherapy with oxaliplatin, fluorouracil and leucovorin as initial treatment of patients with metastatic colorectal cancer. *Cancer* 1999; **85:** 2532–40.

46 Giacchetti S, Perpoint B, Zidani R, *et al.* International Organization of Cancer: phase III multicenter randomized trial of oxaliplatin added to chronomodulated fluorouracil – leucovorin as first line treatment of metastatic colorectal cancer. *J Clin Oncol* 2000; **18:** 136–47.

47 Curé H, Pezet D, Kwiatkowski F, *et al.* 5-fluorouracil (5-FU) and l-folinic acid (l-FA) pharmacokinetics during chronomodulated (CM) delivery. *International Conference on Chemotherapeutic Strategies for Treatment of Colorectal Cancer.* Amsterdam, February 10–12, 1999.

48 Jolivet J, Létourneau Y. Colorectal cancer chronotherapy: implementation in general hospitals and clinical results. Biological clocks: mechanisms and applications. In: Touitou Y (Ed). *Proc Int Cong Chronobiol.* Amsterdam: Elsevier. International Congress Series 1998; **1152:** 483–490.

49 Bertheault-Cvitkovic F, Jami A, Ithzaki M, *et al.* Biweekly intensified ambulatory chronomodulated chemotherapy with oxaliplatin, 5-fluorouracil and leucovorin in patients with metastatic colorectal cancer. *J Clin Oncol* 1996; **14:** 2950–58.

50 Giacchetti S, Itzhaki M, Gruia G, *et al.* Long term survival of patients with unresectable colorectal liver metastases following infusional chemotherapy with 5-fluorouracil, folinic acid, oxaliplatin and surgery. *Ann Oncol* 1999; **10:** 1–7.

51 Lévi F, Zidani R, Llory JF, *et al.* International Organization for Cancer Chronotherapy: final efficacy update at 7 years of flat *vs* chronomodulated infusion (chrono) of oxaliplatin, 5-fluorouracil and leucovorin as first line treatment of metastatic colorectal cancer. *Proc Am Soc Clin Oncol* 2000; **19:** 242a (abstr 936).

52 Stangl R, Altendorf-Hofmann A, Charnley RM, Scheele J. Factors influencing the natural history of colorectal liver metastases. *Lancet* 1994; **343:** 1405–10.

53 Mormont MC, Waterhouse J, Bleuzen P, *et al.* Marked 24-h rest-activity rhythms are associated with better quality of life, better response and longer survival in patients with metastatic colorectal cancer and good performance status. *Clin Cancer Res* 2000; **6:** 3038–45.

54 Sephton SE, Sapolsky RM, Kraemer HC, Spiegel D. Diurnal cortisol rhythm as a predictor of breast cancer survival. *J Natl Cancer Inst* 2000; **92:** 994–1000.

55 Ticher A, Haus E, Ron IG, *et al.* The pattern of hormonal circadian time structure (acrophase) as an assessor of bresat-cancer risk. *Int J Cancer* 1996; **65:** 591–93.

56 Rafnsson V, Hrafnkelsson J, Tulinius H. Incidence of cancer among commercial airline pilots. *Occup Environ Med* 2000; **57:** 175–79.

57 Ballard T, Lagorio S, De Angelis G, Verdecchia A. Cancer incidence and mortality among flight personnel: a meta-analysis. *Aviat Space Environ Med* 2000; **71:** 216–24.

58 Shah PN, Mhatre ML, Kothari L. Effect of melatonin on mammary carcinogenesis in intact and pinealectomized rats in varying photoperiods. *Cancer Res* 1984; **44:** 3403–07.

59 van den Heiligenberg S, Deprés-Brummer P, Barbason H, *et al.* Promoting effect of constant light exposure on diethylnitrosamine-induced hepatocarcinogenesis in rats. *Life Sciences* 1999; **64:** 2523–34.

60 Buyse M, Thirion P, Carlson RW, *et al.* Meta-Analysis Group in Cancer. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. *Lancet* 2000; **356:** 373–78.

61 Rivard G, Infante-Rivard C, Hoyeux C, Champagne J. Maintenance chemotherapy for childhood acute lymphoblastic leukemia: better in the evening. *Lancet* 1985; **2:** 1264–66.

62 Schmiegelow K, Glomstein A, Kristinsson J, *et al.* Impact of morning versus evening schedule for oral methotrexate and 6-mercaptopurine on relapse risk for children with acute lymphoblastic leukemia. *J Pediatr Hematol Oncol* 1997; **19:** 102–09.

63 Bjarnason GA, Jordan R. Circadian variation of cell proliferation and cell protein expression in man: clinical implications. In: Meijer L, Jézéquel A, Ducommun B (Eds). *Progress in cell cycle research* (vol 4). New York: Kluwer Academic Publishers 2000; 193–206.

64 Lévi F, Bourin P, Deprés-Brummer P, Adam R. Chronobiology of the immune system. Implication for the delivery of therapeutic agents. *Clin Immunother* 1994; **2:** 53–64.

65 Lévi F, Soussan A, Adam R, *et al.* A phase I–II trial of a five-day continuous intravenous infusion of 5-fluorouracil delivered at circadian rhythm modulated rate in patients with metastatic colorectal cancer. *J Infus Chemother* 1995; **5:** 153–58.

66 Caussanel JP, Levi F, Brienza S, *et al.* Phase I trial of 5-day continuous venous infusion of oxaliplatin at circadian rhythm-modulated rate compared with constant rate. *J Natl Cancer Inst* 1990; **82:** 1046–50.

Extra information, mentioned in the text, appears on The Lancet Oncology's website: http://oncology.the lancet.com

For personal use. Only reproduce with permission from The Lancet Publishing Group.

52. ROBERT LAIRD JUDSON-TORRES

I am responding: On behalf of myself

Name: ROBERT LAIRD JUDSON-TORRES

Name of Organization: Huntsman Cancer Institute at the University of Utah

Type of Organization: Healthcare Institution

Role: Investigator/Researcher

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: SUMMARY

The single most effective way to improve the rigor of American biomedical science is to stop rewarding investigators for where they publish and start rewarding them for whether their findings hold up. Today, NIH effectively outsources the definition of scientific merit to commercial journals whose business model rewards novelty, citations, and attention — none of which is reproducibility, and several of which actively work against it. Career advancement and future funding hinge on publishing in a small set of high-prestige venues; those venues, in turn, are incentivized toward splash over durability. The result is a system that selects, generation over generation, for productivity at the expense of reliability. This is not a problem of a few bad actors. It is a problem of a misaligned social contract, and it can be fixed by changing what the enterprise measures and who controls the measurement.

NIH is uniquely positioned to fix it, because NIH funds the largest share of biomedical research in the world and already owns the necessary infrastructure (PubMed Central). The proposal below is a public, rigor-gated system for reviewing and recording NIH-funded findings, with a credit currency that rewards reliability rather than prestige. It can be piloted on existing infrastructure and scaled deliberately.

A single principle organizes the recommendations that follow: innovation should be rewarded when an idea is funded; novelty should never be rewarded when a result is evaluated. The worth of a finding is whether it is true and useful, not whether it was first.

RIGOR AND REPRODUCIBILITY

The core recommendation: relocate peer review out of commercial publishers and into a public, rigor-gated review-and-record system, required for work supported by NIH funds.

The current model asks one or a few overburdened, anonymous reviewers to assess an entire multi-disciplinary study — often with dozens of distinct assays — against a journal's appetite for novelty. No single reviewer is competent across all of it; brand name substitutes for verification; and the verdict measures perceived importance rather than reliability. We propose a two-tier structure that decouples those functions:

1) A standing review panel, modeled on a study section: a small group of subject-matter experts who know one another and must defend their judgments to colleagues, not under cover of anonymity. This panel does not itself certify individual experiments.

2) Modular, registry-matched expert review of individual assays: When NIH investigators apply for funding, they register the specific methods and assays in which they hold expertise. When a study is submitted, the standing panel identifies the assays requiring specialized scrutiny and routes each to qualified experts drawn from that registry. These assay-level reviews are anonymous to the author and double-blind, and each assay is certified satisfactory or not on the basis of rigor alone. Crucially, certification is conditioned on access to deposited, runnable data, code, and unprocessed images (see Data, Software, and Model Sharing), so that "satisfactory" means an expert actually checked the evidence rather than the figure. Participation in assay-level review could be made mandatory for good standing in progress reports — the effort to evaluate individual assays should be significantly less than the current effort to review an entire manuscript.

Thus, this division reduces per-reviewer burden (experts assess only what they are expert in), removes brand and perceived-novelty from the rigor judgment entirely, and places accountability where it belongs at each level.

The competence registry.
Two features of this registry deserve emphasis, because the rest of the proposal draws on them. First, reviewer identities are tracked by the system even when blinded to the author. This permits a reviewer's own record to be assessed over time — specifically, whether the assays they certified continue to reproduce — and it is the basis for the portable rigor credential described under Training and Mentorship. Second, because reviewing is not itself a funded activity, the registry can be extended to qualified non-NIH-funded and international experts. This widens access to scarce methodological expertise and, importantly, dilutes the conflicts of interest that arise in small subfields, where the few qualified reviewers and the authors they assess inevitably overlap. Together these features convert peer review from invisible, unrewarded, anonymous labor into a tracked, creditable, and accountable contribution.

Classification, not ranking.
Once every component is certified, the standing panel renders one further judgment — does the overall claim follow from the certified evidence? — and then classifies the study by what it is: reproduction (successful or failed) of prior work, method or assay development, target discovery, mechanistic advance, and so on. This classification step is the natural home for what the RFI calls defining causal research: the panel's explicit task is to assess whether the certified bricks (assays) actually support the building (the conclusions of the full study), including whether a causal claim is warranted by the design. The system records what a study established; it does not anoint a winner. Importantly, it does not assign some concept of 'Impact' at the time the study is published. Impact, by definition, must be assessed longitudinally, and be looked at by reproducibility (captured by the proposed system), citations of the study (as opposed to "Impact Factor" of the journal, and links to emerging technologies, clinical tests, or clinical trials.

Defining causal research and confirmatory claims.
For claims sold as confirmatory, replicative, or near-translational, a registered-report track — in-principle acceptance based on question and methods before results are known — should be available and, where stakes warrant, required. This destroys publication bias and analytic flexibility at the root for precisely the claims where reliability matters most, similar to what is already required of clinical trials. Exploratory and discovery science must remain free to pivot; it should simply be labeled exploratory so it cannot later masquerade as a definitive test. (An NIH grant is, in effect, already a pre-registered protocol; this track formalizes that logic at the reporting stage.)

2. Data, Software, and Model Sharing: In the afore described system, make deposit the precondition for review eligibility, not an optional supplement. A panel cannot be certified unless the underlying data, analysis code, and raw images are deposited and runnable. This reframes sharing not as a compliance burden but as the thing that makes rigorous review possible — it eases the honest path rather than policing the dishonest one. As a measurable indicator, the fraction of an investigator's certified panels backed by independently re-executable data and code is a direct, gameable-resistant signal of rigor that current metrics cannot capture.

3. Training and Mentorship: The competence registry described above doubles as a training and recognition instrument. Performing structured, accountable assay-level review is itself rigorous training; trainees learn what "satisfactory" actually requires by doing it, under the standing panel's oversight. More importantly, a reviewer's record — and specifically whether the panels they certified continue to reproduce over time — becomes a portable, verifiable credential of judgment and rigor. This recognizes a form of scientific labor that is presently invisible and unrewarded, and it gives mentorship and rigor-training a measurable footprint: a mentor's contribution can be assessed in part by the certified, reproducing output of those they train.

4. Collaboration: The indivisibility of authorship is a central pathology of large-team science: on a fifteen-author paper with hundreds of panels, no one is accountable for any specific result, and supervisors cannot — and demonstrably do not — verify all of it. The system above produces accountability as a byproduct of review. Each certified data panel is tied to the authors who stand behind it; supervising investigators stand behind every panel under their direction. This yields a granular, contribution-level record of who is accountable for what — a far more honest substitute for authorship-order conventions, and a concrete mechanism for recognizing genuine team contributions without diluting accountability into anonymity. Recognition scales with verified contribution rather than with position in an author list.

5. Entrepreneurship and Translation: Translation is where reliability carries the highest stakes and where irreproducibility is most expensive. The registered-report track and mandatory data and code deposit are most valuable here: a public, rigor-certified record of what has actually been established accelerates legitimate translation and reuse, while denying downstream developers the false starts that a hype-optimized literature currently supplies. A trustworthy record is an accelerant for entrepreneurship, not a brake on it — capital and effort flow more efficiently when the underlying findings have been verified rather than merely published.

6. Foundational Scientific Exploration: The reform protects high-risk, high-reward work rather than threatening it, provided two safeguards are built in. First, the innovation/novelty distinction: bold,

unprecedented ideas should continue to be rewarded at the funding stage (innovation is a legitimate criterion for backing an idea), while the evaluation of results is stripped of any premium for mere priority. In other words, great solution's value is its greatness, not its novelty and a useless finding is not redeemed by being first. Second, a rigor-rate metric must not punish ambition: an investigator attempting one of the few assays only a handful of labs in the world can perform must not be penalized for a difficult initial verdict. Only the final, post-revision certification should count toward an investigator's record, so that a hard first attempt is a draft, not a permanent scar. Reproducible discovery and bold exploration are then complementary, not competing.

7. Public Impact: This is, fundamentally, a public-trust proposal. The scientific record produced with public funds should be public infrastructure, governed for reliability and accessible to the citizens who paid for it — not rented back from commercial intermediaries whose incentives diverge from the public interest. A system that certifies whether findings hold up, and that records reproduction and replication as first-class outcomes, speaks directly to the erosion of public confidence in science. The most durable measure of impact is not the "impact factor" of the journal is was published in, but whether others can build on a result — and that is precisely what this system would, for the first time, measure directly.

WHAT HAPPENS TO THE JOURNALS?

Commercial journals are not abolished, and they are not the enemy of this proposal — they are simply moved out of a role they were never well suited to. Under this model they continue to do what magazines do well: sell subscriptions and advertising in exchange for curating content. What changes is that they exit the business of gatekeeping peer review and enter the business of surfacing the most interesting results from the public record and explaining them — to the general public, to scientists outside the field, and to clinicians — each audience served by the kind of clear, compelling writing that journals already excel at when freed to do it.

This is a net gain for science on its own terms. There is a well-documented gap between what biomedical research establishes and what the public understands, and that gap is corrosive to the public trust the NIH is explicitly trying to rebuild. A communication-and-curation layer that competes on the quality of its explanation, sitting atop a record that has already been certified for rigor, is far better positioned to close that gap than the current arrangement, in which the same outlets are responsible for both judging science and selling it.

It is worth being candid about the one party that loses. Today the public pays three times over: taxpayers fund the research, then pay again — through institutional subscriptions and publication charges — to put it in print, then pay a third time to read it. Removing peer review and the scientific record from commercial control ends that arrangement. Journals would earn less from taxpayer-subsidized labor and would have to compete harder, on the merits of their communication, for the revenue they keep. That is not a cost to science or to the public; it is the correct alignment of incentives finally reaching the last actor in the chain.

8. Other Areas: SUMMARY OF MEASURABLE INDICATORS OF PROPOSED SYSTEM

1) Rigor rate: the percentage of an investigator's panels that achieve final certification — a direct

measure of reliability that resists padding, because low-quality panels lower the rate.

2) Durability / build-on rate: credit accrues to a panel when others successfully reproduce or build upon it — an impact measure that rewards being right over being first, and that operationalizes the value of doubling down on solid prior observations.

3) Reproducibility-backed deposit rate: the fraction of certified panels supported by independently re-executable data and code.

4) Reviewer reliability: whether the panels a reviewer certified continue to reproduce over time.

These two axes — a rigor rate and a durability rate — deliberately separate the two things the RFI asks NIH to measure: reproducibility discipline and impact. Collapsing them into one number is what produced the current generation of gameable metrics.

UNINTENDED CONSEQUENCES AND FEASIBILITY

A serious proposal must name its own failure modes.

Conflict of interest in small fields. In assays only a few labs perform, reviewers and authors overlap, compete, and co-fund. The mitigations are structural and largely built into the architecture above: assay-level review is double-blind to the author; the reviewer pool extends beyond NIH-funded investigators (see The competence registry, above), diluting conflicts in small subfields; and the standing panel — which is not anonymous to itself — must defend its classifications, deterring capture.

Metric gaming (Goodhart's Law). Any single metric will be gamed. The two-axis design is the defense: the rigor rate punishes panel-proliferation, and durability weighting makes commodity-assay farming worthless, because easy panels are not the ones others build upon.

Fraud is out of scope. This system repairs the incentives that drive otherwise-honest investigators toward poor practice; it is not a fraud-detection apparatus. Deliberate fabrication requires separate detection and enforcement and should be addressed as such. Stating this boundary matters: the goal of this proposal is to repair a broken social contract for the honest majority, not to assume bad faith.

The transition cost and career-stage feasibility. If the proposed system, or any new system is adopted, for an interval, NIH-funded investigators would operate under a standard the rest of the world has not yet adopted. The investigators most exposed are likely also our most vulnerable: post-docs and early-stage investigators at conventional research universities whose promotion-and-tenure committees still weight prestige-journal placement — a domain NIH does not directly control. A related risk is that well-resourced laboratories able to fund themselves through other means decline NIH support to keep pursuing prestige, inverting the intended hierarchy. Mitigations: phase the mandate; condition institutional support on adoption of responsible, venue-independent assessment in promotion (NIH's funding relationship is the lever); and pilot on PubMed Central before scaling. The forcing function is

NIH's own scale — as with public-access policy, an NIH mandate can move the global norm.

Governance and durability. A publicly owned record must not become a politically controlled one. Editorial and certification authority should be vested in a statutorily independent, firewalled entity — chartered for operational independence in the manner of the National Academies, with scientific-society co-governance to import the editorial culture of the most rigorous existing journals. The system must be designed to remain independent of whoever holds power; that durability across administrations is not a vulnerability to be hidden but the central feature that distinguishes a public record from both political and commercial capture.

53. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Member of the Public

Date Submitted: 6/24/2026

1. Rigor and Reproducibility: Incentives to skew scientific findings to make them "look" better than they are (to get into a fancy journal) will never be fixed until metrics focus on the process of doing a good experiment/study and completely separates it from the particular outcome of the study. Reward should be placed on rigorous experimental design (blinding, randomization, sample size estimation, proper positive and negative controls, transparent reporting of data exclusions and analysis plans), on asking an important question that can be answered by the experiment as designed, and on the replication and expansion of collective knowledge in the area by other researchers. The outcome of any single experiment should be irrelevant to crediting the scientist as a "good" scientist. Metrics should extract and consider important experimental design elements within protocols that reduce experimental bias and facilitate the probing of causality, and these metrics should proactively block consideration of the outcome of the experiment. (Outcomes would be more relevant to an overall long-term "impact" metric but should still only be considered in the context of the body of scientific evidence and should not rely on a single experiment or study.) Journal impact factor, journal reputation, citation counts, and other superficial "popularity" metrics are completely irrelevant to the rigor of a study and are, themselves, prone to bias and skewing in a way that does not even correlate with rigorous experimental design.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: In order to promote high-risk, high-reward research, publication counts and other "productivity measures" need to be retired. Similar to the rigor question above, the focus of any metric should be on the design of the study and how likely it is to provide a valid answer to an interesting question, regardless of what that answer happens to be. The process of doing good science should be assessed and not the outcomes of particular lines of research, publications, or individual experiments. Was the protocol rigorously planned? Did the experiments blind, randomize, ensure sufficient sample size, have proper controls, remove subjectivity to the extent possible? Was the design of the experiment actually capable of providing an answer to the specific question posed?

7. Public Impact: The issue with providing any specific metric of public impact (such as contributions to patents or medical guidelines) is that this presumes a single experiment or the work of a single experimenter can or should be linked directly to these outcomes. Science is a collective body of evidence, with multiple lines of research asking orthogonal questions with different experimental designs to build a full context of our knowledge. Single experiments are, by definition, not worth much without this broader context. It is antithetical to the goals of promoting high-quality individual studies that are rigorous and carefully designed so that they properly contribute to the collective knowledgebase to then reward people most highly for producing high-profile, individually recognizable claims of knowledge. That way lies perverse incentives. Rather, models should build collective knowledge by integrating all evidence available, but it should not then be credited to any specific single individuals or studies. Popularity, visibility, and fame do not measure quality. Rather, NIH should instead focus on using these collective knowledgebases to provide a larger story of how NIH-funded research contributes overall to these larger impacts.

8. Other Areas: N/A

54. Edwin S. Monuki

I am responding: On behalf of myself

Name: Edwin S. Monuki

Name of Organization: University of California Irvine

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/25/2026

1. Rigor and Reproducibility: Rather than tactical, the fundamental problem is the disincentivization of R&R throughout research ecosystems, which impacts human behavior at all levels of these ecosystems. Promotion and tenure, private funding, publication, and even training are based principally on impact and productivity. In general, this comes at the expense of rigor and reproducibility. Until there is meaningful cultural movement to combat the disincentives for R&R at multiple levels within these ecosystems, most useful tactics will die on the vine. Actually, in my view, NIH/CSR has been the one meaningful breath of fresh air in this space. NIH/CSR cannot tackle this problem alone, but I do applaud the persistent effort.

For institutions, incentives and infrastructure to support R&R of existing findings could make a difference, but would not address root causes. Just like clinical trials, it is impossible and impractical to have "gold standard" evidence for everything.

Funding that directly (or indirectly) motivates institutions and journals to elevate R&R in their review processes could help, although this may not be within scope for NIH/CSR. Somehow taking advantage of the increasingly-popular preprint mechanisms to modify journal behavior strikes me as an opportunity.

The R&R problem only gets bigger with the promotion of multidisciplinary "large team" science. Thus, another strategy to consider would be shifting project funding back towards small team science, where accountability for R&R is clearer and impact is more disruptive than incremental (Wu et al., Nature 2019; PMID: 30760923).

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: As with grant review, I also applaud NIH/CSR efforts to institutionalize Responsible Conduct in Research (RCR) training, and there is some overlap in the Venn diagram of RCR with R&R trainings. However, in reality, these only partially overlap. Specific training requirements for R&R would be one concrete suggestion.

4. Collaboration: Is the assessment and recognition piece something that NIH/CSR needs to actually worry about? This sound more like an academia problem.

5. Entrepreneurship and Translation: My only comment is that this area, too, comes at the potential cost and diminution of R&R incentives at the front end of the research process. Attention to R&R in this process will be key to avoiding wastage of taxpayer dollars and setting back research fields.

6. Foundational Scientific Exploration: I fully agree. See above comment re: Wu et al., Nature 2019 (PMID: 30760923). Journals and reviewers naturally hold authors to a higher bar for the "burden of proof" for disruptive ideas, which then naturally aligns with R&R incentives. NIH shifting more resources towards rigorous high-risk, high-reward R01s (and not just R21s) seems like a great idea.

7. Public Impact: Come to think of it, this category is also a disincentive to R&R and comes with substantial overpromising risk. Also, while communication with the public is critical, effective communication is hard and public subset-dependent. For NIH/CSR, it would seem that the most relevant subset are the lawmakers. Whatever approaches at lower levels also align with helping NIH message better to the lawmakers would make the most sense. And seems budgetarily feasible.

8. Other Areas: N/A

55. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/25/2026

1. Rigor and Reproducibility: you could look at researchers that have had papers retracted and have harsher penalties for having work retracted if it was NIH funded. There are certain journals that are also more likely, statistically, to have retracted papers within them - this could be looked at.

2. Data, Software, and Model Sharing: NIH has already put in rules to make sure that researchers share their data and tools with others if the work has been funded by NIH research. this isn't always followed. My team and I have come up against several scenarios where we write to corresonding authors for data and tool request and often they don't even email back. there should be something in place that's stronger to make sure PIs share their work.

3. Training and Mentorship: Too often reviewers review training grants or fellowships with an eye toward how many students go on to academic positions . this is an antiquated view, and instead we should be assessing whether graduates are contributing in any way to further knowledge or to the US economy by obtaining their career of choice with strong training.

4. Collaboration: n/a

5. Entrepreneurship and Translation: having a place specifically for patents could demonstrate translation of NIH funded work.

6. Foundational Scientific Exploration: You could weigh innovation of proposals heavier. right now that's buried within significance in the scoring criteria. More specific calls for high risk proposals, such as those funded under the common fund, could also help here.

7. Public Impact: N/A

8. Other Areas: n/a

56. Bycelium

I am responding: On behalf of an organization

Name: Mathïs Fédérico

Name of Organization: Bycelium

Type of Organization: Other

Type of Organization - Other: Initiative / Startup / Wannabe nonprofit

Role: Other

Role - Other: Founder

Date Submitted: 6/25/2026

1. Rigor and Reproducibility: The credibility shift of hypotheses based on evidence, aggregated and compared to follow ups on the same claims.

This comes from Bayesian epistemology.

See https://bycelium.com for more details, or contact me at mathis.federico@bycelium.com

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: The credibility shift of hypotheses based on evidence, aggregated by endorsement of experts, weighted by expertise.

This comes from Bayesian epistemology.

See https://bycelium.com for more details, or contact me at mathis.federico@bycelium.com

7. Public Impact: The credibility shift of hypotheses based on evidence, but gathered from endorsement of the public.

This comes from Bayesian epistemology.

See https://bycelium.com for more details, or contact me at mathis.federico@bycelium.com

8. Other Areas: N/A

57. Ji-Hyun Lee

I am responding: On behalf of myself

Name: Ji-Hyun Lee

Name of Organization: University of Florida

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/26/2026

1. Rigor and Reproducibility: One issue I would like NIH to consider is what happens to study size after a grant is funded.

When investigators submit an NIH proposal, they spend a great deal of time justifying the sample size. The study is designed to detect a clinically meaningful effect with adequate statistical power, and reviewers carefully evaluate those assumptions. The proposed sample size is one of the scientific foundations of the grant.

However, after funding, the reality is often very different. Across many NIH-funded studies that I have collaborated on as a biostatistician, it is actually uncommon for a study to reach its planned enrollment. Recruitment is slower than expected, budgets change, and investigators proceed with a substantially smaller study.

Some reductions are unavoidable. That is not the issue.

The issue is that this has become an accepted part of the research culture. Investigators expect it, reviewers know it happens, and NIH has largely tolerated it as long as the study is completed. As a result, we have become comfortable proposing one study during peer review and conducting another after funding.

I believe this culture should change. If a study undergoes a major reduction in sample size, investigators should transparently justify the change and demonstrate that the revised study can still answer the original scientific question. Scientific success should not be measured only by publications. It should also reflect whether the funded study was conducted with the rigor that justified the award.

2. Data, Software, and Model Sharing: In my experience, I have rarely seen data sharing or software sharing become as valuable as NIH envisions. The reason is not a lack of willingness to share. Rather, high-quality sharing requires a substantial amount of additional work that is often overlooked.

Preparing a dataset for others to use is much more than uploading a file. The data must be cleaned, documented, quality checked, and accompanied by reproducible statistical code. Otherwise, other investigators cannot reproduce the findings or confidently reuse the data.

In many collaborative projects, this work falls largely on statisticians. Yet this effort is rarely reflected in grant budgets, scientific recognition, or measures of research productivity. If NIH truly wants to promote reproducibility through data sharing, it should explicitly recognize the statistical effort required to make shared data scientifically useful, not simply publicly available.

3. Training and Mentorship: None.

4. Collaboration: NIH has been promoting team science and interdisciplinary collaboration for many years. However, I do not think the academic reward system has changed accordingly.

Take a look at tenure and promotion criteria at most universities. They still place the greatest value on being the PI of a grant and the senior or corresponding author of publications. Collaborative contributions are often viewed as secondary, even when they are essential to the success of the research.

This creates a mismatch between what NIH encourages and what academic institutions reward. Why should faculty spend substantial time on collaboration if those efforts are not valued in their careers—or may even work against promotion?

If NIH wants to promote team science, collaboration should be recognized and rewarded not only in NIH funding, but also in the academic evaluation system.

5. Entrepreneurship and Translation: None.

6. Foundational Scientific Exploration: I strongly support basic research and high-risk, high-reward science. However, I wonder whether the current research environment places too much emphasis on the number of funded projects and publications rather than the scientific value they generate.

Biomedical research has grown tremendously over the past several decades. Yet it is not clear that a larger number of grants and publications has always translated into a proportional increase in meaningful scientific discoveries or improvements in human health.

NIH should consider placing greater emphasis on the quality, rigor, and long-term impact of research rather than the quantity of funded projects or published papers. This also requires rewarding the practices that make research more valuable, including conducting studies as originally designed, sharing high-quality and reproducible data and software, and recognizing meaningful contributions to team science. Scientific success should be measured by the knowledge generated and its contribution to society, not simply by research output.

7. Public Impact: None.

8. Other Areas: None.

Description: The comments above reflect my personal views and experiences and are submitted in my individual capacity. They do not represent the views of the University of Florida or any other professional organization with which I am affiliated.

58. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/26/2026

1. Rigor and Reproducibility: Rigorous research should include redundant experimental design, sound statistical analysis, recognition of error, avoidance of logical fallacies, and intellectual honesty (Casadevall and Fang, mBio, 2016). Reproducibility is challenging to assess, but can be facilitated by open science solutions including availability of raw datasets, code and methods.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: GENERAL COMMENT

I am an academic clinician-scientist with 40 years of basic and translational research experience. My professional activities have included more than twenty years as a journal editor, and I have served on many scientific review panels. Therefore, I am very familiar with the challenges in measuring scientific impact. One of the most important limitations is that impact can only be assessed in retrospect, and the amount of time required to fully appreciate a scientific discovery is often in direct proportion to its importance. The ultimate importance of a scientific contribution can be measured by the size of the problem, its practical utility, its integration with other knowledge, and its novelty (Casadevall and Fang, Infect Immun, 2009).

My recommendation is that the NIH focus on research quality, not quantity nor impact. A scientist can ensure the quality of his research, but not the impact. If a broad range of investigator-initiated research is supported, and the work is of high quality, societal rewards will follow. Aggressive, top-down efforts to manage this process with the goal of maximizing impact are likely to backfire. History has repeatedly shown us that we do a poor job of predicting the future of scientific progress, and it is better to cast as wide a net as possible, investing both broadly and deeply.

With regard to the post-hoc assessent of research impact, it would be wise to remember that not everything that counts is easily counted. Bibliometric indicators are only one measure of impact. Beyond the # of publications, citations and h-index, a bibliometric measure should account for first/senior vs co-authorship (Ioannidis et al. PLoS Biology, 2019) and laboratory size. How is it reasonable to compare the publication output of a lab with three members versus one with thirty?

In addition to bibliometrics, for assessment of an individual researcher's productivity I favor the use of multiple "metrics" that include influence, societal engagement, innovation, commercialization, research funding, mentoring activities, rank in the field, etc. No single researcher can contribute in all these respects, but for the research community collectively, each kind of contribution is invaluable. No single metric will fit all researchers.

59. Edwin S. Monuki

I am responding: On behalf of myself

Name: Edwin S. Monuki

Name of Organization: University of California Irvine

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/26/2026

1. Rigor and Reproducibility: Development of RIGOR & REPRODUCIBILITY INDICES to induce cultural change within research ecosystems:

To motivate journals, providing a numerical counterpart to impact factor.
To motivate academic institutions, to provide an objective criterion for promotion & tenure.
To motivate PIs, since their publications and promotions will become more dependent on demonstration of R&R.
To motivate trainees to pay more attention to the R&R of the PIs they choose to train with.

(Apologies for submitting twice, but this only came to me after the first submission)

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

60. Benjamin Clarkson

I am responding: On behalf of myself

Name: Benjamin Clarkson

Name of Organization: Mayo Clinic

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/26/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Response-to-NIH-Request-for-Information.pdf

Response to NIH Request for Information: Measuring and Rewarding Scientific Impact

Biomedical research has become increasingly collaborative, interdisciplinary, and data intensive, yet many of NIH's funding mechanisms and measures of scientific success continue to emphasize individual investigator-led projects. While investigator-initiated research remains an essential driver of innovation, there is an opportunity to complement this model by recognizing and incentivizing collaborative approaches that maximize rigor, reproducibility, efficient use of resources, and progress toward high-priority scientific objectives. Advances in artificial intelligence provide a unique opportunity to help NIH identify complementary scientific efforts and organize research in ways that better leverage the collective strengths of the biomedical research community.

One potential approach would be for NIH to develop an AI-assisted portfolio optimization program that analyzes scientifically meritorious but unfunded grant applications to identify complementary objectives aligned with NIH strategic priorities. Rather than evaluating applications solely as independent projects, the system would identify common hypotheses, complementary expertise, shared technologies, and overlapping scientific goals across multiple proposals. Based on these analyses, AI could generate integrated "plans of work" that decompose ambitious research objectives into coordinated modules assigned to laboratories with demonstrated expertise and strong preliminary data in specific areas. NIH program staff would retain responsibility for selecting research priorities, evaluating AI-generated recommendations, approving project plans, and overseeing program implementation.

Such a model would naturally promote team science by recognizing that many important biomedical questions are best addressed through complementary expertise distributed across multiple laboratories rather than expecting a single laboratory to possess every required capability. Participating laboratories could voluntarily opt into this collaborative funding mechanism and receive milestone-based subawards for completion of clearly defined experimental objectives. This approach would provide opportunities for highly capable investigators whose individual applications narrowly missed funding while allowing NIH to assemble research teams optimized for scientific success rather than institutional boundaries.

This framework also offers a practical mechanism for improving rigor and reproducibility. Instead of treating replication as a post-publication activity, NIH could intentionally design independent replication into selected projects by assigning critical experiments to multiple laboratories using standardized protocols and predefined analysis plans. Independent confirmation of key findings would become an expected component of high-impact research rather than an infrequent follow-up study. Differences in outcomes between laboratories would also provide valuable scientific information by identifying sources of experimental variability early in the research process.

The proposed model would further strengthen data sharing and transparency by encouraging participating laboratories to adopt common experimental protocols, standardized data formats, shared repositories, preregistered analysis plans where appropriate, and coordinated reporting of milestones. Because participating investigators would contribute individual components toward a common scientific objective, success would depend upon open communication and timely sharing of methods, data, and intermediate findings. These collaborative practices would increase the long-term value of federally funded research resources beyond individual publications.

Importantly, this framework suggests that NIH broaden the activities it recognizes and rewards when assessing scientific impact. In addition to publications and citations, NIH could recognize meaningful contributions such as successful completion of collaborative research milestones, independent replication of key findings, generation of reusable datasets and validated protocols, development of shared technologies, participation in multidisciplinary research teams, and demonstrated commitment to open science practices. These activities frequently generate substantial scientific value but are not consistently recognized within traditional measures of academic success.

Several measurable indicators could be used to evaluate the success of such a program. These include rates of successful independent replication, time required to achieve predefined research milestones, adoption of shared datasets and experimental protocols by external investigators, efficiency gains through reduced duplication of effort, successful completion of collaborative project modules, production of broadly reusable research resources, and downstream scientific impact as measured by follow-on discoveries, clinical translation, or incorporation into future NIH-funded research programs.

As with any new funding model, safeguards would be essential. AI should serve only as a decision-support tool rather than replacing peer review or programmatic oversight. Final funding decisions, project design, investigator selection, and milestone evaluation should remain under NIH leadership. Participation should be voluntary, intellectual property and authorship expectations should be established before project initiation, and investigators should retain opportunities to pursue independent research alongside collaborative efforts. Initial implementation through pilot programs within selected NIH Institutes or Centers would allow careful evaluation of feasibility, investigator acceptance, and scientific outcomes before broader adoption.

Overall, an AI-assisted portfolio optimization program offers an opportunity to better align NIH incentives with the collaborative nature of modern biomedical research. By recognizing contributions that improve rigor, reproducibility, data sharing, and complementary team science, NIH could increase the return on public investment while reducing unnecessary duplication and enabling scientifically meritorious ideas that might otherwise remain unfunded to contribute meaningfully toward national biomedical research priorities.

61. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/26/2026

1. Rigor and Reproducibility: NIH should fund dedicated replication studies, recognize replication as a valued contribution in peer review, and build infrastructure for data/code sharing and multi-site validation networks.

2. Data, Software, and Model Sharing: NIH should treat data and tools as core research outputs in peer review, biosketches, and promotion. It should fund data curation and require FAIR practices supported by shared repositories and interoperability standards. Embedding sharing into grant evaluation will accelerate validation, collaboration, and scientific impact.

3. Training and Mentorship: NIH should track trainee career progression and long-term engagement in biomedical research. Additional metrics should include mentorship quality, diversity and inclusion across career stages, and equitable access to funding and leadership opportunities. Tracking trainee outcomes, team science participation, and contributions to collaborative research networks will better reflect training and mentorship.

4. Collaboration: NIH should use structured contributor frameworks to document individual roles in team science. It should also recognize team science leadership and cross-disciplinary integration. It should align funding and promotion criteria to value collective impact alongside individual contributions.

5. Entrepreneurship and Translation: NIH should incentivize entrepreneurship by recognizing patents, startups, licensing, and clinical or commercial adoption as meaningful research outputs in peer review and career advancement. It should fund partnerships that accelerate scale-up and real-world use. Innovation impact metrics should be included in grant evaluations to better align incentives with meaningful translation to patient care.

6. Foundational Scientific Exploration: NIH should reward high-risk research by expanding dedicated funding mechanisms where outcomes are uncertain. Proposal evaluation should emphasize originality, scientific premise, and potential for transformative impact rather than short-term productivity.

7. Public Impact: NIH should adopt a balanced, mission-aligned public impact framework that recognizes biomedical research contributions across six domains: health outcomes, clinical practice, economic value, public trust, accessibility, and societal benefit. The framework should combine quantitative indicators with structured narrative evidence, be tailored by discipline and career stage, and emphasize team-based contributions. Most importantly, NIH place less emphasis on research publications, and

more emphasis on whether the research generates reliable knowledge, usable tools, better care, improved health, reduced disparities, and tangible value for the American public. In the past, NIH's Office of Portfolio Analysis developed iCite to quantify the scientific influence of a research scientist. A new tool should be developed to measure the impact of an investigator's research on the six domains listed above.

8. Other Areas: N/A

62. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Member of the Public

Date Submitted: 6/26/2026

1. Rigor and Reproducibility: The NIH's rigor and reproducibility requirements, while well-intentioned, have created a bureaucratic burden that disproportionately increases administrative workload without meaningfully improving research quality.
The policy mandates extensive documentation across authentication of key biological resources, consideration of sex as a biological variable, scientific premise, and experimental design, requirements that often translate into boilerplate language cut-and-pasted across applications rather than genuine reflection on study design. Reviewers, already stretched thin, lack the bandwidth to critically evaluate these sections, reducing them to checkbox exercises.
Moreover, the policy applies uniformly regardless of study type, scale, or risk of irreproducibility. A small mechanistic study in a well-characterized system faces the same documentation burden as a large multi-site clinical trial, a one-size-fits-all approach that misallocates effort.
The root causes of irreproducibility: underpowered studies, publication bias, inadequate methods reporting, and perverse incentive structures, are largely unaddressed by form-filling. Structural reforms, that address these issues, would more directly target the reproducibility crisis.

In short, the current framework creates the appearance of rigor without the substance of it, and risks training researchers to perform compliance rather than practice it.

Additionally, the form-filling creates a burden for the vast majority of studies that are unfunded. A better system might ask for the reformed rigor and reproducibility documents after a grant has received a fundable score...similar to a Just In Time.

2. Data, Software, and Model Sharing: Again form-filling to satisfy the language the RFA does the field a disservice.
Ask for evidence of data, software, model-sharing.

3. Training and Mentorship: Agreed. There are currently minimal mechanisms or incentives to be a superb mentor.

4. Collaboration: Dedicated MPI RFAs are essential for the future of science

5. Entrepreneurship and Translation: NIH's focus on fostering the translation of discoveries from bench to bedside and on better incentivizing and measuring entrepreneurial impact is both timely and

essential. However, the NIH is conservative and has a limited budget.
It will be necessary to accept risk, assume failure, and provide greater support than current to some studies.

6. Foundational Scientific Exploration: Unbiased basic discovery forms the foundation of everything else.

7. Public Impact: Impact is often delayed sometimes by several decades. Therefore, this metric is not practical

8. Other Areas: N/A

63. Alexander Scheeline

I am responding: On behalf of myself

Name: Alexander Scheeline

Name of Organization: University of Illinois at Urbana-Champaign and SpectroClick, Inc.

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 6/26/2026

1. Rigor and Reproducibility: Because there is inherent diversity among individuals and there is noise in any physical phenomenon or measurement, ensuring irreproducibility at some level, any conclusion reached in medical research should be repeated by individuals other than those who undertook the original study. When the second study differs from the original, which is inevitable, the question should not be, "which study is right," but rather "given that more information is now available, what can we learn from the replicate study that could not be learned from one study alone?" Replica studies consume resources -- researchers, patients, laboratory animals, space, and funds. Thus reproducibility studies are most important when results are unclear. Why might a drug help only 1/2 of patients? Why might a surgical procedure go awry for 1/4 of patients? When a result is clearly positive (say, 90% of patients are helped), then follow-up studies may be unnecessary. The problem will be how to devise repeated study proposals so that they are allowed some creativity while not hitting predictable barriers: lack of enthusiasm by review committees, insufficient success rates of proposals, proposals excessively from second-tier researchers or institutions who may inadequately understand the subtleties of research procedures, or considering those who repeat research as being second-tier researchers (which would discourage all but the most senior people from pursuing such studies, as being a "me too" is fatal to a career).

2. Data, Software, and Model Sharing: No comment on this item.

3. Training and Mentorship: Research assistantships should last for the duration of a graduate program rather than being tethered to the standard length of a research grant. Once a student enters a graduate program, provided they apply themselves, they need continuing access to a living salary and research consumables. The long-time practice of funding students via 3-year (NSF) or 5-year (NIH) research grants makes funding stability precarious. While programs to train entering students (initial two years of graduate school) might well continue the current traineeship programs, the long-standing practice of hanging advanced students out to dry when grants are not renewed has long been an impediment to research completion.

4. Collaboration: No comment on this item.

5. Entrepreneurship and Translation: The "valley of death" between academic progress and societal impact for development of pharmaceuticals, instrumentation, diagnostics, and procedural advances is real. Review committees prevent progress by labeling a proposal too fundamental or too applied, too

narrow or too broad, when one phase of development ends and support of the next phase is sought. SBIR has PIs seeking a pinhole that isn't too focused on business development, too focused on fundamental science, too narrow or too broad in its scientific questions, has too few or too many resources invested in particular enterprise. Translation now works when the pace of that translation matches preconceived agency guidelines. The guidelines should be more flexible. The proposal should state: where is the science and the business development at the moment? What path is envisioned to achieve translation? Review, rather than looking for excuses to decline the proposal, should be focused on incrementally moving technology to where it can help people and businesses. Should Phase I proposals only lead to Phase II SBIR support, or might multiple years of Phase I level support be useful? If the business plan is weak but the science strong, should the science continue to be supported with NIH or NSF helping match the scientist to plausible business partners? If the business plan is strong but the science is weak, might NIH or NSF connect the business to scientists who could strengthen the technical side of the work? If the project is worth doing, the review process should aim to help the work support the public interest. Too often, current proposal review is focused on finding weak spots in the proposals with the goal of saying no to as many approaches as possible, particularly those that don't fit into normal program straitjackets, so that those proposals that succeed fit neatly into bureaucratic or conceptual silos.

One might imagine a graduate student completing a thesis that has commercial possibilities. An SBIR saying, "I want to continue this research and eventually build a business around it" stands little chance of being funded. An approach based on ICorp would focus on finding customers for the incomplete product rather than further developing the technology to the point that a discerning customer would see value in that technology. Meanwhile, business school graduates might want to be part of a startup, but not know what technology needs to be developed. Whether the science needs 6 months or 20 years to develop, whether the scientist is a good salesperson or an utter failure at sales should not be the criterion for subsidizing entrepreneurship and translation. If it looks like a promising idea, back it!

6. Foundational Scientific Exploration: Truly fundamental research has unclear impact, looks absurd to the non-specialist (especially politicians and lawyers), often is at odds with accepted understanding, and may not pay off for extended periods, if ever. The idea that politicians should have any involvement in deciding whether to support high risk or controversial work is absurd. Remember Lysenko in the Former Soviet Union and the work on mRNA vaccines. There are still politicians who, based on dogma, refuse to believe the success of such technology. If the scientific community thinks something is interesting, fund it and leave people alone to follow their noses. If 90% of high risk or fundamental science leads to dead ends, that is useful in the long run by showing where certain ideas or ill-founded. It's the 10% that build the future. And that 10% can only be identified in retrospect.

7. Public Impact: US life expectancy lags that of other advanced societies. Since the science underpinning medical practice has in large measure come from US research since 1945, the problems we see in routine medicine must be structural rather than intellectual. On the other hand, we continue (or, perhaps, continued through 1/19/2025) to lead the world in advanced biochemical and biogenetic technology, providing tools which, if aggressively developed, can improve longevity and disease suppression. Where we appear to be falling short is: community health, broad-based appreciation for rigorous research instead of hunches, snake oil cures, and universal access to care. The public needs to be educated, both in K-12 and later in life, of the success of scientific medicine and how scientific

understanding of folk remedies leads to improvements over holistic or alternative approaches. Favoring our current insurance structure over single-payer coverage appears to be correlated with our high costs and reduced longevity. Alas, a significant fraction of the populace believes otherwise. If the statistics are lying, someone needs to provide an alternative explanation that stands up to critical review. "The way we've always done it" is the worst excuse for doing almost anything. We also need to rebuild the wall between religious preferences and medical practice. If someone wishes to forego medical treatment on religious grounds, there are only limited cases where it may be justified to override that preference. However, having one person's preferences override another's (as is happening in the abortion wars) seems antithetical to freedom.

8. Other Areas: Science thrived from the Second World War through the lunar landings. Thereafter, there has been a gradual drift to lower support by State governments of universities, greater paperwork burdens on proposals, less flexibility in carrying out research, and greater political interference in choosing and executing projects. Some of this is because of the great burden of medical costs on States so that fewer resources are available for education. Some is due to longer lifespans so that pensions consume more resources than in prior decades. Politicians in a hurry for results when more fundamental science is needed is also a concern. Nixon's "War on Cancer" was decades premature. As understanding of biochemistry has progressed, the time for that war is likely now -- except NIH and CDC have been radically resectioned and their momentum broken. Just as Germany had a brain drain in the 1930's that initiated US scientific dominance, current political currents threaten that the US will hemorrhage talent to Europe and East Asia. It is easy to destroy institutions and research groups. It takes decades to recover. Research cannot be turned on and off like a light switch. Bringing a graduate student from basic scientific literacy to the point when research progress is perceptible can take 2-3 years. Getting an independent laboratory to productivity may take a similar amount of time. Fully funding science in 2026 would not result in a return to full productivity until 2029 at the earliest. There are two (or more) election cycles in this time. Unless the society as a whole backs science adequately, continuously, and with minimal interference. the future belongs to other countries that will.

64. Chris Batich

I am responding: On behalf of myself

Name: Chris Batich

Name of Organization: U Florida

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/26/2026

1. Rigor and Reproducibility: Every proposal should have a brief paragraph (half-page) estimating the financial and lifespan benefit if the work is successful. Dialysis for kidney failure is prime example. If one could grow a kidney from a decellularized pig scaffold, that would have a major impact on healthcare costs. When such projects are funded and give positive results, they should be checked by an independent lab. The book series "Organic Synthesis" had a system which worked well. Upon submission (of a funded project), a proposal was offered to an anonymous and independent lab to repeat the work. If successfully replicated, the reviewers would be identified in print. After reading the book "Doctored", I believe that about half of the high impact results are fabricated. They only identified those with 2D-gels from the images. Some consequences should result (loss of funding at least).

2. Data, Software, and Model Sharing: NIH used to have "Consensus Conferences" which did this very nicely. Bring them back?

3. Training and Mentorship: Allow more local authority to distribute some funds. Give grants to institutions and allow departmental chairs to allocate funds. Monitor results, not promises.

4. Collaboration: When I worked for DuPont's Central Research Department, we formed teams based on identified problems and needs. An announcement was made about a problem or need and all were invited to attend if interested. About 30 people usually showed up for a presentation and discussion. Various ideas were suggested and small groups of about 3 to 5 people met to plan how to test solutions. Sometimes additional people were invited to join based on needed skills. Testing for feasibility was developed and carried out. Very successful process. Needs a collegial group or a strong moderator to get everyone's opinion expressed and to keep the louder mouths restrained.

5. Entrepreneurship and Translation: Someone from CMS should add their opinion as to whether the approach proposed would be reimbursed if successful eventually... and what it would take to demonstrate clinical readiness.

6. Foundational Scientific Exploration: Joint NSF-NIH funding and good funding for engineering departments as well as science.

7. Public Impact: Regain trust by getting rid of the fake results and author sharing ("I will add you, if you add me")

8. Other Areas: Keep up the good work and don't get discouraged. You all are the best hope for better lives ahead!

65. PWXPress

I am responding: On behalf of an organization

Name: Gio Saravini

Name of Organization: PWXPress

Type of Organization: Non-Profit Research Organization

Role: Bioethicist

Date Submitted: 6/28/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

66. William Brett Perkison

I am responding: On behalf of myself

Name: William Brett Perkison

Name of Organization: University of Texas School of Public Health

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/29/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: I would recommend broadening the scope of what the NIH considers to be human health research. Grant funding should integrate health outcomes with both financial mechanisms and ecological systems into the scope of work. Our medical research cannot continue to look at findings only in terms of human health.

To elaborate, financial mechanisms need to be in place which can make translational research to practice be financially viable after a grant is completed. Health care system economists, medical billing experts, and other health care systems analysts need to start being part of research teams. Billing, in particular, needs to be an intrinsic part of how an intervention would be put in place. This would help deliver innovations to the people who need them in a much more expedited fashion.

Regarding ecological systems, with meteorological events becoming more intense and more frequent we cannot afford to operate as if human innovation is separate from natural ecosystems. Grant funding needs to include how an intervention will affect ecosystems and how it anticipates upcoming changes in global ecological conditions as well as how intends to off set carbon emissions.

An example would be any program which involves nutrition. What would be the carbon foot print of a nutrition intervention, how would a nutrition intervention improve sustainability in light of increasing droughts, increasing food prices, and decreasing nutritional value in foods due to an increase in atmospheric carbon dioxide levels. All of these factors are critical to incorporate into any true research innovation.

I know this will involve much work for researchers, NIH staff, and cut into research funding but ignoring these factors in research criteria is, in the long run, invalidating the existing research being done.

7. Public Impact: I would recommend including a requirement , as part of the grant and include funding to be able to disseminate findings more thoroughly. This should be included as criteria for reviewers as well. I would recommend not only messaging to the communities affected or participating in the research but also members of our U.S. community who traditionally don't receive any messaging about what research is occurring NIH. Making it a requirement for instance to present findings to members of rural communities , church communities, radio and television. The NIH is doing wonder work but their is huge disconnect between the research community and those who are voting and electing people to office who are not getting enough exposure to the findings.

8. Other Areas: N/A

Description: Discusses Broadening the scope of research funding and widening the dissemination of the findings.

67. Benjamin D. Singer, MD, ATSF

I am responding: On behalf of myself

Name: Benjamin D. Singer, MD, ATSF

Name of Organization: Northwestern University Feinberg School of Medicine

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/29/2026

1. Rigor and Reproducibility: The current incentive structure rewards the first report of a finding far more than its confirmation, even though confirmation is what converts an interesting result into a reliable one. NIH could correct this issue by creating a distinct, citable funding and publication category for replication studies, and by explicitly crediting replication work in grant applications' biosketches and progress reports. For example, there could be a structured field for "independent replications of [investigator]'s findings by other groups," which currently has no home in the application or progress report and is therefore invisible to study sections. This adjustment should not come at the expense of novelty: replication and discovery are complementary, not competing, goods, and NIH's portfolio and review criteria should continue to prioritize and reward genuinely novel, innovative work even as it builds new pathways to credit replication. A field that rewarded only confirmation of existing findings would be just as distorting as one that rewards only novelty.

A second, underappreciated rigor problem is definitional: "causal research" is used loosely across disciplines that differ enormously in what counts as causal evidence – a randomized clinical trial, a Mendelian randomization analysis, and a genetic knockout in a mouse model all claim causality but carry very different epistemic weight. NIH should avoid a single causal-inference standard and instead ask applicants and reviewers to state explicitly which causal framework applies (interventional, quasi-experimental, or mechanistic) and what assumptions that framework requires. This would do more for rigor than a generic "is this causal?" checkbox.

From my own experience developing genetic mouse models to test causal mechanisms in lung injury and then testing the resulting hypotheses in patients, replication is genuinely difficult at every stage, just in different ways. Genetic and mechanistic experiments are constrained by defined alleles and protocols, which makes them somewhat more tractable to replicate than clinical studies, but they remain vulnerable to irreproducibility from genetic background effects, microbiome and vivarium differences across institutions, and subtle variation in technique that rarely makes it into a methods section. Mid-scale translational studies add a further layer of difficulty because they also depend on clinical phenotyping decisions (cohort definitions, timing of sample collection, outcome adjudication) that are even less standardized and even more rarely specified in enough detail to be replicated. NIH could require, as a condition of funding for clinical and translational R01/U01 awards, a structured "phenotyping and replication appendix" specifying these decisions in advance; a lighter-weight analogous requirement for mechanistic awards—specifying genetic background, husbandry conditions,

and reagent sources in a structured, searchable format—would make replication funding proposals easier for others to scope and justify across both domains.

2. Data, Software, and Model Sharing: Data-sharing mandates have so far rewarded the existence of a deposited dataset far more than its actual reuse, which creates an incentive to deposit the minimum compliant dataset rather than a maximally usable one. NIH should track and reward downstream reuse—for example, citation of a dataset's accession number in subsequent papers by other groups—as a distinct, fundable metric, the way citation of a paper is tracked today. This approach is already technically feasible through dbGaP, SRA, and GEO accession tracking and would shift the incentive from compliance to usefulness.

I have directed a multi-site, multi-investigator cohort study of patients with severe lung disease that generates genomics/transcriptomics sequencing data and clinical phenotype data intended for reuse by the broader research community. The single largest barrier to making those data maximally reusable has not been the sequencing or the deposition itself, but the unfunded, ongoing curation work: harmonizing clinical metadata across collecting sites, maintaining computational pipelines as reference databases and tools evolve, and fielding outside investigators' questions about the dataset. None of this work currently has a grant mechanism, a publication outlet, or a credit line. NIH could create a dedicated, renewable "data resource maintenance" supplement or small mechanism (analogous to existing administrative supplements) specifically to fund this curation labor, with progress reports structured around reuse metrics rather than new findings.

For software and analytical tools developed within R01-funded mechanistic work (e.g., pipelines for genomics/transcriptomics/epigenomics, including single-cell analysis), there is currently no NIH mechanism that funds the maintenance of a tool after its first publication, even though tool abandonment after publication is one of the most common and most damaging reproducibility failures in biomedical research. A renewable, lightweight "tool maintenance" supplement, awarded competitively based on evidence of community use (GitHub forks, downloads, or citations of the tool itself rather than the original paper), would address this issue directly and is something program staff could administer with existing infrastructure.

3. Training and Mentorship: After nine years directing a departmental Physician-Scientist Training Program spanning multiple T32 mechanisms, the metric I find most absent from current NIH accounting is mentorship's effect on trainee independence, as distinct from trainee productivity. A T32 progress report today asks how many papers and grants trainees produced; it does not systematically ask whether trainees who left the program went on to obtain independent R-level funding, and on what timeline relative to peers. NIH's own training-grant outcome data could be linked longitudinally to RePORTER funding histories to generate this metric automatically, without adding any new reporting burden on program directors. These data already exist across NIH's own systems.

Mentorship that produces co-senior authorship (i.e., where a mentor steps back from first or sole senior authorship specifically to establish a trainee's standing as an independent investigator) is one of the clearest behavioral signals of investment in the next generation, yet it is invisible in current bibliometric accounting, which simply counts the mentor's overall publication record without distinguishing this kind of deliberate credit-sharing. I have done this deliberately with departing trainees moving to faculty

positions, and there is no field in the biosketch or progress report that captures or rewards it; if anything, current author-order conventions can make a generous mentor's CV look less productive by senior-authorship counts. NIH could add a structured "mentee transition to independence" field to the biosketch, listing trainees, their current positions, and authorship roles a mentor ceded along the way.

Editorial service is similarly undercounted: time spent training the field's reviewers and authors—for example, the substantial editorial labor of triaging and shepherding hundreds of manuscripts through peer review—functions as field-wide mentorship but receives no NIH recognition mechanism comparable to the credit given for formal training grant roles.

4. Collaboration: Team science is currently evaluated almost entirely through the lens of individual attribution (author order, percent effort, role on the award), which forces collaborative contributions into a single-investigator accounting framework that does not fit how the work is actually done. Multi-PI awards, P01 program projects, and large cooperative agreements (U01s, UM1s) already exist as funding mechanisms, but study sections and tenure committees still tend to discount a PI's role on these awards relative to a comparably sized R01, even when the multi-PI award required more coordination, not less. NIH could address this issue directly within its own peer review guidance by instructing reviewers to evaluate multi-PI and program-project contributions on their own terms (e.g., coordination, integration across cores, and scientific leadership of a sub-project) rather than as a diluted version of single-PI productivity.

In assembling and contributing to multidisciplinary proposals—for example, project teams spanning immunology, synthetic biology, and engineering for translational technology development—the highest-friction part of team science is rarely the science itself but the absence of any shared NIH-recognized format for documenting who is responsible for what within a team, beyond the brief "role on project" line in a biosketch. A standardized, structured contribution statement (modeled on the CRediT taxonomy already used by many journals) adopted for NIH biosketches and progress reports would let teams document specific contributions (e.g., study design, data curation, mechanistic interpretation, clinical translation) in a way that travels with the investigator across applications and gives study sections something more substantive than author order to evaluate.

Center grants and large cooperative agreements should also be explicitly exempted from any future per-PI funding caps or productivity benchmarks calibrated to single-PI awards, since these mechanisms are structurally built to require multiple concurrent PI roles.

5. Entrepreneurship and Translation: The single biggest disincentive to translation in my own field is timing: the metrics that determine a mechanistic researcher's success (publications, grant renewals) operate on a one-to-three-year cycle, while translating a finding into a clinical trial or product realistically takes five to ten years and often does not yield a publishable result until very late in that process. A study section reviewing a competing renewal five years into translational work will frequently see "fewer papers" relative to a comparably funded purely mechanistic lab, even when the translational lab has done more total work. NIH could address this by explicitly instructing study sections, in review guidance, to weigh translational progress milestones (IND filings, trial initiation, regulatory interactions) as equivalent evidence of productivity to publications when evaluating renewals of awards with a stated translational aim.

Moving findings from genetic mouse models toward interventional trials in patients has also exposed how little institutional or NIH credit exists for the specific skill of translating a mechanistic finding into a trial-ready hypothesis, which is distinct from either basic discovery or clinical trial execution. This intermediate translational step, often the niche of physician-scientists, currently has no natural funding home (it is often “too applied” for basic mechanism review and “too early” for clinical trial mechanisms) and no natural reward in promotion or tenure review, which tends to recognize either basic-science productivity or trial leadership but not the connective work between them. A dedicated, lower-barrier translational bridge mechanism would fill a real gap and create a citable, fundable category for this work.

6. Foundational Scientific Exploration: High-risk, high-reward and purely mechanistic work is disadvantaged by any metric that rewards near-term, predictable output, because the work that most needs protection is, by definition, the work where a negative or null result is a real and informative possibility. NIH’s existing high-risk mechanisms (e.g., the Pioneer and New Innovator awards) are valuable but small in number and concentrated among a small set of awardees; the cultural problem is broader and affects ordinary R01 renewals, where reviewers often penalize an honestly reported pivot away from a hypothesis that did not pan out.

The clearest fix available without new mechanisms is in review guidance: instruct study sections explicitly that a well-justified pivot away from a non-productive hypothesis, accompanied by rigorous evidence for why the original hypothesis was abandoned, should be scored as evidence of rigor and scientific maturity rather than as a productivity shortfall. This single change in how renewals are read would do more to protect foundational exploration across the entire R01 portfolio than any expansion of dedicated high-risk mechanisms, which by design reach only a small number of investigators.

NIH should also resist the temptation to make “translational potential” an implicit scoring criterion for basic mechanistic R01s reviewed in study sections nominally focused on fundamental biology; the pressure to gesture at downstream clinical relevance in basic science aims pages, even when premature, distorts the foundational work the RFI itself identifies as worth protecting.

7. Public Impact: Public impact is hard to measure well, and NIH should be cautious about adopting metrics that are easy to game (social media mentions, news coverage) rather than metrics that track durable change in practice or policy. A more defensible approach: track and credit specific, attributable downstream uses of NIH-funded findings: citation in a clinical practice guideline, incorporation into an FDA label or guidance document, use in a CDC or WHO recommendation, or adoption in a clinical decision-support tool. These are discrete, verifiable events that already exist in public documents and could be linked to the underlying grant through existing citation-tracking infrastructure, without requiring investigators to self-report soft impact claims.

From my involvement in interventional trials in critically ill patients, the public impact that matters most to me—measurable changes in mortality or recovery time in a defined patient population—is exactly the kind of outcome that does not show up in citation counts or h-indices at all, and is currently recognized, if at all, only informally in narrative biosketch sections that reviewers may or may not weight consistently. NIH could add a structured, optional field to the biosketch for “downstream clinical or policy adoption of NIH-funded findings,” analogous to the existing personal statement, specifically inviting this

kind of concrete, externally verifiable impact claim rather than leaving it to be buried in prose.

Public trust should be treated as a public impact outcome. Transparent, accessible reporting of negative and null results, which currently have almost no path to publication or credit, would do as much for public trust in science as some positive findings.

8. Other Areas: Across the seven areas above, the common thread is that NIH already collects much of the data needed to credit these activities — training outcomes via RePORTER, dataset reuse via repository accession tracking, downstream citation via iCite — but does not yet route that data into the review, renewal, and promotion processes that shape investigator behavior. The highest-leverage changes available to NIH are not new awards or new infrastructure, but explicit instructions to study sections about what to credit, and structured fields in existing documents (the biosketch, the progress report) that make currently invisible contributions visible and fundable.

68. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Member of the Public

Date Submitted: 6/29/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: Research should be publicly available and include a short lay summary.

3. Training and Mentorship: N/A

4. Collaboration: Teams should include professionals in the field as well as those focused solely on research as to promote benefits of research reaching the public.

5. Entrepreneurship and Translation: Universities have startup incubators. Partner with these organizations to encourage discovery. Grants should specify that the technologies are affordable to the public. Many drugs or technologies though helpful, are cost prohibitive. Example: medications without generics

6. Foundational Scientific Exploration: Reduce bureaucracy surrounding guidelines, reporting, etc. In the past two years more guidelines have been implemented (ex research security) that have increased the administrative burden of doing research. This takes time away from scientific discovery and shifts it to paperwork. The unstable environment with layoffs, shut down (in ability to ask questions to NIH staff) and policy changes results in re-work for administrators due to unclear or revised instructions. Clear, straightforward, bi-partisan policy is needed.

7. Public Impact: Focus is needed on comprehensive women's health and how a woman's unique body impacts other health concerns. For example, conditions such as migraine can be influenced by hormone changes during menstruation and this may not be the specialty of a primary care doctor, neurologist, or gynecologist, leaving women guessing about their health. Additionally, research is needed about stroke and pregnancy, endometriosis, PCOS, infertility issues, perimenopause, and many other concerns for women. Diverse populations should be studied including by race, ethnicity and socio-economic status to obtain the best results.

8. Other Areas: N/A

69. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/30/2026

1. Rigor and Reproducibility: All science needs to be evaluated for reproducibility. We need to fund reproducibility studies and publish them. Independent reproduction of key experiments by (ideally) another lab needs to be incorporated into all project plans and publications. This used to be done. But as science has become more competitive and driven towards unrealistic publication counts as the only metric for career advancement and grant funding, best practices for sound science have been lost. Publication counts should no longer be part of grant funding decisions. Ideally, grants should be evaluated in a blinded fashion so that only the science is evaluated, not how many publications a researcher has gotten, how prestigious the PI is, or who their friends are on the review committee. The best scientists publish fewer papers than the sloppiest scientist. But the system for evaluating success in science are broken, and we are rewarding bad scientific practices in the current system. One sloppy practice that has begun infiltrating scientific publications is the presentation and statistical analysis of technical replicates rather than independent experimental repeats. Far too many articles now show "one representative experiment" instead of aggregating the independent experimental repeats. Thus, the scientific literature is now clogged with statistical validation of experimental approaches rather than measurements of biological variability. This must stop!

2. Data, Software, and Model Sharing: This is already a requirement for grant funding, and many journals require it as a condition for publication. I do not see this as a problem that needs to be solved.

3. Training and Mentorship: This is another area of significant problems in academic science. Mentors are basically able to treat their mentees however they want. Sadly, too many PIs treat post-docs as indentured servants rather than PIs in training. I was in a toxic lab environment, and it took me 2 years to transition into another, more supportive lab. Those 2 years counted against me, and led to me not being eligible for K99/R00 awards, which would have been a big boost for my research career. Stop punishing trainees for time lost due to bad mentors. We also need more effective metrics of success for promotion to faculty. It all comes down to publication counts, and that rewards sloppy science. And one of the most important components of promotion is your mentor championing you for promotion. An apathetic mentor or one with inherent bias (often against women, parents, people of color, etc.) can hamper the career advancement of trainees. It is difficult to assess the quality and impact of mentorship. One way would be to ask the mentees if they are getting what they need to advance their training in a supportive environment. But that would need to be done in a way that doesn't threaten their relationship with their mentor. Can there be mentorship committees that have clout and the responsibility to help transition the mentee out of a bad environment without disrupting their career advancement? Can we stop using

publication counts as the metric of success for promotion and instead use reproducibility and sound science instead? Can we reduce bias in grant evaluation through blinding to the identity of the applicant? The quality of mentorship absolutely cannot be based on scientific publication counts or grant funds awarded, as we know both of these metrics are biased, and publication counts rewards sloppy science. Can we instead measure opportunities for advancement through things like (1) designing one's own scientific program with support rather than having projects and experiments assigned by mentor, (2) attending scientific conferences, (3) presenting at scientific conferences, (4) having protected time to complete advanced coursework, (5) having access to support staff like technicians or graduate students or undergraduate research assistants? Can we somehow measure a mentor championing their trainee? Can we measure trainee mental health? It has been known for a long time that academic researchers have higher rates of depression and anxiety than the general public. This must be addressed. Can we begin recognizing post docs as experienced researchers rather than perpetual trainees? Can we begin providing retirement funds to graduate students and post docs? Can we stop financially punishing academic trainees? A major cultural shift needs to occur and how we think about trainees and the academic pipeline's responsibility towards them. If we are going to call them trainees, then we need to train them not exploit them as cheap labor.

4. Collaboration: Reduce emphasis on first- and senior-author positions in publications. Anything else would likely just create more work for the underlings in the lab that are already overwhelmed.

5. Entrepreneurship and Translation: This is an unrealistic expectation for academic researchers at this time. The system is not designed for this, and many academic departments punish researchers for entrepreneurship. Furthermore, academic researchers are already doing multiple jobs and are already burnt out. It is not realistic to expect academic researchers to take on the tremendous responsibility of building a company. For this to work, the entire academic system needs a major overhaul that provides researchers of all levels with an appropriate work-life balance and protected time for additional pursuits, such as entrepreneurship. Academic culture needs to change to see the value in entrepreneurship and support it. Academic researchers also lack the training and support necessary to successfully build companies. This infrastructure needs to first be created.

6. Foundational Scientific Exploration: Yes! Start funding them, then! It is common knowledge that you can only get a grant for something you've already done. Change the mindset at the NIH if you want to see this happen.

7. Public Impact: Yes! Remove political influence from scientific evaluations and awards if you want this to happen.

8. Other Areas: Academic research is broken. Researchers are burnt out and depressed. The system rewards sloppy science. The system pushes out parents, women, and minorities of all kinds. You have to fix the system first. Work–life balance, appropriate financial compensation, and protection from toxic work environments must be built into this system. Publication counts need to stop being a metric used to evaluates one's contribution to science. Instead, the quality of the science—for example, it's rigger and reproducibility, appropriate hypothesis testing and statistical methodology, the innovation and impact—need to be what matters. As the biggest awarder of grants in the US, you hold the purse strings. This gives you tremendous power to shape how we evaluate and reward science and scientists. Use that to transform the system into something sustainable and effective rather than maintaining the status quo that churns out an overwhelming volume of poor scientific quality publications.

70. James Kochenderfer

I am responding: On behalf of myself

Name: James Kochenderfer

Name of Organization: NIH, NCI, CCR, Surgery Branch

Type of Organization: Government Agency

Role: Investigator/Researcher

Date Submitted: 6/30/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: Measurements of translation that are most pertinent are advancing a new therapy to clinical trials and then the efficacy of the intervention in clinical trials. A more long-term measurement of success is FDA approval of a new therapy. Other measures of entrepreneurship are the number of early investigator reports filed, the number of patent applications filed, the number of inventions patented, the number of FDA-approved drugs a lab produces, the number and size of cooperative research and development agreements with industry.

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

71. Yousuke Takahama

I am responding: On behalf of myself

Name: Yousuke Takahama

Name of Organization: National Institutes of Health

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/30/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: In science, collaborations should be basically encouraged. Wide and flat restrictions of individual collaborations based on non-scientific reasons can severely discourage scientists and thereby damage science. Specific concerns, like international conflicts and tax money appropriacies, should be carefully evaluated case-by-case. NIH should remain scientist-centric institute and stay away from politician's playground. Collaborations at NIH may be individually assessed in a manner similar to the ongoing ethics evaluation in NIH.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Foundational scientific exploration is the core of science. It is a beautiful treasure of humankind, thanks to humanity's long and winding history. It is so delicate and vulnerable. It should therefore be evaluated carefully one-by-one by a group of scientific experts, like the site visits at NIH and the manuscript reviews for journal editors. It is important to be aware that a single evaluation can be wrong, so that it is important to install a failsafe mechanism to evaluate and correct, if needed, any prior evaluation. This basic aspect of science should never be evaluated based on fashion-based metrics including journal metrics, which was originally designed for librarians (not at all for individual papers). Honjo's Nobel-prize-winning discovery of PD-1 was published in EMBO Journal, Ohsumi's Nobel-prize-winning discovery of autophagy was published in FEBS Leeters, and Tonegawa's Nobel-prize-winning discovery of V(D)J recombination was published in PNAS. These journals are scientifically well-appreciated journals, but not fashionable journals like Nature, Science, and Cell.

7. Public Impact: N/A

8. Other Areas: N/A

72. Ana-Maria Vranceanu

I am responding: On behalf of myself

Name: Ana-Maria Vranceanu

Name of Organization: Mass General Hospital

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 6/30/2026

1. Rigor and Reproducibility: Release PARs or RFAs specifically designed for replication of prior trials by different research teams.

2. Data, Software, and Model Sharing: All data for all projects should be openly shared, alongside code.

3. Training and Mentorship: For all training and mentoring grants (K24, K12, T32, R25, etc), there should be clear metrics for success that are objective and measurable and collected long term. These should be set out at grant award and submitted at renewal. Things like: 1) nr of mentees with K and R grants; 2) diversity of mentees; 3) mentees who become mentors themselves should be priorities vs papers and conference presentations. Further, unique accomplishments of trainees vs copies of mentors work should be accounted for.

4. Collaboration: Collaborations are going to decrease if the number of grants we can submit or hold as PI or MPI will decrease. Even wit 6 submission, I am now considering single vs MPI grants. Removing restrictions of nr of grants will facilitate collaboration. In addition, developing PARs and RFAs that require MPIs from different disciplines working together on projects will enhance collaboration. Further, funding training programs that are interdisciplinary will also help. One challenge is balancing collaboration with independence. Academic promotion does not incentivize <MPI grants and collaboration. At top university one has to have grants as single PI in order to be promoted.

5. Entrepreneurship and Translation: Developing opportunities for renewal of grants to avoid gaps with new submissions.

6. Foundational Scientific Exploration: NA

7. Public Impact: Mandating return to results and developing protocols for how to do that well will help.

8. Other Areas: Please do not limit nr of grants people can have as PIs. This will destroy collaborations and good science. There will be no incentive to split effort and serve as MPIs. Many of us are 100% in soft money environments and 4 R01s are not enough even as single PI to keep a lab alive and mentor scientists.

73. Peter Gyarmati

I am responding: On behalf of myself

Name: Peter Gyarmati

Name of Organization: University of South Carolina

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/1/2026

1. Rigor and Reproducibility: Any publication should require that the findings are repeated in at least 2 other independent labs.

2. Data, Software, and Model Sharing: It would only make sense if there is a thorough normalization of data to begin with. There is such a broad range of procedures that data/software/model sharing alone would not solve anything.

3. Training and Mentorship: Increased financial support for trainees, number of trainees should be as relevant as eg number of publications. Would require documentation, and a rigorous set of criteria for what a trainee/mentee is

4. Collaboration: Collaboration alone is neutral, does not necessarily mean beneficial therefore shall not be rewarded.

5. Entrepreneurship and Translation: Should be one of the most important criteria when deciding funding. NIH lags behind in funding diagnostic methods/tools. NIH prioritized 'mechanistic research' for decades, this shall end, and practicality, translatability should be prioritized.

6. Foundational Scientific Exploration: More important would be: 'Does this research address a real question?' If NIH really want paradigm-shifting breakthroughs, it would require blank funding of investigators with proven translational potential for a decade.

7. Public Impact: For that, it would be necessary to investigate and evaluate NIH practices in the last several decades, and publish findings.

8. Other Areas: Simplify the grant application processes. PIs spend most of their times on writing applications with a ~12% success rate.

74. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Healthcare Institution

Role: Healthcare Provider

Date Submitted: 7/1/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: It would appear from the scientific review panel that the only metric that the NIH appears to use to measure productivity is the number of published papers. This is an extremely faulty metric that leads to the same people getting more and more funding, being incentivized to produce bad science that gets published in terrible journals. It has no real impact on patient care, but they can state that a grant led to 200 publications, and that's all the review panel looks at. If nothing else is changed, and even if that can be changed, the NIH review and what we invest in will be better.

75. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 7/1/2026

1. Rigor and Reproducibility: Journal reputation is a measure of scientific impact that is seriously underutilized by the NIH. The measure is imperfect but superior to other measures that have been suggested or tried.

Journals build their reputation over time primarily based on the impact of the studies that they publish. Rigor and reproducibility have always been critical to scientific impact over time.

NIH should encourage, recognize, and incentivize publishing in highly reputable journals. Journal reputation can be assessed by a combination of citation metrics (e.g., impact factor) and surveys of scientists (e.g., Nature Index) and other informed users of science.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: Collaboration is a means to an end. NIH evaluation should focus on the end (scientific impact), not on the means.

5. Entrepreneurship and Translation: Journal reputation is a measure of scientific impact that is seriously underutilized by the NIH. The measure is imperfect but superior to other measures that have been suggested or tried.

Journals build their reputation over time primarily based on the impact of the studies that they publish. Translation into products is part of the scientific impact that enhances journal reputation over time.

NIH should encourage, recognize, and incentivize publishing in highly reputable journals. Journal reputation can be assessed by a combination of citation metrics (e.g., impact factor) and surveys of scientists (e.g., Nature Index) and other informed users of science.

6. Foundational Scientific Exploration: Journal reputation is a measure of scientific impact that is seriously underutilized by the NIH. The measure is imperfect but superior to other measures that have been suggested or tried.

Journals build their reputation over time primarily based on the impact of the studies that they publish.

Foundational scientific breakthroughs are already key elements of scientific impact and journal reputation over time.

NIH should encourage, recognize, and incentivize publishing in highly reputable journals. Journal reputation can be assessed by a combination of citation metrics (e.g., impact factor) and surveys of scientists (e.g., Nature Index) and other informed users of science.

7. Public Impact: Journal reputation is a measure of scientific impact that is seriously underutilized by the NIH. The measure is imperfect but superior to other measures that have been suggested or tried.

Journals build their reputation over time primarily based on the impact of the studies that they publish. Public impact is already part of the scientific impact that enhances journal reputation over time.

NIH should encourage, recognize, and incentivize publishing in highly reputable journals. Journal reputation can be assessed by a combination of citation metrics (e.g., impact factor) and surveys of scientists (e.g., Nature Index) and other informed users of science.

8. Other Areas: N/A

76. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 7/2/2026

1. Rigor and Reproducibility: Ideas:
1) create a branch of the NIH who's only responsibility is replicating study findings that are a) replicable, b) most likely to have public health impact, c) contradictory to established literature, or d) new. The work may be done by staff scientists or contracted out to other sites. The prior NIH-funded investigative team must comply with the branch to ensure completion. Publish results with a specific identifier in the title or abstract.
2) increase the quality of scientific review such that methodological details and results are provided clearly so studies can be reproduced--this is too big of a problem for the NIH to tackle, but it's a genuine problem.

2. Data, Software, and Model Sharing: This takes times and not everyone knows how to do this well, especially at smaller institutions or non-academic universities where there is not as much administrative support or specialization. Consider having NIH staff experts assigned who's only job is to help virtually on on-site to ensure this happens.

3. Training and Mentorship: N/A

4. Collaboration: It can be hard for new researchers or institutions to break into this space, especially when it's dominated by larger universities and/or richer institutions that can send their researchers to more conferences. It's also hard to know who what specific expertise, equipment, grad students to support execution or creating collaborations--which are harder when they're outside your organization. Finding and curating solid collaborations also take time. In general, the field often knows what those really important studies that should be done but seem impossible due to funding restrictions or difficulty recruiting enough participants. We need to find a way to expedite the logistical and bureaucratic hurdles. Besides compensation for a site's time in recruiting for a study, perhaps there should be a formal CV-worthy acknowledgment besides authorship that still incentives sites to contribute.

5. Entrepreneurship and Translation: This is NIH's greatest weakness, in my opinion. I've been part of review sessions where discussed grants in the top half (now top third), including those that score well, feel like research for research sake. Too much is incentivized when some technology sounds cool or will gather all types of information. Often it's done with such expensive equipment or by people with such specific, non-scalable processes that it'll never be implementing into real life. And so many grants are funded that are simply correlational. Many researchers don't have training in causality, at least from

observational data. Covariates or confounders are rarely fully identified by the grant submitters and grant reviewers rarely ask for them. Or the Significance section might mention A, B, and C are all related to outcome Y but no specifics are given or required to make it clear that maybe A explains 50% of the variance and factors B and C only explain 5% each and this study wants to study factors B and C but not model A in their statistics. It inflates the importance of B and C in the scheme of public health and doesn't provide a clear picture on intervening on which factor gives the best ROI. Repeatability or measurement error is often not acknowledged sufficiently with people providing ICCs rather than absolute reliability/agreement metrics.
The NIH should create a team follows up with funded researchers each year for up to 10 years to see if and how their research has been implemented/translated--this is a hard metric(s) to quantify, I get. A public report should be created. Simple language. Not "how many times was your paper cited". The NIH should be more like PCORI who funds Implementation projects after solid, large-scale CER studies. If it's clinical practice or public policy/health that needs to be changed as a result of scientific findings, that takes a long time and a lot of money. The NIH needs to do a better job of funding good science from the start, independently reproducing it, then ensuring its implementation. I think the NIH needs to play a more authoritative role in telling researchers what metrics are important. It's not just manuscripts or publishing something with little impact because an RPPR.

6. Foundational Scientific Exploration: This is challenging because I agree but who gets funded seems biased in favor of researchers or students from prestigious universities or a strong track record. Maybe the NIH needs to be more involved and check in more quickly and be able to decide to pull funding sooner if a project doesn't seem to be yielding results.

7. Public Impact: Share the public benefit publicly! I'm a researcher and fully believe it in, but I agree that the ROI isn't clear. I've reviewed too many grants and too many journal manuscripts (not necessarily NIH-funded) that are just crappy or minimally useful science and/or with unclear or minimal clinical/public health impact. Researchers need to stop trying to oversell the importance or impact of their results if their study is only n=10 and cross-sectional, for instance. Uncertainty/confidence intervals need to be clearly communicated.

I think the NIH needs to drive more of the questions researchers ask -- e.g., health dollars spent, unmet public needs (e.g., mental health), premature death, health disparities.

8. Other Areas: The NIH needs to switch to a 2-phased scientific review: Phase 1: review a 1pg letter of intent. Phase 2: only invited LOIs submit a full application. This will save investigators (and grant reviewers) a lot of time and less burn out and stress and false hope. It will likely also save the NIH money.

77. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Healthcare Institution

Role: Investigator/Researcher

Date Submitted: 7/2/2026

1. Rigor and Reproducibility: Funding is needed to conduct investigations that reproduce or replicate the work of other investigators. However, the peer-review system rewards innovation, so obtaining funding to repeat the work of others is unlikely. It would also be exorbitantly expensive to repeat every investigation. I recommend that special funding opportunities be created to repeat work that is of particularly high impact - such as the discovery of a new drug or technology with large health consequences. This would require vetting scientific discoveries and selecting those with the highest impact, and perhaps more modest data/evidence that merit replication.

Also, repeating the work of others can be expedited (and hence cost less) if the original data are available. However, archiving and curating data costs money. Thus, funding needs to be available to support the data sharing and management mandates.

2. Data, Software, and Model Sharing: The costs associated with these activities can be significant, and with across-the-board budget cuts, investigators are forced to choose between conducting their proposed research plan to generate new knowledge or managing and sharing data. Free archive infrastructure need to be readily available to make sure data sharing can happen efficiently.

3. Training and Mentorship: Gathering data on how many trainees go on to participate in the biomedical research community is an important way to assess the impact of training efforts. Trainees/mentees who go on to participate in research and development within an industry setting need to be acknowledged, in addition to those who pursue traditional academic routes.

4. Collaboration: Assessing the number of disciplines involved in a project is one metric for team science, though successful programs built around a single discipline should not be penalized if the work can be conducted effectively without an interdisciplinary team.

5. Entrepreneurship and Translation: More credit should be given to patents, especially those that are licensed to commercial entities. Getting new technology into the commercial sector is key to having a dramatic impact on public health.

6. Foundational Scientific Exploration: Journal editors are not keen on publishing many negative findings. So, even though knowing what doesn't work is valuable to the scientific community, programs that experience failure in a high-risk activity have more difficulty in getting their work published in

leading journals, if at all. A culture change is needed to allow short communications of negative findings in the peer-reviewed literature.

7. Public Impact: Rewarding the grants that produce drugs or devices that are translated into clinical practice is essential to ensuring that discoveries impact human health. Measures of health outcomes on a population scale is thus essential. Secondary measures, such as economic growth resulting from discoveries, also help to measure societal impact.

8. Other Areas: The traditional metrics for scientific productivity should not be abandoned, but rather supplemented with metrics that measure commercialization success, translation into clinical practice, and impact on health outcomes for the target disease.

78. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/2/2026

1. Rigor and Reproducibility: As a requirement of NIH awards, authors should be required to include in their acknowledgements the PMIDs of papers with data replicated or data sets used in the new study. NIH could then link these PMIDs to the grants acknowledged in the referenced prior papers, serving as a measure of reproducibility

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: As it stands, there is not a perfect mechanism for tracking collaborations between NIH-funded investigators supported by different awards. Typically, productivity is tracked by publications, which are associated with grants by award identifiers. But often, a collaborating investigator will perform a key experiment that falls outside of the scope of their own grants and is thus not funded by the NIH or attributed to the NIH in the publication. Thus, RPPRs are blind to this collaboration.

NIH should create a formal mechanism to track and reward general collaborative activity between its funded investigators, even if it falls outside the scope of NIH-supported work. For NIH-supported investigators, they should have two categories of research products to report. The current class of research products funded by the NIH award, plus a new class of general collaborative manuscripts not supported by the award. This would allow NIH to capture the grey area of collaboration not directly funded by NIH awards, the bulk of which is currently missed.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

79. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/3/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: Recently, NIH appears to be pulling back from supporting infrastructure such as the model organism databases that have been critical to integrating datasets generated by different groups and making the data readily accessible to all researchers in the field (as well as anyone in the public who would be interested in accessing the data). Rather than ceasing this support, I believe NIH should expand its support to ensure that such databases are well-designed, well-functioning, and appropriately curated by experts in the field. To enhance transparency, data within the databases should be readily traceable to its origin. NIH funding to support the databases can require/reward adherence to best practices for making sure the data are accessible, transparent, and freely available. Large-scale data generation projects can be tied to additional funding appropriate for supporting this dissemination.
For those whose research focuses on method generation, key metrics of success should include not only publication but also adoption of methods by the community over time to analyze data, not simply compare to other methods.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: It should be easier to obtain a small amount of seed funding for new approaches or to expand a lab's research into a new scientific area. Review of such proposals should truly be independent of preliminary data, but instead be based on novelty, rigor of the approach, and potential for real innovation. Early investigators, early-mid career investigators, and small labs should be prioritized for receiving such funding.

7. Public Impact: N/A

8. Other Areas: N/A

80. Gregory Edward Cox

I am responding: On behalf of myself

Name: Gregory Edward Cox

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 7/6/2026

1. Rigor and Reproducibility: Historically, NIH has prioritized novelty when evaluating proposals, but criteria could be amended to explicitly include replication and production of new research tools as sources of intellectual merit for proposals. Note that I am not advocating for new categories of grants. Rather, I'm saying that a proposal that includes an Aim to replicate a prior result---along with a clear rationale for doing so---should be counted as having equivalent intellectual merit to a proposal that includes an Aim to test a novel hypothesis (again, assuming there is a clear rationale for doing so). Similarly for grants including an Aim to develop a new research tool, method, or material.

2. Data, Software, and Model Sharing: Part of the budget granted by NIH funded proposals could be explicitly allocated to costs associated with sharing materials. For example, NIH funds should be able to be used to support open-access publishing, public data repositories, and other mechanisms that enable the sharing of data and software.

3. Training and Mentorship: NIH should support students from a broad range of backgrounds, particularly those underrepresented in STEM fields. Doing so increases the intellectual strength of our scientific workforce because the greater the variety of perspectives and skills within the community, the greater the chance for innovation and discovery. It also fosters transparency and openness, and therefore public trust, by increasing access to scientific knowledge and skills beyond an elite few.

4. Collaboration: N/A

5. Entrepreneurship and Translation: As a federal agency funded by taxpayer dollars, the best function the NIH can serve is to build a robust body of knowledge that can then be rapidly iterated on and applied by private corporations and nonprofit entities. The NIH plays the long game.

6. Foundational Scientific Exploration: Because the NIH does not need to turn a profit, it is ideally situated to support basic research and potentially risky proposals. These should be accorded higher funding priority than proposals that seek to translate existing knowledge into products, since other entities are better suited to accomplish that goal.

7. Public Impact: The NIH makes its public impact known through the research it supports. Therefore, supporting a broad array of initiatives focusing on training new scientists and developing basic knowledge and techniques will ultimately produce the greatest public impact, through new knowledge,

through expanding representation in the sciences, and through the public-facing applications of the research supported by NIH.

8. Other Areas: N/A

81. Katherine Sauder

I am responding: On behalf of myself

Name: Katherine Sauder

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/6/2026

1. Rigor and Reproducibility: Reproducibility requires funding of larger studies with greater hetereogeneity in settings and populations, which has higher costs. More funding mechanisms to support large, multi-site studies are needed.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: Funding opportunities that require both bench and clinical aims will incentivize basic and clinical scientists to collaborate, accelerating the maturation of discoveries.

6. Foundational Scientific Exploration: N/A

7. Public Impact: Funding opportunities to conduct systematic reviews and meta-analyses will incentivize scientists to focus on synthesis of research that leads to practice statements and clinical guidelines. Rather than simple metrics of total publication counts and citation metrics, counts of publications that are cited in systematic reviews, meta-analyses,, practice statements, and clinical guidelines will reveal which publications are most impactful. Counts of publications that are cited in grant applications also captures the nearer term influence of published work (as opposed to the ultimate publication from completed study that cites papers published before it was funded).

8. Other Areas: N/A

82. Donald Allen Williamson, Ph.D.

I am responding: On behalf of myself

Name: Donald Allen Williamson, Ph.D.

Name of Organization: Pennington Biomedical Research Center

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/8/2026

1. Rigor and Reproducibility: The reliability (reproducibility and replicability) of scientific findings is critical for innovation and for public confidence in science and medicine. There are many potential causes of unreliable scientific findings, and many solutions have been proposed over the past two decades. Incentives from federal funding agencies for research initiatives should be defined by the priorities of the funding agency, e.g., as currently expressed by priority scores. Priorities for funding projects can be based (primarily) on the potential scientific significance (including scientific rigor) of the research findings and the experimental design of the proposed study. One broad framework related to experimental design would be to classify proposed studies as: 1. Controlled Experiments or 2. Exploratory Studies. Controlled experiments should have priority over Exploratory Studies because they directly test cause-and-effect relationships. Controlled Experiments should compare two or more interventions (treatment arms), often including a control group, with adequate statistical power to test an explicitly stated experimental hypothesis or primary aim. Controlled Experiments should (ideally) randomly assign participants to treatment arms. Controlled Experiments could include traditional randomized controlled trials (RCTs) as well as quasi-experimental Real-World studies such as case-controlled studies that are often used in econometric research where randomization of participants to interventions is not possible for logistical or ethical reasons. Exploratory Studies typically use correlation designs, which include long-term observational studies and cross-sectional studies. Exploratory Studies can be viewed as having a primary function of generating hypotheses that can then be tested using Controlled Experiments. Cross-validation of the findings of Exploratory Studies could strengthen decisions to progress from Exploratory Studies to Controlled Experiments. There is need for both basic research and applied research, with applied research having priority for NIH. Coordination with the National Science Foundation would facilitate crosstalk and strategic planning for the translation of basic research to tests of practical application. Priority can be provided for proposed studies that replicate (or reproduce) and extend previous findings from controlled experiments. To compare findings across studies, reporting standardized effects sizes for all results should be made mandatory. Publishing the null results of well-designed experiments can also be encouraged. Findings of weak or null effect sizes can facilitate an understanding of failures of replication. The reliability (replicability/reproducibility) of results from a series of studies can be tested via meta-analysis and/or systematic reviews. Utilization of the concept of a fail-safe number of null results could be encouraged to operationally define unreliable findings.

2. Data, Software, and Model Sharing: Sharing data, software, and previous data analysis can be made mandatory for federal funding of research. Priority for clear descriptions of this shared data capacity can be made explicit in NIH policies. Also, explicit descriptions of independent (intervention/control) and dependent (outcome) variables can be prioritized, including specification of the reliability and validity of outcome/dependent measures. To facilitate data sharing and transparency across studies, NIH might seek competitively funded data coordinating centers (DCCs) that could be utilized by independent investigators or multi-site investigations (e.g., DPP Research Group, 2002; Look AHEAD Research Group, 2013). These DCCs can be conceptualized as centers of excellence for experimental expertise ranging from genotyping and phenotyping to measures of behavioral adherence (e.g., Williamson et al, 2010). DCCs could assist in data collection, storage, and data sharing. Making DCCs available to all interested investigators should facilitate collaboration and enhance the likelihood of testing for replication and extension of original findings. It could also reduce the overall costs of conducting scientific research and enhance the transparency of scientific rigor. One of the mandates of DCCs might be to plan studies that facilitate making causal inferences among variables using controlled experimental methods.

3. Training and Mentorship: Funding support for junior investigators should be a continuing priority for NIH. Within each of the NIH Institutes, an executive committee could be formed to establish a long-term strategy for priorities and for the process of attaining objectives related to mentoring. One part of this strategy would be to establish a vision for mentorship of junior scientists and support institutions that will ensure a non-interrupted pipeline of high-quality research and researchers that can be supported by government, foundation, and private sources. This plan could include funding for promising post-doctoral fellows who might later be granted funding as junior investigators. The infrastructure of long-term mentorship should be flexible, including regular meta-science evaluation of the success of the mentoring program using objective measures, e.g., per cent of funded post-doctoral fellows who receive future funding. NIH can also facilitate the training of early career scientists by providing virtual and in-person training related to the design of Controlled Experiments, philosophy of science, and the optimal data analytic strategies for testing a diversity of hypotheses. This training can emphasize the importance of the nuances of experimental design, e.g., the importance of external controls such as control groups or putative mediators. Also, the training could emphasize the significance of internal control methods, e.g., blinding investigators regarding treatment arm assignment of participants, the reliability and validity of outcome measures, the impact of repeated measurement, and the requirements for using multivariate statistical analyses.

4. Collaboration: Current research initiatives are often complex and require considerable scale, e.g., large sample sizes. These factors are especially relevant for Controlled Experiments that address significant health concerns of a population. For these reasons, priority for highly qualified multi-disciplinary teams of scientists is important. Peer review panels could evaluate research teams, research support institutions, and proposed research plans, utilizing the principles of priority that have been described. These review panels could include senior and mid-career scientists as well as public information advocates that can address the public response to the program of study and its hypothesized results. In some areas of health-related research, the body of knowledge may be quite advanced. In such cases, the need for large-scale, national scope, multi-site studies can be prioritized, Funding for such multi-site projects can be “carved-out” of the overall budget of an NIH Institute so that investigators will be motivated to establish collaborative networks that are likely of work together in a cooperative manner (e.g., DPP Research Group, 2002; Look AHEAD Research Group, 2013). In such

studies, priority should be given to studies that test an explicit mechanism of action. Such experimental tests might be guided by the precision of experimental design, e.g., the design of intervention and control arms, or using mediation analysis (e.g., Williamson et al, 2009) that tests for covariation within and across intervention/control arms. Such mediation analysis might hypothesize genotypic or phenotypic variables, that include factors such as adherence to the intervention. Identification of primary mechanisms of action from meditation analyses can inform the development of hypotheses for future basic or applied controlled experiments.

5. Entrepreneurship and Translation: In addition to promoting Controlled Experiments, NIH could prioritize applied health-related research. Collaboration with NSF and other public and private institutions could involve strategic planning to link basic health-related research and applied research. The long-term strategy for applied research projects would include proof-of-concept studies, scaled-up Real-World Experiments (e.g., Horstmann et al, 2021), and innovative research development projects that are likely to yield commercially viable health enhancement programs. This pipeline of applied research could include pharmacological and non-pharmacological interventions. In this context, technological innovations such as artificial intelligence (AI) and social media could be viewed as tools for delivering/disseminating specific health enhancement interventions, e.g., how the internet and streaming media were tested as innovative delivery strategies in the first two decades of this century. From this perspective, the benefits and harms of these tools could be evaluated in terms of specific objectives, e.g., prevention of type 2 diabetes or suicide, and unintended negative consequences could be objectively studied using experimental controls. For example, a control group could be used to evaluate the natural historical occurrence of adverse events associated with or without AI or social media delivery strategies. If studies used similar research designs, e.g., RCTs, and comparable interventions and outcome measures, meta-analyses and systematic literature reviews could confidently test the reliability, reproducibility, and strength of the benefits and harms of innovative interventions. It may be necessary to establish NIH-supported Centers of Excellence across research domains to systematically collate such data for meaningful results from meta-analysis and reviews and to explain the significance of the conclusions to the public. Centers of Excellence could enhance the likelihood that null results of well-designed RCTs were available for inclusion in meta-analyses. This comprehensive strategy would enhance the probability that innovative health-related interventions were empirically proven to be effective, safe, and available to the public. This process could be illuminated by public health advocates so that public trust in science and medicine can be restored.

6. Foundational Scientific Exploration: There can be no doubt that basic research and exploratory research forms the foundation for high-quality applied research. For this reason, prioritized funding for basic and exploration research should be maintained. A question for NIH is who the primary source of funding for basic research might be. There are many potential funding sources for different types of research, ranging from NSF, USDA, DOD, foundations, and private institutions. Coordination among these public and private agencies will be critical to produce streamlined coordination and relative lack of overlap. Another potential solution might be to plan critical basic and exploration research in the internal laboratories of NIH Institutes. For example, the hypothesized targets of interventions to prevent/delay the onset of cognitive impairment of senior citizens might originate from the National Institute of Aging. The Centers of Excellence may also play a role in the planning and development of such research.

7. Public Impact: Translation of proof-of-concept research into population-based Real-World strategies that enhance the health of entire communities will be essential to ensure maximal public impact and a widespread trust in the integrity and utility of federally funded research initiatives. One strategy to attain this objective will be to carve out and prioritize Real-World quasi-experimental studies such as case-controlled trials that are common in economics, climatology, and other fields with practical constraints on randomization of participants to treatment arms. To preserve data analytic integrity, such studies must be careful to utilize cluster randomization strategies when individuals are naturally grouped together (e.g., students in a school or patients in a clinic). Public education about the strengths and weaknesses of different types of experimental designs and the implication of weak effect sizes is badly needed to disentangle confusion about what is often interpreted by experts and the public as conflicting results. There is need for continuing public education about complex scientific concepts, e.g., the probabilistic nature of empirical knowledge. Good targets for scientific education include making logical errors, e.g., making causal inferences from correlations, significance of standardized effect sizes, and the limitations of statistical controls. Another important educational distinction will be to differentiate scientific results that are based on empirical observations versus results that are based on statistical modelling of future events. Such statistical modelling should be viewed as Exploratory Studies that generate hypotheses that can be tested using Controlled Experiments. For maximal public impact, there is a need to systematically evaluate and collate a research base using unbiased methodology. Over the past two decades great strides in objectively grading the quality of research using consensus-based rules for judging the quality and bias of research studies. These grading systems often yield a letter grading format, e.g., A to F, that is easily understood by the public. This type of evaluation could be managed by Centers of Excellence or other expert panels that include public information advocates. If these strategies could be systematically employed, public trust in science and health care could benefit.

8. Other Areas: Generalization of the results of Controlled Experiments to the general population is very important. The experimental design of prioritized experiments should include internal control protocols for recruitment of participants that are representative of the population of interest. These methods do not necessarily assume a uniform response to interventions across individuals or sub-groups of people. Large-scale studies can be designed to test the relative efficacy and safety of an intervention for different types of people by inclusion of planned evaluations of the impact of the intervention on specified sub-groups, Sub-group analyses must be planned from participant recruitment through the intervention phase(s) to ensure that adequate numbers of sub-group members are recruited and retained. Statistical power analysis for sub-group studies can be used to guide this effort. Priority can be given to proposals that plan to study demographic and non-demographic variables. Ideally, these non-demographic variables can be operationally defined as dimensions, i.e., not categories. Sub-groups can be viewed as moderator variables in the overall analysis of benefits and harms of an intervention. Examples of different moderators include genotypes and phenotypes as well as behavioral factors, and social factors, e.g., level of education. Evaluation of protocols should include protection of human subjects, e.g., safety, privacy, and ethics.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/LA-2013-CVD-effects.pdf

Description: Experimental Design is Key to Enhanced Scientific Impact

NIH Public Access
Author Manuscript
N Engl J Med. Author manuscript; available in PMC 2014 July 11.

Published in final edited form as:
N Engl J Med. 2013 July 11; 369(2): 145–154. doi:10.1056/NEJMoa1212914.

Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes

The Look AHEAD Research Group

Abstract

BACKGROUND—Weight loss is recommended for overweight and obese individuals with type 2 diabetes based on short-term studies, but long-term effects on cardiovascular disease remain unknown. We examined whether intensive lifestyle intervention for weight loss decreased cardiovascular morbidity and mortality in overweight or obese adults with type 2 diabetes.

METHODS—We randomly assigned 5,145 overweight or obese individuals with type 2 diabetes recruited at 16 US centers to intensive lifestyle intervention (the intervention group), which promoted weight loss through decreased calorie intake and increased physical activity, or diabetes support and education (the control group). The primary outcome was the first post-randomization occurrence of a composite cardiovascular outcome (cardiovascular death, non-fatal myocardial infarction, non-fatal stroke, or hospitalized angina) over a planned maximum follow-up of 13.5 years.

RESULTS—The trial was stopped early based on a futility analysis when median follow-up was 9.6 years. Weight loss was greater in the intervention group than the control group throughout (8.6% vs. 0.7% at 1 year; 6.0% vs. 3.5% at study end). Intensive lifestyle intervention also produced greater reductions in hemoglobin A1c and greater initial improvements in fitness and all cardiovascular risk factors, except LDL cholesterol. The primary outcome occurred in 403 patients in the intervention group and in 418 in the control group (1.83/100 person-years and 1.92/100 person-years, respectively; hazard ratio 0.95; 95% CI 0.83 to 1.09, p=0.505).

CONCLUSION—In our study, intensive lifestyle intervention focused on weight loss did not reduce cardiovascular events in overweight or obese adults with type 2 diabetes. (Funded by the Department of Health and Human Services and others; ClinicalTrials.gov number, NCT00017953.)

Weight loss is recommended for overweight and obese individuals with type 2 diabetes.[1] This recommendation is based on short-term studies showing numerous benefits of weight loss, including improvements in glycemic control, cardiovascular disease risk factors, quality of life and other obesity-related comorbidities.[2] However, it is unknown whether weight loss reduces the risk of cardiovascular morbidity and mortality in individuals with type 2 diabetes. Epidemiological studies in individuals with diabetes provide conflicting results, perhaps due to confounding from unintentional weight loss.[3] A recent meta-analysis[4] of cohort studies concluded that moderate intentional weight loss was associated with reduced mortality in “unhealthy” individuals, which included individuals with diabetes. The Swedish Obesity Subjects (SOS) study (ClinicalTrials.gov number NCT01479452) reported lower cardiovascular event rates over a mean follow-up of 13.3 years in patients with type 2 diabetes who underwent bariatric surgery;[5] however, this was a non-randomized study, and the results achieved through surgery cannot be generalized to other approaches to weight loss.

Thus, a critical question remains: Does an intensive lifestyle intervention designed to produce weight loss through caloric restriction and increased physical activity decrease cardiovascular morbidity and mortality in overweight and obese adults with type 2 diabetes?

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Look AHEAD (Action for Health in Diabetes) addressed this question in a multicenter randomized clinical trial.

METHODS

Study Design

The study methods have been published previously.[6, 7] Look AHEAD was a randomized trial conducted at 16 clinical sites in the United States (see the Supplementary Appendix). The trial was designed and conducted by the authors, and all analyses were completed by the coordinating center. The study was approved by institutional review boards at each center. The trial was not blinded, but clinical assessors and end point adjudicators were masked to treatment assignment. The authors vouch for the accuracy and completeness of the data and all analyses and for the fidelity of this report to the trial protocol, which is available with the full text of this article at NEJM.org.

The study was sponsored by the National Institutes of Health, with additional support from other federal partners and the clinical research centers of several participating institutions. Major contributions were provided by the FedEx Corporation; Health Management Resources; LifeScan, Inc., a Johnson & Johnson Company; Nestle HealthCare Nutrition, Inc.; Hoffmann-La Roche Inc.; Abbott Nutrition; and Unilever North America. None of the corporate sponsors had any role in the trial design, data analysis, or reporting of results.

Participants

To be eligible for participation in the trial, volunteers were required to be 45 to 76 years of age and to have all of the following: self-reported type 2 diabetes verified by use of glucose-lowering medication, physician report, or glucose levels; body-mass index (BMI) of 25.0 kg/m^2 or greater (27 kg/m^2 or greater in individuals taking insulin); hemoglobin A1c less than 11%; systolic blood pressure less than 160 mmHg; diastolic blood pressure less than 100 mmHg; triglycerides less than 600 mg/dl; ability to complete a valid maximal exercise test suggesting it was safe to exercise; and a primary care provider. Participants could be using any type of glucose-lowering medication, but the percentage of participants using insulin was limited to less than 30%. Participants with and without a history of cardiovascular disease were included to increase generalizeability of results. Additional eligibility criteria are described elsewhere [6] and in the Supplementary Appendix.

Interventions

Eligible individuals were randomized to either an intensive lifestyle intervention (the intervention group) or diabetes support and education (the control group); randomization was stratified by clinical site. Curricula for both treatment arms were developed centrally and have been described in detail [6, 8] (see the Supplementary Appendix). Intensive lifestyle intervention aimed at achieving and maintaining at least a 7% weight loss by focusing on reduced caloric intake and increased physical activity. The program included both group and individual counseling sessions, occurring weekly during the first six months, with decreasing frequency over the course of the trial. Specific intervention strategies included a calorie goal of 1200 to 1800 kcal/day (with less than 30% of calories from fat and more than 15% from protein), use of meal replacement products, and at least 175 minutes per week of moderate intensity physical activity. A toolbox of strategies was available for participants having difficulty achieving the weight loss goals.

Diabetes support and education featured three group sessions per year focused on diet, exercise, and social support during years 1 to 4. In subsequent years the frequency was reduced to one session annually.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

All medication adjustments were made by the participant's health care provider, with the exception of temporary changes in glucose-lowering medications made by study staff to reduce the risk of hypoglycemia in the intensive lifestyle intervention group. Participants and their health care providers received annual reports on the participants' updated cardiovascular risk factors and the goals recommended by the American Diabetes Association.[1]

Assessments and Outcomes

Certified staff, masked to participants' intervention assignment, measured weight, waist circumference and blood pressure, assessed medication use, and obtained blood for analysis at the central laboratory annually.[6] Maximal exercise tests were performed on the full cohort prior to randomization. Submaximal exercise tests were performed on the full cohort at years 1 and 4 and on a subset at year 2.

During annual visits and 6-month phone calls, participants were queried by staff masked to treatment assignment about all medical events and hospitalizations. These queries were augmented with searches of national databases on deaths. Hospital and other records were obtained for potential cardiovascular events and adjudicated according to standard criteria by reviewers masked to intervention assignment (see Supplementary Appendix).

The primary end point was the first post-randomization occurrence of a composite cardiovascular outcome. Initially, the composite cardiovascular outcome included cardiovascular death, nonfatal myocardial infarction, and nonfatal stroke, and the anticipated maximal follow-up interval was 11.5 years. During the first two years of the trial, the primary event rate in the control group was lower than expected.[9] Therefore, hospitalized angina was added to the primary outcome and planned follow-up was extended to a maximum of 13.5 years. Three composite secondary cardiovascular outcomes were also examined: cardiovascular death, myocardial infarction, or stroke (the original primary outcome); death (all causes), myocardial infarction, stroke, or hospitalized angina; and death (all causes), myocardial infarction, stroke, hospitalized angina, coronary artery bypass grafting, percutaneous coronary angioplasty, hospitalization for heart failure, or peripheral vascular disease.

Statistical Analyses

With a planned sample size of 5000 participants, Look AHEAD had greater than 80% probability of detecting an 18% difference in major cardiovascular events between the two groups, assuming a two-sided alpha level of 0.05, a primary outcome rate of 2% per year in the control group, and a planned maximum follow-up of 13.5 years. The 18% difference was chosen based on reductions in mortality in individuals with type 2 diabetes with voluntary weight loss in an observational study,[10] effect sizes chosen for trials with similar outcomes,[11] feasibility, and public health significance.

On September 14, 2012, Look AHEAD's primary sponsors, based on futility analysis and recommendation of the data and safety monitoring board, instructed the study investigators to terminate the intervention. All data analyses extend to this date. At that time, the probability of observing a significant positive result at planned end of follow-up, assuming the actual hazard ratio to be 0.82, was estimated to be 1%.

Baseline characteristics and key safety outcomes were compared between study groups using chi-square, Fisher's exact, Wilcoxon rank-sum, two-sample t-tests and Poisson regression. Physical and laboratory measurements and medication use from baseline through 10 years were modeled with generalized linear regression and generalized estimating equations. Center was included as a covariate, covariance was unstructured, and linear

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

contrasts were used to compare groups throughout post-baseline follow-up and at annual visits. Analyses of primary and secondary outcomes were performed using time-to-event methods according to the intention-to-treat principle as pre-specified in the protocol. Kaplan–Meier estimates were used to calculate the proportion of participants who had an event over time. First occurrences of primary and secondary outcomes in the two groups were compared with hazard ratios and 95% confidence intervals. Two-sided p-values were calculated with likelihood-ratio tests from Cox proportional-hazards regression, with models containing terms for clinical site, history of cardiovascular disease, and intervention assignment. The consistency of intervention effects on the primary outcome among three pre-specified subgroups (sex, race or ethnicity and history of cardiovascular disease) was evaluated using interaction tests. Results were not adjusted for multiple comparisons and $p<0.05$ was used. Statistical analyses were conducted using S-Plus software, version 8.0 (Insightful) or SAS software, and version 9.1 (SAS Institute).

RESULTS

Study Participants

Between August 2001 and April 2004, a total of 5145 participants were enrolled and randomly assigned to intensive lifestyle intervention (N=2,570) or to diabetes support and education (N=2,575) (Supplementary Appendix Fig. S1). Groups were similar at baseline (Table 1). The average age was 58.7 years, 59% were women, and the mean BMI was 36.0 kg/m^2. The median duration of diabetes was 5 years, and 14% reported a history of cardiovascular disease. Additional baseline data have been published previously.[12]

When the intervention was stopped on September 14, 2012, the median follow-up was 9.6 years (interquartile range 8.9 to 10.3) and fewer than 4% of all randomized participants had been lost to follow-up.

Weight, Waist Circumference and Fitness

Intensive lifestyle intervention produced significantly greater changes than diabetes support and education in weight, waist circumference and fitness (Fig. 1A-C) and Supplementary Appendix Table S1). Differences in mean weight loss were greatest at 1 year (8.6% in the intervention group vs. 0.7% in the control group) but remained statistically significant throughout the trial. When intervention ended, mean weight loss from baseline was 6.0% in the intervention group compared with 3.5% in the control group.

Cardiovascular Disease Risk Factors

During the first year of follow-up, intensive lifestyle intervention produced greater improvements than diabetes support and education in all measured cardiovascular risk factors except low-density lipoprotein (LDL) cholesterol (Fig. 1D and Supplementary Appendix Fig. S2 and Table S1). The difference in cardiovascular risk factors between intervention groups diminished over time, with hemoglobin A1c and systolic blood pressure showing the most sustained differences. LDL cholesterol levels were lower in the control group than the intervention group (mean difference of 1.6 mg/dl over the 10 years). Use of antihypertensive medications, statins and insulin was lower in the intervention group than in the control group (Supplementary Appendix Fig. S2 and Fig. S3 and Table S1).

Clinical Outcomes

The composite primary outcome of the first occurrence of hospitalized angina, nonfatal myocardial infarction, nonfatal stroke, or death from cardiovascular causes occurred in 821 participants (403 in the intensive lifestyle intervention group and 418 in the diabetes support and education group). Rates were not different between groups (1.83/100 person-years and

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

1.92/100 person-years, respectively; hazard ratio =0.95; 95% CI 0.83 to 1.09, p=0.505) (Table 2 and Fig. 2). There were also no significant differences between the two trial groups with respect to the pre-specified composite secondary outcomes or any of the individual cardiovascular events making up the composite outcomes (Table 2). There were no significant interactions among the pre-specified subgroups (Fig. 3).

Adverse Events

Severe hypoglycemia, gallstones, fractures, amputations and congestive heart failure were monitored, as plausibly being affected by intensive lifestyle intervention. Although the rate of self-reported fractures differed significantly between groups (2.16/100 person-years in the control group versus 2.51/100 person-years in the intervention group; p=0.011), the incidence rate for adjudicated fractures did not (1.64/100 person-years and 1.66/100 person-years, respectively; p=0.833) (Supplementary Appendix Table S2).

DISCUSSION

The Look AHEAD trial compared the effect of an intensive lifestyle intervention with a control regimen of diabetes support and education in overweight and obese participants with type 2 diabetes. At a median follow-up of almost 10 years, there was no significant difference between the two groups in cardiovascular morbidity and mortality.

Look AHEAD showed that overweight and obese adults with type 2 diabetes can lose weight and maintain modest weight losses over 10 years. Multi-component lifestyle interventions in clinical trial settings typically produce initial weight losses of 7 to 10%;[13, 14] with maximal weight loss at 1 year, followed by gradual regain.[13-15] However, few studies have provided ongoing intervention for an extended time.[16] In our trial, the initial mean weight loss in the intensive lifestyle intervention group was 8.6%. This was followed by weight regain through year 5 and then a subsequent gradual decrease in weight, resulting in an average weight loss of 6.0% at the end of the trial. The control group experienced a gradual but consistent weight loss throughout the study, resulting in an average weight loss of 3.5% at end of the trial. The intervention group also had greater improvements in fitness, particularly at 1 year.

We have considered several possible explanations for the lack of a significant difference in the cardiovascular event rates between groups. One possibility is that the study lacked sufficient power. We do not feel this explains the negative result; the 95% confidence interval for the primary outcome excluded the benefit of 18% or more targeted in the trial's design. Another possibility is that sustained weight losses larger than those achieved in Look AHEAD may be required to reduce the risk of cardiovascular disease. In this regard, it is noteworthy that the differential weight loss between the two trial groups averaged 4% over the course of the study but only 2.5% at the end. However, our trial was planned to test the effects of intensive lifestyle intervention and the weight losses achieved in the intervention group are representative of the best that has been achieved with current lifestyle approaches. Third, educational sessions provided to the control group assigned to diabetes support and education, the greater use of statinsin the control group, and the overall intensification of medical management of cardiovascular risk factors[17] may have made a relative benefit of intensive lifestyle intervention more difficult to demonstrate. The intervention may also have had different effects in different subgroups. Although none of the interactions with subgroups were statistically significant, our data suggest that the event rate for the primary outcome was non-significantly lower in ILI than in DSE among individuals with no history of CVD at baseline, but non-significantly higher in ILI than in DSE among individuals with CVD at baseline.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

There are several limitations to these findings. We used a specific lifestyle intervention that focused on achieving weight loss through caloric restriction and increased physical activity. It is unclear whether an intervention focused on changes in dietary composition, for example the Mediterranean diet,[18] might yield different outcomes. In addition, Look AHEAD recruited participants with type 2 diabetes who were motivated to lose weight through lifestyle intervention and who could successfully complete a maximal fitness test at baseline. Thus, the results cannot be generalized to all individuals with type 2 diabetes.

The finding that intensive lifestyle intervention, compared with diabetes support and education, did not reduce the risk of cardiovascular morbidity and mortality must be considered within the context of other positive effects observed with this intervention. Intensive lifestyle intervention produced clinically meaningful improvements in hemoglobin A1c, especially over the first year of the intervention, that were at least partly sustained throughout follow-up. This positive effect is noteworthy since it may explain why participants in the intervention group were less likely to be treated with insulin over this period. Further, we recently reported that individuals in the intervention group of Look AHEAD were more likely to experience partial diabetes remission during the first four years of the trial.[19] Other benefits identified during the early years of the trial have included improvements in urinary incontinence,[20] sleep apnea,[21] depression,[22] quality of life,[23] physical function,[24] and mobility. Intensive lifestyle intervention has also been shown to prevent or delay the development of type 2 diabetes in other studies.[15, 26]

In conclusion, the Look AHEAD trial found that intensive lifestyle intervention did not reduce the risk of cardiovascular morbidity or mortality compared with a control program of diabetes support and education in overweight and obese participants with type 2 diabetes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Look AHEAD Research Group acknowledges the contributions of Drs. Frederick Brancati and Richard Rubin, who passed away during the completion of this manuscript. Both colleagues were fundamental to the successful development and execution of our trial.

Appendix

Authors

Rena R.Wing, Ph.D.; Paula Bolin, R.N., M.C.; Frederick L.Brancati, M.D., M.H.S.#; George A.Bray, M.D.; Jeanne M.Clark, M.D., M.P.H.; MaceCoday, Ph.D.; Richard S.Crow, M.D#; Jeffrey M.Curtis, M.D., M.P.H.; CaitlinEgan, M.S.; Mark A.Espeland, Ph.D.; MaryEvans, Ph.D.; John P.Foreyt, Ph.D.; SiranGhazarian, M.D.; Edward W.Gregg, Ph.D.; BarbaraHarrison, M.S.; Helen P.Hazuda, Ph.D.; James O.Hill, Ph.D.; Edward S.Horton, M.D.; Van S.Hubbard, M.D., Ph.D.; John M.Jakicic, Ph.D.; Robert W.Jeffery, Ph.D.; Karen C.Johnson, M.D., M.P.H.; Steven E.Kahn, M.B., Ch.B.; Abbas E.Kitabchi, Ph.D., M.D.; William C.Knowler, M.D., Dr.P.H.; Cora E.Lewis, M.D., M.S.P.H.; Barbara J.Maschak-Carey, M.S.N., C.D.E.; Maria G.Montez, R.N., M.S.H.P., C.D.E.; AnneMurillo, B.S.; David M.Nathan, M.D.; JenniferPatricio, M.S.; AnnePeters, M.D.; XavierPi-Sunyer, M.D.; HenryPownall, Ph.D.; DavidReboussin, Ph.D.; JudithRegensteiner, Ph.D.; Amy D.Rickman, Ph.D., R.D., L.D.N.; Donna H.Ryan, M.D.; MonikaSafford, M.D.; Thomas A.Wadden, Ph.D.; Lynne E.Wagenknecht, Dr.P.H.; Delia S.West, Ph.D.; David F.Williamson, Ph.D.; Susan Z.Yanovski, M.D.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Look AHEAD Funding and Support

This study is supported by the Department of Health and Human Services through the following cooperative agreements from the National Institutes of Health: DK57136, DK57149, DK56990, DK57177, DK57171, DK57151, DK57182, DK57131, DK57002, DK57078, DK57154, DK57178, DK57219, DK57008, DK57135, and DK56992. The following federal agencies have contributed support: National Institute of Diabetes and Digestive and Kidney Diseases; National Heart, Lung, and Blood Institute; National Institute of Nursing Research; National Center on Minority Health and Health Disparities; NIH Office of Research on Women's Health; and the Centers for Disease Control and Prevention. This research was supported in part by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases. The Indian Health Service (I.H.S.) provided personnel, medical oversight, and use of facilities. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the I.H.S. or other funding sources.

Additional support was received from The Johns Hopkins Medical Institutions Bayview General Clinical Research Center (M01RR02719); the Massachusetts General Hospital Mallinckrodt General Clinical Research Center and the Massachusetts Institute of Technology General Clinical Research Center (M01RR01066); the University of Colorado Health Sciences Center General Clinical Research Center (M01RR00051) and Clinical Nutrition Research Unit (P30 DK48520); the University of Tennessee at Memphis General Clinical Research Center (M01RR0021140); the University of Pittsburgh General Clinical Research Center (GCRC) (M01RR000056), the Clinical Translational Research Center (CTRC) funded by the Clinical & Translational Science Award (UL1 RR 024153) and NIH grant (DK 046204); the VA Puget Sound Health Care System Medical Research Service, Department of Veterans Affairs; and the Frederic C. Bartter General Clinical Research Center (M01RR01346).

The following organizations have committed to make major contributions to Look AHEAD: FedEx Corporation; Health Management Resources; LifeScan, Inc., a Johnson & Johnson Company; OPTIFAST® of Nestle HealthCare Nutrition, Inc.; Hoffmann-La Roche Inc.; Abbott Nutrition; and Slim-Fast Brand of Unilever North America.

Some of the information contained herein was derived from data provided by the Bureau of Vital Statistics, New York City Department of Health and Mental Hygiene.

REFERENCES

1. American Diabetes Association. Clinical Practice Recommendations. American Diabetes Association; Alexandria, VA: 2013.
2. NHLBI. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults-The evidence report. Obes Res. 1998; 6(S2):51S–210S. [PubMed: 9813653]
3. Williamson DF. Weight loss and mortality in persons with type-2 diabetes mellitus: A review of the epidemiological evidence. Exp Clin Endocrinol Diabetes. 1998; 106(Suppl 2):14–21. [PubMed: 9792476]
4. Harrington M, Gibson S, Cottrell RC. A review and meta-analysis of the effect of weight loss on all-cause mortality risk. Nurt Res Rev. 2009; 22(1):93–108.
5. Romeo S, Maglio C, Burza MA, et al. Cardiovascular events after bariatric surgery in obese subjects with type 2 diabetes. Diabetes Care. 2012; 35(12):2613–7. [PubMed: 22855732]
6. Look AHEAD Research Group. Look AHEAD: Action for Health in Diabetes: Design and methods for a clinical trial of weight loss for the prevention of cardiovascular disease in type 2 diabetes. Control Clin Trials. 2003; 24:610–28. [PubMed: 14500058]
7. Look AHEAD Action for Health in Diabetes. Accessed at https://www.lookaheadtrial.org

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

8. Look AHEAD Research Group. The Look AHEAD study: A description of the lifestyle intervention and the evidence supporting it. Obesity (Silver Spring). 2006; 14(5):737–52. [PubMed: 16855180]
9. Brancati FL, Evans M, Furberg CD, et al. Midcourse correction to a clinical trial when the event rate is underestimated: the Look AHEAD (Action for Health in Diabetes) Study. Clin Trials. 2012; 9(1): 113–24. [PubMed: 22334468]
10. Williamson DF, Thompson TJ, Thun M, Flanders D, Pamuk E, Byers T. Intentional weight loss and mortality among overweight individuals with diabetes. Diabetes care. 2000; 23(10):1499–504. [PubMed: 11023143]
11. Buse JB, Bigger JT, Byington RP, et al. Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial: design and methods. Am J Cardiol. 2007; 99(12A):21i–33i.
12. Bray G, Gregg E, Haffner S, et al. Baseline characteristics of the randomized cohort from the Look AHEAD (Action for Health in Diabetes) study. Diab Vasc Dis Res. 2006; 3(3):202–15. [PubMed: 17160917]
13. Wing, RR. Behavioral approaches to the treatment of obesity. In: Bray, G.; Bouchard, C., editors. Handbook of Obesity: Clinical Applications. 2nd ed.. Marcel Dekker, Inc.; New York: 2004.
14. Wadden TA, Webb VL, Moran CH, Bailer BA. Lifestyle modification for obesity: new developments in diet, physical activity, and behavior therapy. Circulation. 125(9):1157–70. [PubMed: 22392863]
15. Knowler WC, Barrett-Connor E, Fowler SE, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002; 346(6):393–403. [PubMed: 11832527]
16. Knowler WC, Fowler SE, Hamman RF, et al. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. Lancet. 2009; 374(9702):1677–86. [PubMed: 19878986]
17. Ali MK, Bullard KM, Saaddine JB, Cowie CC, Imperatore G, Gregg EW. Achievement of goals in U.S. diabetes care, 1999-2010. N Engl J Med. 2013; 368(17):1613–24. [PubMed: 23614587]
18. Estruch R, Ros E, Salas-Salvado J, et al. Primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med. 2013
19. Gregg EW, Chen H, Wagenknecht LE, et al. Association of an intensive lifestyle intervention with remission of type 2 diabetes. JAMA. 2012; 308(23):2489–96. [PubMed: 23288372]
20. Phelan S, Kanaya AM, Subak LL, et al. Weight loss prevents urinary incontinence in women with type 2 diabetes: results from the Look AHEAD trial. J Urol. 2012; 187(3):939–44. [PubMed: 22264468]
21. Foster GD, Borradaile KE, Sanders MH, Millman R, Zammit G, Newman AB, Wadden TA, Kelley D, Wing RR, Pi-Sunyer FX, Reboussin D, Kuna ST. A randomized study on the effects of weight loss on obstructive sleep apnea among obese patients with type 2 diabetes: the Sleep AHEAD study. Arch Intern Med. 2009; 169(17):1619–26. [PubMed: 19786682]
22. Faulconbridge LF, Wadden TA, Rubin RR, et al. One-year changes in symptoms of depression and weight in overweight/obese individuals with type 2 diabetes in the Look AHEAD study. Obesity (Silver Spring). 2012; 20(4):783–93. [PubMed: 22016099]
23. Williamson DA, Rejeski J, Lang W, Van Dorsten B, Fabricatore AN, Toledo K. Impact of a weight management program on health-related quality of life in overweight adults with type 2 diabetes. Arch Intern Med. 2009; 169(2):163–71. [PubMed: 19171813]
24. Foy CG, Lewis CE, Hairston KG, et al. Intensive lifestyle intervention improves physical function among obese adults with knee pain: findings from the Look AHEAD trial. Obesity (Silver Spring). 2011; 19(1):83–93. [PubMed: 20559303]
25. Rejeski WJ, Ip EH, Bertoni AG, et al. Lifestyle change and mobility in obese adults with type 2 diabetes. N Engl J Med. 2012; 366(13):1209–17. [PubMed: 22455415]
26. Tuomilehto J, Lindstrom J, Eriksson J, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerence. N Engl J Med. 2001; 344:1343–50. [PubMed: 11333990]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

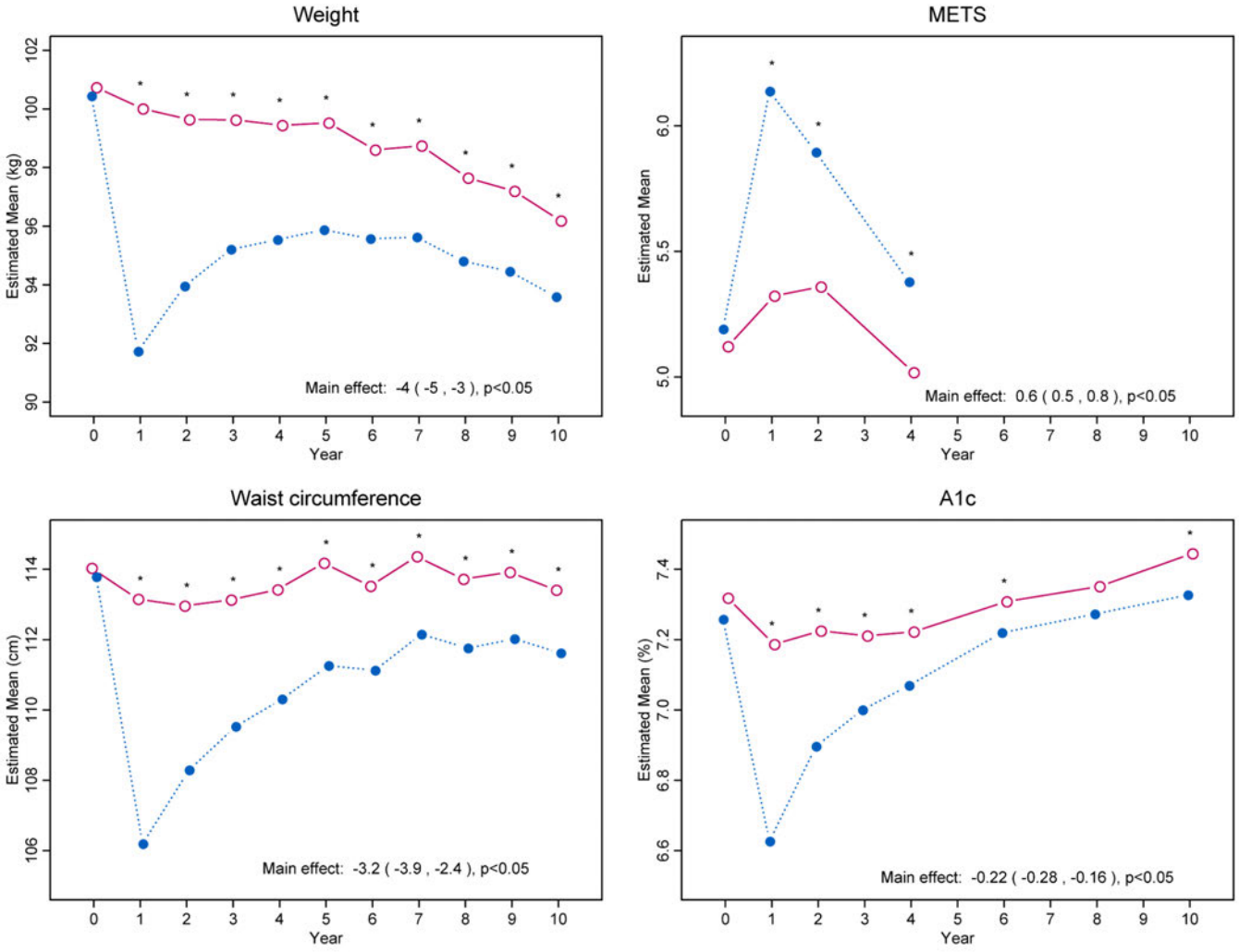


Figure 1.
Changes in Weight, Physical Fitness, Waist Circumference, and Glycated Hemoglobin Levels during 10 Years of Follow-up.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

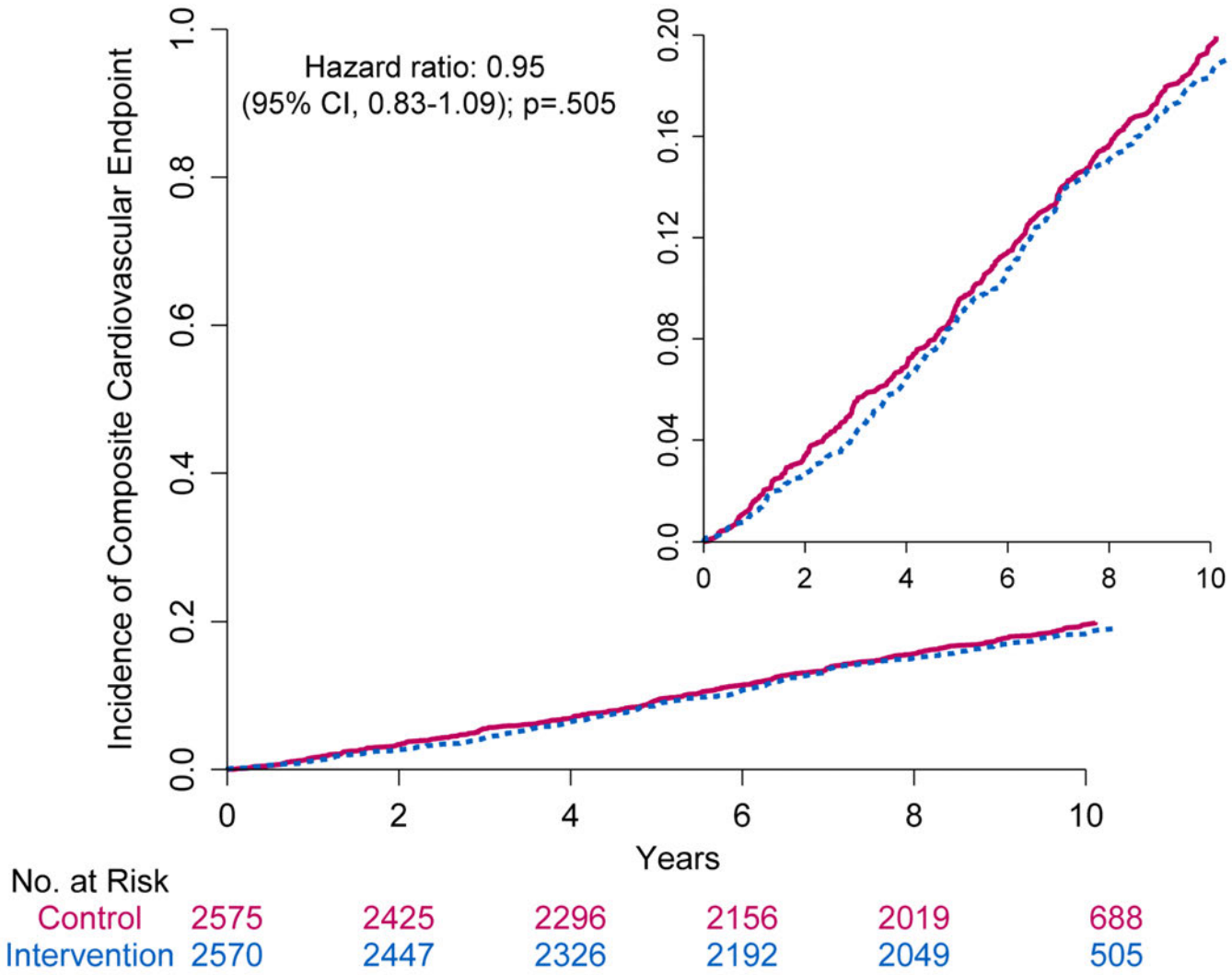


Figure 2.
Cumulative Hazard Curves for the Primary Composite End Point.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

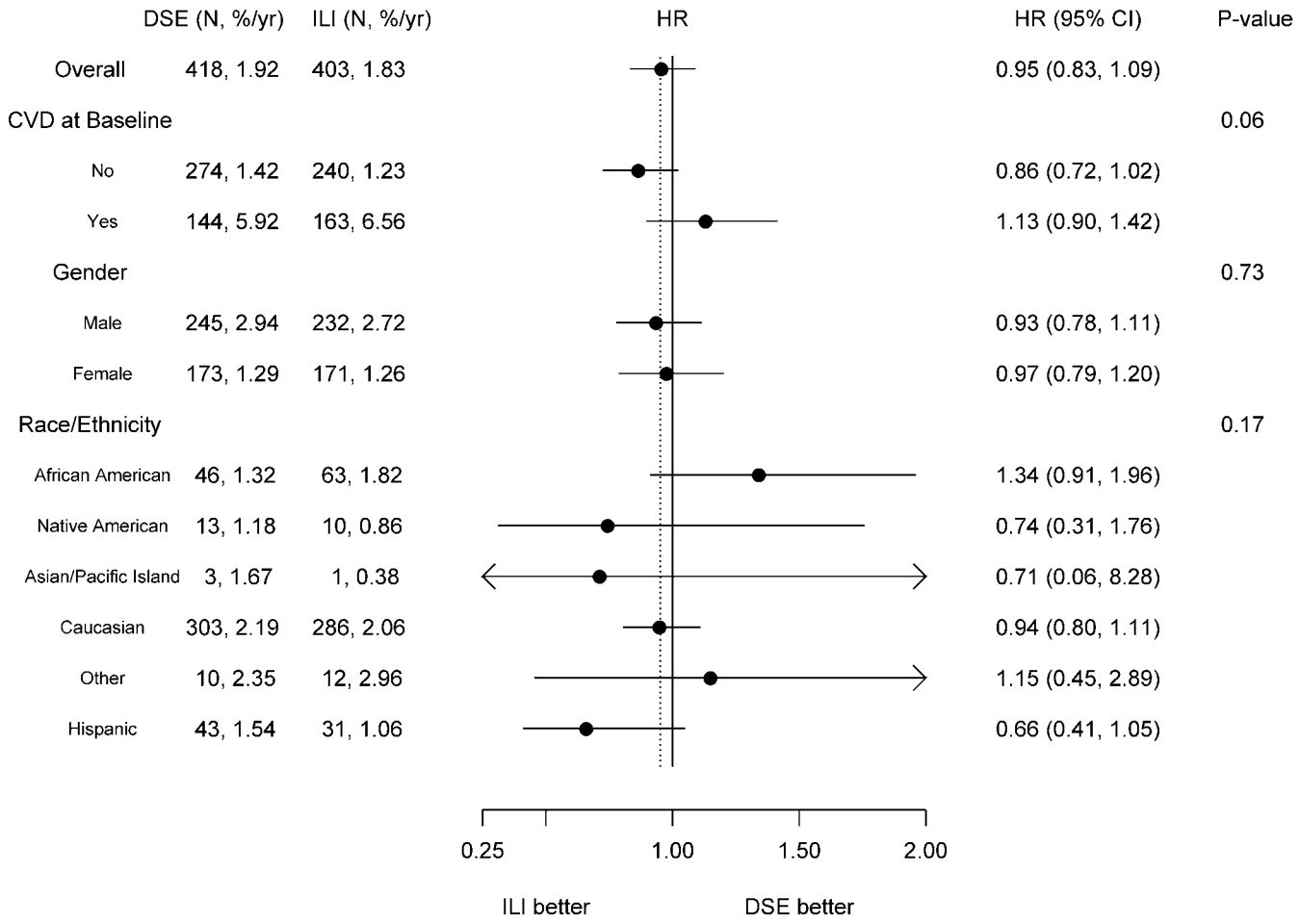


Figure 3.
Primary Outcome in Prespecified Subgroups.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Comparison of intervention groups at baseline.

Variable		DSE	ILI
Age	Mean (SD)	58.9 (6.9)	58.6 (6.8)
Female	N (%)	1537 (59.7)	1526 (59.4)
Race/ethnicity	N (%)		
	African American	404 (15.7)	400 (15.6)
	American Indian	128 (5.0)	130 (5.1)
	Asian/Pacific Islander	21 (0.8)	29 (1.1)
	White	1631 (63.3)	1621 (63.1)
	Hispanic	340 (13.2)	340 (13.2)
	Other/Mixed	51 (2.0)	49 (1.9)
History of CVD *	N (%)	348 (13.5)	366 (14.2)
Using insulin	N (%)	410 (16.5)	382 (15.4)
Smoking	N (%)	110 (4)	117 (5)
Duration of diabetes (years)	Median (Q1,Q3)	5.0 (2.0, 10)	5.0 (2.0, 10)
Weight (kg)	Mean (SD)	101 (19)	101 (20)
BMI (kg/m^2)	Mean (SD)	36.0 (5.8)	35.9 (6.0)
Waist circumference (cm)	Mean (SD)	114 (14)	114 (14)
A1c (%)	Mean (SD)	7.3 (1.2)	7.2 (1.1)
SBP (mm Hg) **	Mean (SD)	129 (17)	128 (17)
DBP (mm Hg)	Mean (SD)	70.4 (9.6)	69.9 (9.5)
HDL-C (mg/dl)	Mean (SD)	43.5 (12)	43.4 (12)
LDL-C (mg/dl)	Mean (SD)	112 (32)	112 (32)
Triglycerides (mg/dl)	Median (Q1,Q3)	152 (107, 218)	155 (110, 221)

ILI = Intensive Lifestyle Intervention. DSE = Diabetes Support and Education.

* Prior CVD was defined as history of myocardial infarction, stroke, congestive heart failure or CVD procedures (CABG, PTCA, carotid endarterectomy, angioplasty of a lower extremity artery, or aortic aneurysm repair).

** indicates significant difference between arms with $p < 0.01$.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 2

Primary and secondary outcomes and other cardiovascular outcomes.

Outcome	N	DSE N, events/100pyrs	ILI N, events/100 pyrs	Hazard ratio, 95% CI, p-value
Primary and secondary outcomes				
Primary: MI, stroke, hospitalized angina and CVD death	821	418, 1.92	403, 1.83	0.95 (0.83, 1.09), p=.505
Secondary 1: MI, Stroke and CVD Death	550	283, 1.25	267, 1.17	0.93 (0.79, 1.10), p=.417
Secondary 2: primary + total mortality	1025	529, 2.43	496, 2.25	0.93 (0.82, 1.05), p=.229
Secondary 3: primary + CABG/PTCA, hospitalized CHF, carotid endarterectomy, PVD and total mortality	1177	600, 2.81	577, 2.67	0.94 (0.84, 1.05), p=.294
Other CV outcomes				
Death	376	202, 0.86	174, 0.73	0.85 (0.69, 1.04), p=.112
Death from any CV cause	109	57, 0.24	52, 0.22	0.88 (0.61, 1.29), p=.520
All MI (fatal and non-fatal)	354	191, 0.84	163, 0.71	0.84 (0.68, 1.04), p=.107
Fatal MI	16	11, 0.05	5, 0.02	0.44 (0.15, 1.26), p=.127
Non-fatal MI	342	183, 0.80	159, 0.69	0.86 (0.69, 1.06), p=.156
Hospitalized Angina	390	196, 0.87	194, 0.85	0.97 (0.80, 1.19), p=.785
All Stroke (fatal and non-fatal)	165	80, 0.34	85, 0.36	1.05 (0.77, 1.42), p=.775
All CHF (fatal and non-fatal)	218	119, 0.51	99, 0.42	0.80 (0.61, 1.04), p=.096
CABG	525	269, 1.21	256, 1.14	0.93 (0.78, 1.10), p=.407

N Engl J Med. Author manuscript; available in PMC 2014 July 11.

83. fleming scientific

I am responding: On behalf of an organization

Name: john Heironimus

Name of Organization: fleming scientific

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Other

Role - Other: founder and inventor of polymathic state space mapping of chaotic hemodynamical turbulence

Date Submitted: 7/8/2026

1. Rigor and Reproducibility: The AI gold rush has created conflicted reviewers across public, private and academic organizations. These reviewers may be economically, academically or emotionally conflicted. They may be indirectly conflicted by the investments in their 401k.

Leading institutions in AI have long histories of anti-competitive behavior as well and the circular economy which represents half of stock market value is well documented. As AI represents the only growing area of the US economy, our government has clearly chosen to look the other way on issues that should have been brought to the DOJ which included reviewer supply chain collusion.

Although WEO 14110 ordered NIH to cease collusion with major technology companies and foster innovation by small businesses, the order was clearly ignored by NIH and GAO. In this environment, it has been all but impossible for small businesses to introduce non AI innovations that are safer (even if preferred by the FDA), more powerful and more efficient/environmentally friendly. Compounding the problem is that many reviewers with clinical research background seem to know the language of AI but not actually what the words mean. Reviewers with computationally background are the most hostile to non AI approaches because they know that AI is a poor option for certain applications like chaotic time series observation.

The greed, dysfunction, misinformation and just plain stupidity is unprecedented in US history. Current resource allocation away from "real science" and to AI pretty well assures that AI will always be needed. However, to use AI, when a physics based model is possible, is to be artificially stupid and not old fashioned. This corruption of scientific process and fiduciary responsibility is not in the best interest of healthcare, science or the American people.

AI is useful when there are no real science alternatives but not as a replacement for real science or math. Our goal should be to replace AI with superior knowledge and not the reverse. I suggest that NIH at least consider DARPA documentation on this topic because they focus on what works and not what makes money for powerful stakeholders. They do not persecute innovators with potentially game changing new ideas.

Perhaps the greatest travesty with AI stakeholders holding all the cards is that it is in their best interest to improve real science to to the point where AI can perform better but not to the point that AI is obsoleted and that is what the future brings.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

84. Annette Christine Frenk Oquendo

I am responding: On behalf of myself

Name: Annette Christine Frenk Oquendo

Name of Organization: The University of Texas Health Science Center at Houston

Type of Organization: Academic Institution

Role: Research Trainee/Fellow

Date Submitted: 7/8/2026

1. Rigor and Reproducibility: The best way to test for replication and reproducibility, is to make sure that there is sufficient funding for it. But this should be in-addition to existing funding, not taking away from other scientists.

2. Data, Software, and Model Sharing: Sure, as long as data is secure and analyzed by experts.

3. Training and Mentorship: I think you need to let mentors mentor, and trainees be trained. You're looking to trim a brisket that's already lean enough.

4. Collaboration: Collaboration is important, as long as it is done by those who are trained and familiar within their fields.

5. Entrepreneurship and Translation: Translational research is important, as long as we're using a scientific and rigorous approach.

6. Foundational Scientific Exploration: High-risk high reward makes sense when there is knowledge and appropriate scientific background. Otherwise you just get a Theranos case.

7. Public Impact: If you care about public trust, please do not allow anti-vaxxer quacks into the high ranks of our scientific leadership. That would really help.

8. Other Areas: N/A

85. Rob Holston

I am responding: On behalf of myself

Name: Rob Holston

Name of Organization: T.H.E. Inc.

Type of Organization: Other

Type of Organization - Other: Health Acvocate

Role: Investigator/Researcher

Date Submitted: 7/8/2026

1. Rigor and Reproducibility: PREMISE

A 2014 Student Growth Screening Survey indicates about 42% of Ketchikan district students, pre-kindergarten through 9th grade, are either overweight or obese. An additional 6.5% of the student population is ranked as severely obese (MORBID OBESITY.) That totals 48.5%. Barbara McCarthy was the Ketchikan School District's Wellness Coordinator and said "….with parental permission, 88 percent of all students enrolled in the Ketchikan School District participated in the survey."

According to the most recent Behavioral Risk Factor Surveillance System (BRFSS) data, adult obesity rates now exceed 35% in 23 states. West Virginia has the highest adult obesity rate at 41.2% and the District of Columbia (D.C.) has the lowest at 23.5%. Between 2022 and 2023, three states experienced statistically significant increases in their adult obesity rate: Alaska, Arkansas, and Oregon. No state experienced a decline. (Assumed fact- "Many Overweight students will become obese adults.")

"Consistently studies show that students who eat well and students who are active will perform better academically, they are more likely to go on to higher education and they have fewer classroom behavior issues" according to McCarthy.

In 2011–2012, approximately 1 in 6 (17.7%) US children aged 6 to 11 were obese. Compared with their non-obese peers, these children are more likely to experience low self-esteem and depression, discrimination by other children, and academic struggles; they are also more likely to become obese adults. The ability to monitor childhood obesity prevalence at the state and local levels is critical for state public health programs to assess trends and the success or failure of interventions.
Obesity can negatively affect a student's academic learning through several pathways, including biological factors like compromised executive function and memory, indirect factors such as increased school absenteeism and bullying, and teacher bias which can lead to lower expectations and grading. This can result in lower grades, poorer test scores in reading and math, and less likelihood of graduating from college.

6.5% of Ketchikan's young students were measured as having MORBID OBESITY. Morbid obesity at an

early age can significantly reduce a child's life expectancy, with some studies suggesting it could be cut by half. This is because severe early-onset obesity is linked to developing serious health conditions like heart disease & type 2 diabetes at an earlier age, increasing the risk of mortality later in life. The impact is profound and highlights the need for early and comprehensive prevention and treatment to improve both the quality and length of life.

A childhood diagnosis of Type 2 diabetes significantly reduces life expectancy due to the long-term impact of high blood sugar, increasing the risk of early and severe complications like heart disease, stroke, and kidney disease. Studies estimate that life expectancy could be reduced by 15 to 19 years or more. Diabesity leads to Chronic Disease and Chronic Disease leads to disabilities, increased insurance claims and premature death.
pastedGraphic.png
Alaska Nutrition, Physical Activity, and Obesity Facts Report- 2025
“For over two decades, the percentage of adult and high school-aged Alaskans with obesity has continued to rise. Physical activity and healthy eating can prevent overweight, obesity, and several chronic diseases, including many cancers, heart disease, and type 2 diabetes. It can also impact mental health. Healthy eating patterns have been associated with a reduced risk of depression. The mental health benefits of physical activity—for adults and children—include better mood, sleep and focus.” ALASKA Physical Activity, Nutrition & Obesity Facts Report.

PROPOSAL
To combat Obesity amongst Alaskan School students:

At the high school and junior high/middle school levels: Reassign all P.E. & health teachers to become Health & P.E. Concierges (HPEC). Their responsibilities will include assisting students and homeroom teachers to introduce, implement and maintain each student’s individual P.E. & Health Journal (PE&HJ), in a manner similar to IEP’s for special needs students. The Journal will remain the property of the student. The monthly markers of change will be shared with each student’s concierge. The concierge along with the student will decide upon a health, diet and physical education program that meets the student where they are and plans for maintenance or improvement in weight, strength, mobility, flexibility, body-fat-%, heart rate, blood pressure, etc. all of which can be measured with a body analysis scale preferably located in each home room and each Health & P.E. Concierge’s office. These concierges will provide students with lists of foods and beverages to avoid and also lists and portions of foods and beverages to consume. This protocol could also be implemented at the elementary school level. Faculty & staff would be encouraged to participate also thus lowering the incidence and complications of chronic disease.

Activity: for 8 minutes each morning and 8 minutes each afternoon all students (& staff?) will participate in a nonstop physical activity; walking, jogging, jumping, 6 count burpee's, jumping jacks, etc. or combinations thereof agreed to by the student with the guidance of their concierge. The idea is to move for 8 minutes and work for improvement in: distance traveled, numbers of repetitions, difficulty of tasks performed, etc.

Compare this to a typical P.E. class on softball where 9 players are on the field and the pitcher and catcher are the only ones even coming close to getting any actual exercise. The OTHER half of the class is in the dugout, SITTING! Consider ever popular dodgeball where the obese slow player is sidelined only to view the athletes continue to get exercise.

Completing the PE&HJ for all four years of H.S. attendance should become the new graduation requirement, replacing the current failed health and P.E. classes requirements most districts now have. An extensive health & lifestyle bibliography could be incorporated into most other subjects of academia: math, English, science, reading, etc.

Please compare this PROPOSAL with ALL other class requirements. There is NO OTHER SUBJECT MATTER taught that has such practical application to each student's life every single day of their lives. It teaches and guides students to make informed choices on what to eat and not eat, what to drink and not drink and whether or not to take the stairs as opposed to the elevator. These are daily choices that are impacting our students multiple times per day and will, daily for the rest of their lives. District staff members should also benefit from the district's focus on health and longevity.

Concierges will enter weights & body fate % data into a spreadsheet to calculate school wide improvements, while protecting each student's identity. Please realize that this plan is subject to changes and improvements but also realize that the existing curriculum is NOT solving the problems of student obesity.

Alaska School Districts are encouraged to abandon their existing Health and P.E. curriculums and replace the present 1/2 semester each of Health & P.E. as graduation requirements with the requirement to complete 4 years of a student's PE&HJ (P.E. & HEALTH JOURNAL).

I'm suggesting that the Alaska School Boards and the Alaska Department of Education & Early Development take initial steps to address this issue and form task force committees to examine the PREMISE and thoroughly vet my PROPOSAL. I'm willing to serve.

Kindly submitted by:

Robert B. Holston Jr. "Rob" 907-617-2262 holston@kpunet.net Health & Longevity Coach
I am a volunteer :) "Willing to serve."

2. Data, Software, and Model Sharing: PREMISE

A 2014 Student Growth Screening Survey indicates about 42% of Ketchikan district students, pre-kindergarten through 9th grade, are either overweight or obese. An additional 6.5% of the student population is ranked as severely obese (MORBID OBESITY.) That totals 48.5%. Barbara McCarthy was the Ketchikan School District's Wellness Coordinator and said "….with parental permission, 88 percent of all students enrolled in the Ketchikan School District participated in the survey."

According to the most recent Behavioral Risk Factor Surveillance System (BRFSS) data, adult obesity rates now exceed 35% in 23 states. West Virginia has the highest adult obesity rate at 41.2% and the District of

Columbia (D.C.) has the lowest at 23.5%. Between 2022 and 2023, three states experienced statistically significant increases in their adult obesity rate: Alaska, Arkansas, and Oregon. No state experienced a decline. (Assumed fact- “Many Overweight students will become obese adults.”)

“Consistently studies show that students who eat well and students who are active will perform better academically, they are more likely to go on to higher education and they have fewer classroom behavior issues” according to McCarthy.

In 2011–2012, approximately 1 in 6 (17.7%) US children aged 6 to 11 were obese. Compared with their non-obese peers, these children are more likely to experience low self-esteem and depression, discrimination by other children, and academic struggles; they are also more likely to become obese adults. The ability to monitor childhood obesity prevalence at the state and local levels is critical for state public health programs to assess trends and the success or failure of interventions.
Obesity can negatively affect a student's academic learning through several pathways, including biological factors like compromised executive function and memory, indirect factors such as increased school absenteeism and bullying, and teacher bias which can lead to lower expectations and grading. This can result in lower grades, poorer test scores in reading and math, and less likelihood of graduating from college.

6.5% of Ketchikan’s young students were measured as having MORBID OBESITY. Morbid obesity at an early age can significantly reduce a child's life expectancy, with some studies suggesting it could be cut by half. This is because severe early-onset obesity is linked to developing serious health conditions like heart disease & type 2 diabetes at an earlier age, increasing the risk of mortality later in life. The impact is profound and highlights the need for early and comprehensive prevention and treatment to improve both the quality and length of life.

A childhood diagnosis of Type 2 diabetes significantly reduces life expectancy due to the long-term impact of high blood sugar, increasing the risk of early and severe complications like heart disease, stroke, and kidney disease. Studies estimate that life expectancy could be reduced by 15 to 19 years or more. Diabesity leads to Chronic Disease and Chronic Disease leads to disabilities, increased insurance claims and premature death.
pastedGraphic.png
Alaska Nutrition, Physical Activity, and Obesity Facts Report- 2025
“For over two decades, the percentage of adult and high school-aged Alaskans with obesity has continued to rise. Physical activity and healthy eating can prevent overweight, obesity, and several chronic diseases, including many cancers, heart disease, and type 2 diabetes. It can also impact mental health. Healthy eating patterns have been associated with a reduced risk of depression. The mental health benefits of physical activity—for adults and children—include better mood, sleep and focus.”
ALASKA Physical Activity, Nutrition & Obesity Facts Report.

PROPOSAL
To combat Obesity amongst Alaskan School students:

At the high school and junior high/middle school levels: Reassign all P.E. & health teachers to become Health & P.E. Concierges (HPEC). Their responsibilities will include assisting students and homeroom teachers to introduce, implement and maintain each student's individual P.E. & Health Journal (PE&HJ), in a manner similar to IEP's for special needs students. The Journal will remain the property of the student. The monthly markers of change will be shared with each student's concierge. The concierge along with the student will decide upon a health, diet and physical education program that meets the student where they are and plans for maintenance or improvement in weight, strength, mobility, flexibility, body-fat-%, heart rate, blood pressure, etc. all of which can be measured with a body analysis scale preferably located in each home room and each Health & P.E. Concierge's office. These concierges will provide students with lists of foods and beverages to avoid and also lists and portions of foods and beverages to consume. This protocol could also be implemented at the elementary school level. Faculty & staff would be encouraged to participate also thus lowering the incidence and complications of chronic disease.

Activity: for 8 minutes each morning and 8 minutes each afternoon all students (& staff?) will participate in a nonstop physical activity; walking, jogging, jumping, 6 count burpee's, jumping jacks, etc. or combinations thereof agreed to by the student with the guidance of their concierge. The idea is to move for 8 minutes and work for improvement in: distance traveled, numbers of repetitions, difficulty of tasks performed, etc.

Compare this to a typical P.E. class on softball where 9 players are on the field and the pitcher and catcher are the only ones even coming close to getting any actual exercise. The OTHER half of the class is in the dugout, SITTING! Consider ever popular dodgeball where the obese slow player is sidelined only to view the athletes continue to get exercise.

Completing the PE&HJ for all four years of H.S. attendance should become the new graduation requirement, replacing the current failed health and P.E. classes requirements most districts now have. An extensive health & lifestyle bibliography could be incorporated into most other subjects of academia: math, English, science, reading, etc.

Please compare this PROPOSAL with ALL other class requirements. There is NO OTHER SUBJECT MATTER taught that has such practical application to each student's life every single day of their lives. It teaches and guides students to make informed choices on what to eat and not eat, what to drink and not drink and whether or not to take the stairs as opposed to the elevator. These are daily choices that are impacting our students multiple times per day and will, daily for the rest of their lives. District staff members should also benefit from the district's focus on health and longevity.

Concierges will enter weights & body fate % data into a spreadsheet to calculate school wide improvements, while protecting each student's identity. Please realize that this plan is subject to changes and improvements but also realize that the existing curriculum is NOT solving the problems of student obesity.

Alaska School Districts are encouraged to abandon their existing Health and P.E. curriculums and replace the present 1/2 semester each of Health & P.E. as graduation requirements with the requirement to

complete 4 years of a student's PE&HJ (P.E. & HEALTH JOURNAL).

I'm suggesting that the Alaska School Boards and the Alaska Department of Education & Early Development take initial steps to address this issue and form task force committees to examine the PREMISE and thoroughly vet my PROPOSAL. I'm willing to serve.

Kindly submitted by:

Robert B. Holston Jr. "Rob" 907-617-2262 holston@kpunet.net Health & Longevity Coach
I am a volunteer :) "Willing to serve."

3. Training and Mentorship: PREMISE

A 2014 Student Growth Screening Survey indicates about 42% of Ketchikan district students, pre-kindergarten through 9th grade, are either overweight or obese. An additional 6.5% of the student population is ranked as severely obese (MORBID OBESITY.) That totals 48.5%. Barbara McCarthy was the Ketchikan School District's Wellness Coordinator and said "….with parental permission, 88 percent of all students enrolled in the Ketchikan School District participated in the survey."

According to the most recent Behavioral Risk Factor Surveillance System (BRFSS) data, adult obesity rates now exceed 35% in 23 states. West Virginia has the highest adult obesity rate at 41.2% and the District of Columbia (D.C.) has the lowest at 23.5%. Between 2022 and 2023, three states experienced statistically significant increases in their adult obesity rate: Alaska, Arkansas, and Oregon. No state experienced a decline. (Assumed fact- "Many Overweight students will become obese adults.")

"Consistently studies show that students who eat well and students who are active will perform better academically, they are more likely to go on to higher education and they have fewer classroom behavior issues" according to McCarthy.

In 2011–2012, approximately 1 in 6 (17.7%) US children aged 6 to 11 were obese. Compared with their non-obese peers, these children are more likely to experience low self-esteem and depression, discrimination by other children, and academic struggles; they are also more likely to become obese adults. The ability to monitor childhood obesity prevalence at the state and local levels is critical for state public health programs to assess trends and the success or failure of interventions.
Obesity can negatively affect a student's academic learning through several pathways, including biological factors like compromised executive function and memory, indirect factors such as increased school absenteeism and bullying, and teacher bias which can lead to lower expectations and grading. This can result in lower grades, poorer test scores in reading and math, and less likelihood of graduating from college.

6.5% of Ketchikan's young students were measured as having MORBID OBESITY. Morbid obesity at an early age can significantly reduce a child's life expectancy, with some studies suggesting it could be cut by half. This is because severe early-onset obesity is linked to developing serious health conditions like heart disease & type 2 diabetes at an earlier age, increasing the risk of mortality later in life. The impact is profound and highlights the need for early and comprehensive prevention and treatment to improve

both the quality and length of life.

A childhood diagnosis of Type 2 diabetes significantly reduces life expectancy due to the long-term impact of high blood sugar, increasing the risk of early and severe complications like heart disease, stroke, and kidney disease. Studies estimate that life expectancy could be reduced by 15 to 19 years or more. Diabesity leads to Chronic Disease and Chronic Disease leads to disabilities, increased insurance claims and premature death.
pastedGraphic.png
Alaska Nutrition, Physical Activity, and Obesity Facts Report- 2025
“For over two decades, the percentage of adult and high school-aged Alaskans with obesity has continued to rise. Physical activity and healthy eating can prevent overweight, obesity, and several chronic diseases, including many cancers, heart disease, and type 2 diabetes. It can also impact mental health. Healthy eating patterns have been associated with a reduced risk of depression. The mental health benefits of physical activity—for adults and children—include better mood, sleep and focus.”
ALASKA Physical Activity, Nutrition & Obesity Facts Report.

PROPOSAL
To combat Obesity amongst Alaskan School students:

At the high school and junior high/middle school levels: Reassign all P.E. & health teachers to become Health & P.E. Concierges (HPEC). Their responsibilities will include assisting students and homeroom teachers to introduce, implement and maintain each student’s individual P.E. & Health Journal (PE&HJ), in a manner similar to IEP’s for special needs students. The Journal will remain the property of the student. The monthly markers of change will be shared with each student’s concierge. The concierge along with the student will decide upon a health, diet and physical education program that meets the student where they are and plans for maintenance or improvement in weight, strength, mobility, flexibility, body-fat-%, heart rate, blood pressure, etc. all of which can be measured with a body analysis scale preferably located in each home room and each Health & P.E. Concierge’s office. These concierges will provide students with lists of foods and beverages to avoid and also lists and portions of foods and beverages to consume. This protocol could also be implemented at the elementary school level. Faculty & staff would be encouraged to participate also thus lowering the incidence and complications of chronic disease.

Activity: for 8 minutes each morning and 8 minutes each afternoon all students (& staff?) will participate in a nonstop physical activity; walking, jogging, jumping, 6 count burpee's, jumping jacks, etc. or combinations thereof agreed to by the student with the guidance of their concierge. The idea is to move for 8 minutes and work for improvement in: distance traveled, numbers of repetitions, difficulty of tasks performed, etc.

Compare this to a typical P.E. class on softball where 9 players are on the field and the pitcher and catcher are the only ones even coming close to getting any actual exercise. The OTHER half of the class is in the dugout, SITTING! Consider ever popular dodgeball where the obese slow player is sidelined only to view the athletes continue to get exercise.

Completing the PE&HJ for all four years of H.S. attendance should become the new graduation requirement, replacing the current failed health and P.E. classes requirements most districts now have. An extensive health & lifestyle bibliography could be incorporated into most other subjects of academia: math, English, science, reading, etc.

Please compare this PROPOSAL with ALL other class requirements. There is NO OTHER SUBJECT MATTER taught that has such practical application to each student's life every single day of their lives. It teaches and guides students to make informed choices on what to eat and not eat, what to drink and not drink and whether or not to take the stairs as opposed to the elevator. These are daily choices that are impacting our students multiple times per day and will, daily for the rest of their lives. District staff members should also benefit from the district's focus on health and longevity.

Concierges will enter weights & body fate % data into a spreadsheet to calculate school wide improvements, while protecting each student's identity. Please realize that this plan is subject to changes and improvements but also realize that the existing curriculum is NOT solving the problems of student obesity.

Alaska School Districts are encouraged to abandon their existing Health and P.E. curriculums and replace the present 1/2 semester each of Health & P.E. as graduation requirements with the requirement to complete 4 years of a student's PE&HJ (P.E. & HEALTH JOURNAL).

I'm suggesting that the Alaska School Boards and the Alaska Department of Education & Early Development take initial steps to address this issue and form task force committees to examine the PREMISE and thoroughly vet my PROPOSAL. I'm willing to serve.

Kindly submitted by:

Robert B. Holston Jr. "Rob" 907-617-2262 holston@kpunet.net Health & Longevity Coach
I am a volunteer :) "Willing to serve."

4. Collaboration: PREMISE

A 2014 Student Growth Screening Survey indicates about 42% of Ketchikan district students, pre-kindergarten through 9th grade, are either overweight or obese. An additional 6.5% of the student population is ranked as severely obese (MORBID OBESITY.) That totals 48.5%. Barbara McCarthy was the Ketchikan School District's Wellness Coordinator and said "….with parental permission, 88 percent of all students enrolled in the Ketchikan School District participated in the survey."

According to the most recent Behavioral Risk Factor Surveillance System (BRFSS) data, adult obesity rates now exceed 35% in 23 states. West Virginia has the highest adult obesity rate at 41.2% and the District of Columbia (D.C.) has the lowest at 23.5%. Between 2022 and 2023, three states experienced statistically significant increases in their adult obesity rate: Alaska, Arkansas, and Oregon. No state experienced a decline. (Assumed fact- "Many Overweight students will become obese adults.")

“Consistently studies show that students who eat well and students who are active will perform better academically, they are more likely to go on to higher education and they have fewer classroom behavior issues” according to McCarthy.

In 2011–2012, approximately 1 in 6 (17.7%) US children aged 6 to 11 were obese. Compared with their non-obese peers, these children are more likely to experience low self-esteem and depression, discrimination by other children, and academic struggles; they are also more likely to become obese adults. The ability to monitor childhood obesity prevalence at the state and local levels is critical for state public health programs to assess trends and the success or failure of interventions.
Obesity can negatively affect a student's academic learning through several pathways, including biological factors like compromised executive function and memory, indirect factors such as increased school absenteeism and bullying, and teacher bias which can lead to lower expectations and grading. This can result in lower grades, poorer test scores in reading and math, and less likelihood of graduating from college.

6.5% of Ketchikan’s young students were measured as having MORBID OBESITY. Morbid obesity at an early age can significantly reduce a child's life expectancy, with some studies suggesting it could be cut by half. This is because severe early-onset obesity is linked to developing serious health conditions like heart disease & type 2 diabetes at an earlier age, increasing the risk of mortality later in life. The impact is profound and highlights the need for early and comprehensive prevention and treatment to improve both the quality and length of life.

A childhood diagnosis of Type 2 diabetes significantly reduces life expectancy due to the long-term impact of high blood sugar, increasing the risk of early and severe complications like heart disease, stroke, and kidney disease. Studies estimate that life expectancy could be reduced by 15 to 19 years or more. Diabesity leads to Chronic Disease and Chronic Disease leads to disabilities, increased insurance claims and premature death.
pastedGraphic.png
Alaska Nutrition, Physical Activity, and Obesity Facts Report- 2025
“For over two decades, the percentage of adult and high school-aged Alaskans with obesity has continued to rise. Physical activity and healthy eating can prevent overweight, obesity, and several chronic diseases, including many cancers, heart disease, and type 2 diabetes. It can also impact mental health. Healthy eating patterns have been associated with a reduced risk of depression. The mental health benefits of physical activity—for adults and children—include better mood, sleep and focus.” ALASKA Physical Activity, Nutrition & Obesity Facts Report.

PROPOSAL
To combat Obesity amongst Alaskan School students:

At the high school and junior high/middle school levels: Reassign all P.E. & health teachers to become Health & P.E. Concierges (HPEC). Their responsibilities will include assisting students and homeroom teachers to introduce, implement and maintain each student’s individual P.E. & Health Journal (PE&HJ), in a manner similar to IEP’s for special needs students. The Journal will remain the property of the

student. The monthly markers of change will be shared with each student's concierge. The concierge along with the student will decide upon a health, diet and physical education program that meets the student where they are and plans for maintenance or improvement in weight, strength, mobility, flexibility, body-fat-%, heart rate, blood pressure, etc. all of which can be measured with a body analysis scale preferably located in each home room and each Health & P.E. Concierge's office. These concierges will provide students with lists of foods and beverages to avoid and also lists and portions of foods and beverages to consume. This protocol could also be implemented at the elementary school level. Faculty & staff would be encouraged to participate also thus lowering the incidence and complications of chronic disease.

Activity: for 8 minutes each morning and 8 minutes each afternoon all students (& staff?) will participate in a nonstop physical activity; walking, jogging, jumping, 6 count burpee's, jumping jacks, etc. or combinations thereof agreed to by the student with the guidance of their concierge. The idea is to move for 8 minutes and work for improvement in: distance traveled, numbers of repetitions, difficulty of tasks performed, etc.

Compare this to a typical P.E. class on softball where 9 players are on the field and the pitcher and catcher are the only ones even coming close to getting any actual exercise. The OTHER half of the class is in the dugout, SITTING! Consider ever popular dodgeball where the obese slow player is sidelined only to view the athletes continue to get exercise.

Completing the PE&HJ for all four years of H.S. attendance should become the new graduation requirement, replacing the current failed health and P.E. classes requirements most districts now have. An extensive health & lifestyle bibliography could be incorporated into most other subjects of academia: math, English, science, reading, etc.

Please compare this PROPOSAL with ALL other class requirements. There is NO OTHER SUBJECT MATTER taught that has such practical application to each student's life every single day of their lives. It teaches and guides students to make informed choices on what to eat and not eat, what to drink and not drink and whether or not to take the stairs as opposed to the elevator. These are daily choices that are impacting our students multiple times per day and will, daily for the rest of their lives. District staff members should also benefit from the district's focus on health and longevity.

Concierges will enter weights & body fate % data into a spreadsheet to calculate school wide improvements, while protecting each student's identity. Please realize that this plan is subject to changes and improvements but also realize that the existing curriculum is NOT solving the problems of student obesity.

Alaska School Districts are encouraged to abandon their existing Health and P.E. curriculums and replace the present 1/2 semester each of Health & P.E. as graduation requirements with the requirement to complete 4 years of a student's PE&HJ (P.E. & HEALTH JOURNAL).

I'm suggesting that the Alaska School Boards and the Alaska Department of Education & Early Development take initial steps to address this issue and form task force committees to examine the

PREMISE and thoroughly vet my PROPOSAL. I'm willing to serve.

Kindly submitted by:

Robert B. Holston Jr. "Rob" 907-617-2262 holston@kpunet.net Health & Longevity Coach
I am a volunteer :) "Willing to serve."

5. Entrepreneurship and Translation: PREMISE

A 2014 Student Growth Screening Survey indicates about 42% of Ketchikan district students, pre-kindergarten through 9th grade, are either overweight or obese. An additional 6.5% of the student population is ranked as severely obese (MORBID OBESITY.) That totals 48.5%. Barbara McCarthy was the Ketchikan School District's Wellness Coordinator and said "….with parental permission, 88 percent of all students enrolled in the Ketchikan School District participated in the survey."

According to the most recent Behavioral Risk Factor Surveillance System (BRFSS) data, adult obesity rates now exceed 35% in 23 states. West Virginia has the highest adult obesity rate at 41.2% and the District of Columbia (D.C.) has the lowest at 23.5%. Between 2022 and 2023, three states experienced statistically significant increases in their adult obesity rate: Alaska, Arkansas, and Oregon. No state experienced a decline. (Assumed fact- "Many Overweight students will become obese adults.")

"Consistently studies show that students who eat well and students who are active will perform better academically, they are more likely to go on to higher education and they have fewer classroom behavior issues" according to McCarthy.

In 2011–2012, approximately 1 in 6 (17.7%) US children aged 6 to 11 were obese. Compared with their non-obese peers, these children are more likely to experience low self-esteem and depression, discrimination by other children, and academic struggles; they are also more likely to become obese adults. The ability to monitor childhood obesity prevalence at the state and local levels is critical for state public health programs to assess trends and the success or failure of interventions.
Obesity can negatively affect a student's academic learning through several pathways, including biological factors like compromised executive function and memory, indirect factors such as increased school absenteeism and bullying, and teacher bias which can lead to lower expectations and grading. This can result in lower grades, poorer test scores in reading and math, and less likelihood of graduating from college.

6.5% of Ketchikan's young students were measured as having MORBID OBESITY. Morbid obesity at an early age can significantly reduce a child's life expectancy, with some studies suggesting it could be cut by half. This is because severe early-onset obesity is linked to developing serious health conditions like heart disease & type 2 diabetes at an earlier age, increasing the risk of mortality later in life. The impact is profound and highlights the need for early and comprehensive prevention and treatment to improve both the quality and length of life.

A childhood diagnosis of Type 2 diabetes significantly reduces life expectancy due to the long-term impact of high blood sugar, increasing the risk of early and severe complications like heart disease,

stroke, and kidney disease. Studies estimate that life expectancy could be reduced by 15 to 19 years or more. Diabesity leads to Chronic Disease and Chronic Disease leads to disabilities, increased insurance claims and premature death.
pastedGraphic.png
Alaska Nutrition, Physical Activity, and Obesity Facts Report- 2025
“For over two decades, the percentage of adult and high school-aged Alaskans with obesity has continued to rise. Physical activity and healthy eating can prevent overweight, obesity, and several chronic diseases, including many cancers, heart disease, and type 2 diabetes. It can also impact mental health. Healthy eating patterns have been associated with a reduced risk of depression. The mental health benefits of physical activity—for adults and children—include better mood, sleep and focus.” ALASKA Physical Activity, Nutrition & Obesity Facts Report.

PROPOSAL
To combat Obesity amongst Alaskan School students:

At the high school and junior high/middle school levels: Reassign all P.E. & health teachers to become Health & P.E. Concierges (HPEC). Their responsibilities will include assisting students and homeroom teachers to introduce, implement and maintain each student’s individual P.E. & Health Journal (PE&HJ), in a manner similar to IEP’s for special needs students. The Journal will remain the property of the student. The monthly markers of change will be shared with each student’s concierge. The concierge along with the student will decide upon a health, diet and physical education program that meets the student where they are and plans for maintenance or improvement in weight, strength, mobility, flexibility, body-fat-%, heart rate, blood pressure, etc. all of which can be measured with a body analysis scale preferably located in each home room and each Health & P.E. Concierge’s office. These concierges will provide students with lists of foods and beverages to avoid and also lists and portions of foods and beverages to consume. This protocol could also be implemented at the elementary school level. Faculty & staff would be encouraged to participate also thus lowering the incidence and complications of chronic disease.

Activity: for 8 minutes each morning and 8 minutes each afternoon all students (& staff?) will participate in a nonstop physical activity; walking, jogging, jumping, 6 count burpee's, jumping jacks, etc. or combinations thereof agreed to by the student with the guidance of their concierge. The idea is to move for 8 minutes and work for improvement in: distance traveled, numbers of repetitions, difficulty of tasks performed, etc.

Compare this to a typical P.E. class on softball where 9 players are on the field and the pitcher and catcher are the only ones even coming close to getting any actual exercise. The OTHER half of the class is in the dugout, SITTING! Consider ever popular dodgeball where the obese slow player is sidelined only to view the athletes continue to get exercise.

Completing the PE&HJ for all four years of H.S. attendance should become the new graduation requirement, replacing the current failed health and P.E. classes requirements most districts now have. An extensive health & lifestyle bibliography could be incorporated into most other subjects of academia:

math, English, science, reading, etc.

Please compare this PROPOSAL with ALL other class requirements. There is NO OTHER SUBJECT MATTER taught that has such practical application to each student's life every single day of their lives. It teaches and guides students to make informed choices on what to eat and not eat, what to drink and not drink and whether or not to take the stairs as opposed to the elevator. These are daily choices that are impacting our students multiple times per day and will, daily for the rest of their lives. District staff members should also benefit from the district's focus on health and longevity.

Concierges will enter weights & body fate % data into a spreadsheet to calculate school wide improvements, while protecting each student's identity. Please realize that this plan is subject to changes and improvements but also realize that the existing curriculum is NOT solving the problems of student obesity.

Alaska School Districts are encouraged to abandon their existing Health and P.E. curriculums and replace the present 1/2 semester each of Health & P.E. as graduation requirements with the requirement to complete 4 years of a student's PE&HJ (P.E. & HEALTH JOURNAL).

I'm suggesting that the Alaska School Boards and the Alaska Department of Education & Early Development take initial steps to address this issue and form task force committees to examine the PREMISE and thoroughly vet my PROPOSAL. I'm willing to serve.

Kindly submitted by:

Robert B. Holston Jr. "Rob" 907-617-2262 holston@kpunet.net Health & Longevity Coach
I am a volunteer :) "Willing to serve."

6. Foundational Scientific Exploration: PREMISE

A 2014 Student Growth Screening Survey indicates about 42% of Ketchikan district students, pre-kindergarten through 9th grade, are either overweight or obese. An additional 6.5% of the student population is ranked as severely obese (MORBID OBESITY.) That totals 48.5%. Barbara McCarthy was the Ketchikan School District's Wellness Coordinator and said "….with parental permission, 88 percent of all students enrolled in the Ketchikan School District participated in the survey."

According to the most recent Behavioral Risk Factor Surveillance System (BRFSS) data, adult obesity rates now exceed 35% in 23 states. West Virginia has the highest adult obesity rate at 41.2% and the District of Columbia (D.C.) has the lowest at 23.5%. Between 2022 and 2023, three states experienced statistically significant increases in their adult obesity rate: Alaska, Arkansas, and Oregon. No state experienced a decline. (Assumed fact- "Many Overweight students will become obese adults.")

"Consistently studies show that students who eat well and students who are active will perform better academically, they are more likely to go on to higher education and they have fewer classroom behavior issues" according to McCarthy.

In 2011–2012, approximately 1 in 6 (17.7%) US children aged 6 to 11 were obese. Compared with their non-obese peers, these children are more likely to experience low self-esteem and depression, discrimination by other children, and academic struggles; they are also more likely to become obese adults. The ability to monitor childhood obesity prevalence at the state and local levels is critical for state public health programs to assess trends and the success or failure of interventions.
Obesity can negatively affect a student's academic learning through several pathways, including biological factors like compromised executive function and memory, indirect factors such as increased school absenteeism and bullying, and teacher bias which can lead to lower expectations and grading. This can result in lower grades, poorer test scores in reading and math, and less likelihood of graduating from college.

6.5% of Ketchikan's young students were measured as having MORBID OBESITY. Morbid obesity at an early age can significantly reduce a child's life expectancy, with some studies suggesting it could be cut by half. This is because severe early-onset obesity is linked to developing serious health conditions like heart disease & type 2 diabetes at an earlier age, increasing the risk of mortality later in life. The impact is profound and highlights the need for early and comprehensive prevention and treatment to improve both the quality and length of life.

A childhood diagnosis of Type 2 diabetes significantly reduces life expectancy due to the long-term impact of high blood sugar, increasing the risk of early and severe complications like heart disease, stroke, and kidney disease. Studies estimate that life expectancy could be reduced by 15 to 19 years or more. Diabesity leads to Chronic Disease and Chronic Disease leads to disabilities, increased insurance claims and premature death.
pastedGraphic.png
Alaska Nutrition, Physical Activity, and Obesity Facts Report- 2025
"For over two decades, the percentage of adult and high school-aged Alaskans with obesity has continued to rise. Physical activity and healthy eating can prevent overweight, obesity, and several chronic diseases, including many cancers, heart disease, and type 2 diabetes. It can also impact mental health. Healthy eating patterns have been associated with a reduced risk of depression. The mental health benefits of physical activity—for adults and children—include better mood, sleep and focus."
ALASKA Physical Activity, Nutrition & Obesity Facts Report.

PROPOSAL
To combat Obesity amongst Alaskan School students:

At the high school and junior high/middle school levels: Reassign all P.E. & health teachers to become Health & P.E. Concierges (HPEC). Their responsibilities will include assisting students and homeroom teachers to introduce, implement and maintain each student's individual P.E. & Health Journal (PE&HJ), in a manner similar to IEP's for special needs students. The Journal will remain the property of the student. The monthly markers of change will be shared with each student's concierge. The concierge along with the student will decide upon a health, diet and physical education program that meets the student where they are and plans for maintenance or improvement in weight, strength, mobility, flexibility, body-fat-%, heart rate, blood pressure, etc. all of which can be measured with a body analysis

scale preferably located in each home room and each Health & P.E. Concierge's office. These concierges will provide students with lists of foods and beverages to avoid and also lists and portions of foods and beverages to consume. This protocol could also be implemented at the elementary school level. Faculty & staff would be encouraged to participate also thus lowering the incidence and complications of chronic disease.

Activity: for 8 minutes each morning and 8 minutes each afternoon all students (& staff?) will participate in a nonstop physical activity; walking, jogging, jumping, 6 count burpee's, jumping jacks, etc. or combinations thereof agreed to by the student with the guidance of their concierge. The idea is to move for 8 minutes and work for improvement in: distance traveled, numbers of repetitions, difficulty of tasks performed, etc.

Compare this to a typical P.E. class on softball where 9 players are on the field and the pitcher and catcher are the only ones even coming close to getting any actual exercise. The OTHER half of the class is in the dugout, SITTING! Consider ever popular dodgeball where the obese slow player is sidelined only to view the athletes continue to get exercise.

Completing the PE&HJ for all four years of H.S. attendance should become the new graduation requirement, replacing the current failed health and P.E. classes requirements most districts now have. An extensive health & lifestyle bibliography could be incorporated into most other subjects of academia: math, English, science, reading, etc.

Please compare this PROPOSAL with ALL other class requirements. There is NO OTHER SUBJECT MATTER taught that has such practical application to each student's life every single day of their lives. It teaches and guides students to make informed choices on what to eat and not eat, what to drink and not drink and whether or not to take the stairs as opposed to the elevator. These are daily choices that are impacting our students multiple times per day and will, daily for the rest of their lives. District staff members should also benefit from the district's focus on health and longevity.

Concierges will enter weights & body fate % data into a spreadsheet to calculate school wide improvements, while protecting each student's identity. Please realize that this plan is subject to changes and improvements but also realize that the existing curriculum is NOT solving the problems of student obesity.

Alaska School Districts are encouraged to abandon their existing Health and P.E. curriculums and replace the present 1/2 semester each of Health & P.E. as graduation requirements with the requirement to complete 4 years of a student's PE&HJ (P.E. & HEALTH JOURNAL).

I'm suggesting that the Alaska School Boards and the Alaska Department of Education & Early Development take initial steps to address this issue and form task force committees to examine the PREMISE and thoroughly vet my PROPOSAL. I'm willing to serve.

Kindly submitted by:

Robert B. Holston Jr. “Rob” 907-617-2262 holston@kpunet.net Health & Longevity Coach
I am a volunteer :) "Willing to serve."

7. Public Impact: PREMISE

A 2014 Student Growth Screening Survey indicates about 42% of Ketchikan district students, pre-kindergarten through 9th grade, are either overweight or obese. An additional 6.5% of the student population is ranked as severely obese (MORBID OBESITY.) That totals 48.5%. Barbara McCarthy was the Ketchikan School District’s Wellness Coordinator and said “….with parental permission, 88 percent of all students enrolled in the Ketchikan School District participated in the survey.”

According to the most recent Behavioral Risk Factor Surveillance System (BRFSS) data, adult obesity rates now exceed 35% in 23 states. West Virginia has the highest adult obesity rate at 41.2% and the District of Columbia (D.C.) has the lowest at 23.5%. Between 2022 and 2023, three states experienced statistically significant increases in their adult obesity rate: Alaska, Arkansas, and Oregon. No state experienced a decline. (Assumed fact- “Many Overweight students will become obese adults.”)

“Consistently studies show that students who eat well and students who are active will perform better academically, they are more likely to go on to higher education and they have fewer classroom behavior issues” according to McCarthy.

In 2011–2012, approximately 1 in 6 (17.7%) US children aged 6 to 11 were obese. Compared with their non-obese peers, these children are more likely to experience low self-esteem and depression, discrimination by other children, and academic struggles; they are also more likely to become obese adults. The ability to monitor childhood obesity prevalence at the state and local levels is critical for state public health programs to assess trends and the success or failure of interventions.
Obesity can negatively affect a student's academic learning through several pathways, including biological factors like compromised executive function and memory, indirect factors such as increased school absenteeism and bullying, and teacher bias which can lead to lower expectations and grading. This can result in lower grades, poorer test scores in reading and math, and less likelihood of graduating from college.

6.5% of Ketchikan’s young students were measured as having MORBID OBESITY. Morbid obesity at an early age can significantly reduce a child's life expectancy, with some studies suggesting it could be cut by half. This is because severe early-onset obesity is linked to developing serious health conditions like heart disease & type 2 diabetes at an earlier age, increasing the risk of mortality later in life. The impact is profound and highlights the need for early and comprehensive prevention and treatment to improve both the quality and length of life.

A childhood diagnosis of Type 2 diabetes significantly reduces life expectancy due to the long-term impact of high blood sugar, increasing the risk of early and severe complications like heart disease, stroke, and kidney disease. Studies estimate that life expectancy could be reduced by 15 to 19 years or more. Diabesity leads to Chronic Disease and Chronic Disease leads to disabilities, increased insurance claims and premature death.
pastedGraphic.png

Alaska Nutrition, Physical Activity, and Obesity Facts Report- 2025
“For over two decades, the percentage of adult and high school-aged Alaskans with obesity has continued to rise. Physical activity and healthy eating can prevent overweight, obesity, and several chronic diseases, including many cancers, heart disease, and type 2 diabetes. It can also impact mental health. Healthy eating patterns have been associated with a reduced risk of depression. The mental health benefits of physical activity—for adults and children—include better mood, sleep and focus.” ALASKA Physical Activity, Nutrition & Obesity Facts Report.

PROPOSAL
To combat Obesity amongst Alaskan School students:

At the high school and junior high/middle school levels: Reassign all P.E. & health teachers to become Health & P.E. Concierges (HPEC). Their responsibilities will include assisting students and homeroom teachers to introduce, implement and maintain each student’s individual P.E. & Health Journal (PE&HJ), in a manner similar to IEP’s for special needs students. The Journal will remain the property of the student. The monthly markers of change will be shared with each student’s concierge. The concierge along with the student will decide upon a health, diet and physical education program that meets the student where they are and plans for maintenance or improvement in weight, strength, mobility, flexibility, body-fat-%, heart rate, blood pressure, etc. all of which can be measured with a body analysis scale preferably located in each home room and each Health & P.E. Concierge’s office. These concierges will provide students with lists of foods and beverages to avoid and also lists and portions of foods and beverages to consume. This protocol could also be implemented at the elementary school level. Faculty & staff would be encouraged to participate also thus lowering the incidence and complications of chronic disease.

Activity: for 8 minutes each morning and 8 minutes each afternoon all students (& staff?) will participate in a nonstop physical activity; walking, jogging, jumping, 6 count burpee's, jumping jacks, etc. or combinations thereof agreed to by the student with the guidance of their concierge. The idea is to move for 8 minutes and work for improvement in: distance traveled, numbers of repetitions, difficulty of tasks performed, etc.

Compare this to a typical P.E. class on softball where 9 players are on the field and the pitcher and catcher are the only ones even coming close to getting any actual exercise. The OTHER half of the class is in the dugout, SITTING! Consider ever popular dodgeball where the obese slow player is sidelined only to view the athletes continue to get exercise.

Completing the PE&HJ for all four years of H.S. attendance should become the new graduation requirement, replacing the current failed health and P.E. classes requirements most districts now have. An extensive health & lifestyle bibliography could be incorporated into most other subjects of academia: math, English, science, reading, etc.

Please compare this PROPOSAL with ALL other class requirements. There is NO OTHER SUBJECT MATTER taught that has such practical application to each student’s life every single day of their lives. It teaches

and guides students to make informed choices on what to eat and not eat, what to drink and not drink and whether or not to take the stairs as opposed to the elevator. These are daily choices that are impacting our students multiple times per day and will, daily for the rest of their lives. District staff members should also benefit from the district's focus on health and longevity.

Concierges will enter weights & body fate % data into a spreadsheet to calculate school wide improvements, while protecting each student's identity. Please realize that this plan is subject to changes and improvements but also realize that the existing curriculum is NOT solving the problems of student obesity.

Alaska School Districts are encouraged to abandon their existing Health and P.E. curriculums and replace the present 1/2 semester each of Health & P.E. as graduation requirements with the requirement to complete 4 years of a student's PE&HJ (P.E. & HEALTH JOURNAL).

I'm suggesting that the Alaska School Boards and the Alaska Department of Education & Early Development take initial steps to address this issue and form task force committees to examine the PREMISE and thoroughly vet my PROPOSAL. I'm willing to serve.

Kindly submitted by:

Robert B. Holston Jr. "Rob" 907-617-2262 holston@kpunet.net Health & Longevity Coach
I am a volunteer :) "Willing to serve."

8. Other Areas: PREMISE

A 2014 Student Growth Screening Survey indicates about 42% of Ketchikan district students, pre-kindergarten through 9th grade, are either overweight or obese. An additional 6.5% of the student population is ranked as severely obese (MORBID OBESITY.) That totals 48.5%. Barbara McCarthy was the Ketchikan School District's Wellness Coordinator and said "….with parental permission, 88 percent of all students enrolled in the Ketchikan School District participated in the survey."

According to the most recent Behavioral Risk Factor Surveillance System (BRFSS) data, adult obesity rates now exceed 35% in 23 states. West Virginia has the highest adult obesity rate at 41.2% and the District of Columbia (D.C.) has the lowest at 23.5%. Between 2022 and 2023, three states experienced statistically significant increases in their adult obesity rate: Alaska, Arkansas, and Oregon. No state experienced a decline. (Assumed fact- "Many Overweight students will become obese adults.")

"Consistently studies show that students who eat well and students who are active will perform better academically, they are more likely to go on to higher education and they have fewer classroom behavior issues" according to McCarthy.

In 2011–2012, approximately 1 in 6 (17.7%) US children aged 6 to 11 were obese. Compared with their non-obese peers, these children are more likely to experience low self-esteem and depression, discrimination by other children, and academic struggles; they are also more likely to become obese adults. The ability to monitor childhood obesity prevalence at the state and local levels is critical for state

public health programs to assess trends and the success or failure of interventions.
Obesity can negatively affect a student's academic learning through several pathways, including biological factors like compromised executive function and memory, indirect factors such as increased school absenteeism and bullying, and teacher bias which can lead to lower expectations and grading. This can result in lower grades, poorer test scores in reading and math, and less likelihood of graduating from college.

6.5% of Ketchikan's young students were measured as having MORBID OBESITY. Morbid obesity at an early age can significantly reduce a child's life expectancy, with some studies suggesting it could be cut by half. This is because severe early-onset obesity is linked to developing serious health conditions like heart disease & type 2 diabetes at an earlier age, increasing the risk of mortality later in life. The impact is profound and highlights the need for early and comprehensive prevention and treatment to improve both the quality and length of life.

A childhood diagnosis of Type 2 diabetes significantly reduces life expectancy due to the long-term impact of high blood sugar, increasing the risk of early and severe complications like heart disease, stroke, and kidney disease. Studies estimate that life expectancy could be reduced by 15 to 19 years or more. Diabesity leads to Chronic Disease and Chronic Disease leads to disabilities, increased insurance claims and premature death.
pastedGraphic.png
Alaska Nutrition, Physical Activity, and Obesity Facts Report- 2025
"For over two decades, the percentage of adult and high school-aged Alaskans with obesity has continued to rise. Physical activity and healthy eating can prevent overweight, obesity, and several chronic diseases, including many cancers, heart disease, and type 2 diabetes. It can also impact mental health. Healthy eating patterns have been associated with a reduced risk of depression. The mental health benefits of physical activity—for adults and children—include better mood, sleep and focus."
ALASKA Physical Activity, Nutrition & Obesity Facts Report.

PROPOSAL
To combat Obesity amongst Alaskan School students:

At the high school and junior high/middle school levels: Reassign all P.E. & health teachers to become Health & P.E. Concierges (HPEC). Their responsibilities will include assisting students and homeroom teachers to introduce, implement and maintain each student's individual P.E. & Health Journal (PE&HJ), in a manner similar to IEP's for special needs students. The Journal will remain the property of the student. The monthly markers of change will be shared with each student's concierge. The concierge along with the student will decide upon a health, diet and physical education program that meets the student where they are and plans for maintenance or improvement in weight, strength, mobility, flexibility, body-fat-%, heart rate, blood pressure, etc. all of which can be measured with a body analysis scale preferably located in each home room and each Health & P.E. Concierge's office. These concierges will provide students with lists of foods and beverages to avoid and also lists and portions of foods and beverages to consume. This protocol could also be implemented at the elementary school level. Faculty & staff would be encouraged to participate also thus lowering the incidence and complications of chronic

disease.

Activity: for 8 minutes each morning and 8 minutes each afternoon all students (& staff?) will participate in a nonstop physical activity; walking, jogging, jumping, 6 count burpee's, jumping jacks, etc. or combinations thereof agreed to by the student with the guidance of their concierge. The idea is to move for 8 minutes and work for improvement in: distance traveled, numbers of repetitions, difficulty of tasks performed, etc.

Compare this to a typical P.E. class on softball where 9 players are on the field and the pitcher and catcher are the only ones even coming close to getting any actual exercise. The OTHER half of the class is in the dugout, SITTING! Consider ever popular dodgeball where the obese slow player is sidelined only to view the athletes continue to get exercise.

Completing the PE&HJ for all four years of H.S. attendance should become the new graduation requirement, replacing the current failed health and P.E. classes requirements most districts now have. An extensive health & lifestyle bibliography could be incorporated into most other subjects of academia: math, English, science, reading, etc.

Please compare this PROPOSAL with ALL other class requirements. There is NO OTHER SUBJECT MATTER taught that has such practical application to each student's life every single day of their lives. It teaches and guides students to make informed choices on what to eat and not eat, what to drink and not drink and whether or not to take the stairs as opposed to the elevator. These are daily choices that are impacting our students multiple times per day and will, daily for the rest of their lives. District staff members should also benefit from the district's focus on health and longevity.

Concierges will enter weights & body fate % data into a spreadsheet to calculate school wide improvements, while protecting each student's identity. Please realize that this plan is subject to changes and improvements but also realize that the existing curriculum is NOT solving the problems of student obesity.

Alaska School Districts are encouraged to abandon their existing Health and P.E. curriculums and replace the present 1/2 semester each of Health & P.E. as graduation requirements with the requirement to complete 4 years of a student's PE&HJ (P.E. & HEALTH JOURNAL).

I'm suggesting that the Alaska School Boards and the Alaska Department of Education & Early Development take initial steps to address this issue and form task force committees to examine the PREMISE and thoroughly vet my PROPOSAL. I'm willing to serve.

Kindly submitted by:

Robert B. Holston Jr. "Rob" 907-617-2262 holston@kpunet.net Health & Longevity Coach
I am a volunteer :) "Willing to serve."

86. Antony Schwartz

I am responding: On behalf of myself

Name: Antony Schwartz

Name of Organization:

Type of Organization: Academic Institution

Role: Institutional Review Oversight Committee Member

Date Submitted: 7/8/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: Please see attachment.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Comment-on-NOT-OD-26-087-RFI_07-2026.pdf

Description: Recognizing Biosafety, Biosecurity, and Safety Oversight as Measurable Indicators of Rigorous and High Impact Science

Recognizing Biosafety, Biosecurity, and Safety Oversight as Measurable Indicators of Rigorous and High Impact Science

Submitted in response to: NIH Request for Information NOT-OD-26-087: Measuring and Rewarding Scientific Impact

Submitted by:
Antony Schwartz, dr.schwartz.safety@gmail.com
Randy Albrecht, randyalbrecht2018@gmail.com
David Gillum, david.gillum@gmail.com
Wyndi Klement, klement.wyndi@gmail.com
Rebecca Moritz, rlmoritz5@gmail.com
Megan Shoff, megan.shoff@tamu.edu

Core Recommendation

NIH should explicitly recognize biosafety, biosecurity, and safety oversight as measurable leading indicators of rigorous, reproducible, transparent, and publicly responsible biomedical research.

These activities are not separate from scientific impact. They are part of the conditions that determine whether research can be interpreted, repeated, translated, and trusted. A study that uses infectious agents, recombinant or synthetic nucleic acids, biological toxins, human materials, animals, arthropods, environmental samples, or emerging biotechnologies is shaped not only by its experimental design, but also by the risk assessment, containment practices, oversight review, personnel competency, material control, and institutional systems under which the work is conducted.

NIH has an opportunity to align scientific incentives with the full research lifecycle by recognizing that gold standard science requires gold standard biosafety.

This recommendation is also consistent with NIH's ongoing modernization of the NIH Guidelines for Research Involving Recombinant or Synthetic Nucleic Acid Molecules, as directed by the NIH Biosafety Modernization Initiative and related policy updates. As NIH modernizes biosafety oversight to reflect current technologies, risk paradigms, and institutional practices, there is a parallel opportunity to ensure that these expectations are consistently applied and visible within how scientific impact is defined and rewarded.

Rationale

Scientific impact is often measured through publications, citations, journal prestige, grant productivity, patents, data sharing, and translation of scientific findings into practical application. These measures capture important dimensions of productivity and influence, but they do not fully account for whether research was conducted under conditions that support safety, security, reproducibility, transparency, and public trust. When the scientific enterprise is supported by public funding, these considerations are equally important, if not more important, to sustaining public support and maintaining international leadership in science.

Biosafety and biosecurity oversight are foundational to responsible biomedical research. Existing federal and institutional frameworks already require risk assessment, appropriate containment, training, incident

response, institutional review, and continuing oversight for many categories of life sciences research. Institutional Biosafety Committees, biosafety professionals, animal care and use programs, occupational health programs, research compliance offices, and other institutional partners collectively support the safe and responsible conduct of science. Many of these oversight processes are also required as a condition of federal funding.

Despite this, these activities are rarely visible in how scientific impact is measured. In many publications, critical safety-related methodological details are absent, inconsistently reported, or difficult to interpret. Examples include the biosafety level used, containment or facility controls, inactivation methods for materials intended for downstream use, PPE and exposure mitigation strategies, biological material traceability, institutional review status, and whether additional biosecurity or dual use review was completed when applicable. Requiring these oversight activities to be documented in research publications would not only show that lead institutions are conducting work responsibly, but also promote similar standards among collaborators and other laboratories that may replicate or build upon such work.

This gap matters because biosafety conditions can influence experimental reproducibility and interpretation. Safety considerations are embedded throughout the research process, from experimental design and risk assessment to the procedures, materials, equipment, and personnel involved in the work. Risk assessment must account not only for individual agents or materials, but also for how they interact with experimental manipulations, routes of exposure, handling practices, containment level, inactivation methods, personnel competency, equipment controls, and facility conditions. These factors may all affect the quality, reliability, and interpretability of research outputs. They also affect public confidence in federally funded research, particularly when work involves pathogens, emerging technologies, dual use potential, biological materials of concern, or research conducted under heightened public scrutiny.

Proposed Measurable Leading Indicators

NIH could incorporate biosafety, biosecurity, and safety oversight into scientific impact frameworks through tiered, risk proportionate indicators. In this context, leading indicators are evidence that appropriate risk assessment, oversight, controls, training, and accountability were in place before and during the conduct of research, rather than measures that only identify failures after they occur. This approach aligns with the direction of NIH's broader biosafety modernization efforts, which emphasize risk-based oversight, clarity of expectations, and applicability to emerging biotechnologies. These indicators should scale with the nature of the research and should rely as much as possible on existing institutional processes rather than creating duplicative administrative systems. Where applicable, reporting expectations should be standardized across research sectors so readers can interpret biosafety and biosecurity requirements consistently across studies while allowing institutional processes to remain risk proportionate and context specific. This is essential to avoid additional burden for institutions that are already stretched by existing oversight requirements.

A. Project and Protocol Level Indicators

NIH could recognize whether a project includes:

1. Documented biological risk assessment appropriate to the agents, materials, procedures, host range, routes of exposure, scale, and research setting.

2. Institutional review and approval where applicable, including IBC review, DURC or PEPP review, select agent review, animal care and use review, occupational health review, or other relevant institutional oversight.

3. Clear linkage between risk assessment and the selected controls, including containment level, engineering controls, PPE, work practices, exposure response, and decontamination or final disposition controls when materials are not intended for further use.

4. Personnel training and competency documentation for individuals performing high-consequence or specialized procedures, such as work in Biosafety Level 3 or Biosafety Level 4 laboratories.

5. Material accountability and traceability, where relevant, including biological material provenance, transfer approvals, material transfer agreements, inventory expectations, and chain of custody for materials of concern.

6. Documentation of inactivation methods when biological materials are intended for downstream analysis, transfer, storage, reuse, or removal from containment.

7. Evidence of ongoing oversight, such as post approval monitoring, biosafety inspections, corrective actions, learning from incidents, and periodic protocol updates.

B. Publication and Dissemination Indicators

A biosafety and biosecurity methods statement could be required for federally funded research outputs when the work involves biological hazards, recombinant or synthetic nucleic acids, biological toxins, animals, human specimens, or other materials subject to institutional biosafety and biosecurity review. As NIH and the broader research ecosystem move toward clearer and more consistent expectations for transparency in biosafety-relevant methodologies, standardized reporting of applicable requirements, institutional review, and oversight verification would help produce concise, interpretable statements that support rigor, reproducibility, and public trust.

This statement could be concise and risk proportionate. It might include:

1. Applicable institutional approvals, such as IBC, IACUC, DURC or PEPP, select agent, or other relevant reviews.

2. Biosafety level or containment approach used for the work, as documented through applicable institutional review or risk assessment processes.

3. Key facility, equipment, or procedural controls necessary to interpret or reproduce the work.

4. Inactivation methods for materials intended for downstream analysis, transfer, storage, reuse, or removal from containment, when those methods affect interpretation, reproducibility, or downstream application.

5. Decontamination, disinfection, sterilization, destruction, or waste treatment approaches when relevant to responsible final disposition of biological materials, without requiring disclosure of sensitive operational details.

6. Personnel training or competency expectations for specialized procedures.

7. Material transfer, provenance, or traceability information when relevant.
8. A statement that sensitive security details have been withheld when disclosure could create misuse risk.

This would be analogous in spirit to existing publication reporting expectations for animal research, clinical trials, data sharing, conflict of interest, funding acknowledgment, and author contribution statements. The purpose would not be to disclose sensitive operational details. The purpose would be to make safety relevant methodological information and oversight controls visible enough to support rigor, reproducibility, and public trust.

C. Investigator and Team Science Indicators

NIH could recognize biosafety and biosecurity contributions as part of team-based science. In many institutions, the quality of research depends on contributions from biosafety professionals, IBC administrators, veterinarians, occupational health experts, facility engineers, biocontainment staff, research security professionals, data and cyberbiosecurity experts, and compliance personnel.

Possible indicators include:

1. Evidence that investigators engaged appropriate institutional expertise early in study design.
2. Documentation that biosafety or biosecurity considerations informed experimental planning.
3. Inclusion of biosafety or biorisk management expertise in complex or higher risk projects.
4. Recognition of safety professionals and oversight contributors in acknowledgments, methods sections, grant progress reports, or contributor statements when their input significantly shaped the research.

D. Institutional Indicators

NIH could also recognize institutional commitment to biosafety and biosecurity as part of gold standard science. Potential indicators include:

1. Centralized systems for protocol review, biological material oversight, training, occupational health involvement, and incident tracking.
2. Adequate staffing and resourcing for IBC administration, biosafety programs, biosecurity review, and compliance monitoring.
3. Routine use of data and metrics to evaluate program effectiveness.
4. Processes for learning from incidents, near misses, audit findings, and corrective actions.
5. Institutional mechanisms for escalating research with higher biosafety, biosecurity, or dual use concern.
6. Systems for distinguishing and validating materials that are inactivated for downstream use from materials that are decontaminated, destroyed, or otherwise treated for final disposition.
7. Evidence that safety oversight supports research rather than merely delaying or policing it.

To address concerns that institutions with fewer biosafety resources could be disproportionately affected by the inclusion of these metrics, NIH could consider a two-tier approach. The first tier could include minimum biosafety and biosecurity reporting expectations for all institutions receiving federal funding. The second tier could be a voluntary recognition model for institutions with more mature and well-resourced systems. This approach could function as an NIH aligned gold standard biosafety recognition framework rather than a rigid accreditation requirement. Such a model could incentivize institutions to invest in safety infrastructure, improve internal coordination, strengthen public confidence in federally funded research, and continuously improve existing safety systems.

Potential Incentives

NIH could reward biosafety and biosecurity excellence without turning safety into a punitive or overly bureaucratic scoring exercise.

Potential incentives include:

1. Incorporating biosafety and biosecurity reporting into existing rigor and reproducibility expectations.
2. Adding biosafety and biosecurity methods statements to publication and public access expectations for relevant research.
3. Encouraging progress reports to describe major biosafety or biosecurity milestones for higher risk work.
4. Supporting administrative supplements or infrastructure funding for institutions seeking to strengthen biorisk management systems.
5. Recognizing investigators and institutions that model exemplary integration of safety, security, transparency, and scientific rigor.
6. Developing common reporting elements for biosafety and biosecurity, similar to other structured reporting frameworks.
7. Supporting training modules that help investigators understand how safety and security decisions affect research quality and public trust.
8. Encouraging journals and preprint servers to adopt risk proportionate biosafety reporting expectations for federally funded research.

Benefits

Recognizing biosafety, biosecurity, and safety oversight as measurable indicators of scientific impact would provide several benefits.

1. Stronger Reproducibility

Key biosafety variables, such as review of experimental conditions, risk assessment of individual components and system level issues, containment level, handling practices, inactivation methods for

materials intended for downstream use, PPE, equipment controls, and facility conditions, can affect whether a study can be accurately interpreted or repeated.

2. Greater Transparency

A concise biosafety and biosecurity methods statement would help readers understand the conditions under which research was conducted without requiring disclosure of sensitive operational details.

3. Improved Public Trust

The public is more likely to trust federally funded research when institutions can demonstrate that work was conducted under appropriate oversight, accountability, and risk management systems.

4. Better Institutional Coordination

If NIH recognizes biosafety and biosecurity as part of scientific impact, institutions will have a stronger incentive to centralize systems, appropriately size and fund oversight programs, create efficiencies, reduce facility and administrative costs, and reduce silos among research administration, safety, security, occupational health, animal care, and compliance functions.

5. Stronger Safety Culture

Rewarding safety as part of excellent science reinforces the idea that investigators, institutions, and funders share responsibility for responsible research. Beyond institutional benefits, this approach would reinforce a broader culture in which the scientific enterprise recognizes its responsibility to advance science in a manner that is safe, secure, ethical, and publicly accountable.

6. Better Support for Emerging Technologies

Fields such as synthetic biology, gene editing, pathogen research, AI enabled biology, organoids, advanced animal models, and environmental sampling increasingly require integrated scientific, safety, and security review. NIH can help normalize this integration as part of modern biomedical research.

Implementation Considerations

As NIH implements updates to the NIH Guidelines and related biosafety oversight frameworks, incorporating these concepts into scientific impact measures could be phased alongside those policy changes. This would allow institutions to align reporting expectations with evolving federal guidance without creating parallel or conflicting systems. Additionally, these requirements could be incorporated into new and renewing award terms, grant policy statements, and notices of funding opportunities where appropriate. NIH should implement these updates in a way that is practical, scalable, and risk proportionate.

1. Use existing institutional systems where possible. IBC records, training systems, DURC or PEPP review records, occupational health requirements, and inspection findings already exist in many institutions.
2. Avoid creating a one-size-fits-all requirement. Expectations should scale based on the risk profile of the research. A two-tier system, as described above, may help balance broad applicability with institutional flexibility.

3. Protect sensitive security information. Public reporting should not disclose facility vulnerabilities, detailed security controls, exact inventories of materials of concern, or information that could enable misuse.
4. Pilot before broad implementation. NIH could begin with higher relevance categories, such as research involving high-consequence infectious agents, recombinant or synthetic nucleic acids, biological toxins, select agents, PEPP relevant research, or high containment facilities.
5. Engage journals and preprint servers. NIH could work with publishers to develop consistent reporting language that supports transparency without compromising safety or security.
6. Provide templates, checklists, and examples. Investigators and institutions would benefit from short model statements, decision trees, and common data elements.
7. Recognize resource differences across institutions. Smaller institutions, teaching-focused institutions, minority serving institutions, and institutions with limited biosafety infrastructure may need additional guidance and support.

Potential Unintended Consequences and Mitigation

1. Administrative Burden

Risk: New requirements could become another paperwork exercise.

Mitigation: Use existing institutional documentation, create concise reporting templates, and limit detailed reporting to research where biosafety or biosecurity information is relevant to interpretation, reproducibility, or public trust.

2. Inconsistent Institutional Capacity

Risk: Institutions vary widely in biosafety staffing, systems, and documentation maturity.

Mitigation: NIH could develop guidance, common elements, training resources for biosafety personnel, and expect implementation in phases.

3. Disclosure of Sensitive Information

Risk: Publication or reporting expectations could inadvertently reveal private or security-sensitive information that should remain protected.

Mitigation: NIH should clearly distinguish between transparency that supports scientific interpretation and operational details that should not be publicly disclosed.

4. Perverse Incentives

Risk: Investigators or institutions may treat safety reporting as reputation management rather than genuine risk reduction and may suppress reporting of genuine safety concerns.

Mitigation: Indicators should focus on evidence of meaningful risk assessment, oversight, corrective action, and continuous improvement rather than superficial statements of compliance. NIH could also consider direct confidential reporting pathways for safety or oversight concerns, with appropriate protections against retaliation.

5. Overly Rigid Review Criteria

Risk: Safety indicators could be applied mechanically across diverse research fields.

Mitigation: NIH should use a tiered approach that allows flexibility by discipline, material, procedure, scale, and risk level. Not all indicators will be relevant to all fields of biomedical research. NIH should involve biosafety professionals who manage these differences in day-to-day operations to help identify relevant, practical, and meaningful indicators.

Conclusion

Biosafety, biosecurity, and safety oversight should be recognized as measurable components of rigorous and impactful biomedical research. They are not ancillary administrative functions. They shape how research is designed, conducted, interpreted, reproduced, translated, and trusted.

As NIH modernizes its biosafety policy framework to reflect the realities of contemporary life sciences research, incorporating biosafety and biosecurity into measures of scientific impact represents a natural and necessary evolution of that effort. As NIH modernizes its biosafety policy framework to reflect contemporary life sciences research, it has a parallel opportunity to align scientific impact measures with the responsibilities that come with publicly funded science. Gold standard science requires more than discovery. It requires transparent methods, reproducible practices, institutional accountability, responsible innovation, and public trust.

By incorporating biosafety, biosecurity, and safety oversight into its scientific impact framework, NIH can strengthen research quality, promote institutional culture, protect public investments, support emerging technologies, and reinforce a movement in which safe science is understood as excellent science.

87. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Other

Role - Other: Program Director

Date Submitted: 7/10/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: To train a biomedical workforce, there needs to be training in all career types (academia, clinical, industry government, consulting, teaching). There has been a shift in recent years to train more predoctoral fellows on T32s instead of postdoctoral fellows /clinical scientists. This leaves a huge gap in protected time for clinician-scientists, who now more than ever have less protected time for research from clinical duties. More T32s should be provided for postdoctoral level clinical and translational scientists. Benchmarks for success for each program should be based on the specifics of the post-program career types the program produces, as it would vary greatly between clinical coordinators, clinician-scientists, translational researchers, and basic scientists, for examples. At the T/F level of training, better benchmarks are career pathway-oriented, such as positions. At the K level of training, time to promotion in academia/industry and more traditional benchmarks such as publications and subsequent grants may be more appropriate.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Public impact is hard to assess short-term. Measures created should have long-term impact in mind.

8. Other Areas: N/A

88. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/10/2026

1. Rigor and Reproducibility: NIH would need to completely overhaul grant review to do this. Because right now where you publish and how often you publish run counter to the goal of rigor and reproducibility.

2. Data, Software, and Model Sharing: More databases like Pubmed and GEO are needed.

3. Training and Mentorship: Mentorship should apply to ALL levels of trainees. Smaller schools are going to graduate less PhD students but if their undergrads go on to amazing scientific careers that means they are good mentors. Indicators are outcomes of all trainees, types of papers (one or two student versus 50 people).

4. Collaboration: N/A

5. Entrepreneurship and Translation: NIH money that funds trainees should not also fund patent work. It's a huge conflict of interest and detrimental to students.

6. Foundational Scientific Exploration: Stop funding academics that have had their careers. Most faculty hit their prime well before 70. Yes so much money is going to PIs who refuse to retire and have no moved the needle.

7. Public Impact: N/A

8. Other Areas: There is no equity, review and budget decisions are a sham. And most importantly, the modular budget is woefully inadequate.

89. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Member of the Public

Date Submitted: 7/13/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: I am strongly opposed to Section 200.205 - political appointees review of grants. While not employed as a scientist, I understand the scientific method and earned a Ph.D. in 2011. I value my government's investment in and recognition of expert opinion. I have absolutely no desire to see political appointees reviewing grants. This would politicize the scientific process and is absolutely unacceptable. It would result in a degradation of the public trust in NIH research and decision making. We, the public, need to feel confident that our tax dollars are spent / invested wisely and that information provided is based on objective and rigorous methodology.

I am also opposed to 200.454 which makes journal subscriptions unallowable. I value the process of peer review and publication and I understand how these facilitate the dissemination of reliable information within and across fields of study. I want researchers to have access to current, reliable information so that they will stay current in their field of study and reliable, important information will be disseminated within and across fields.

7. Public Impact: N/A

8. Other Areas: I am also opposed to section 200.340 termination provisions. Again, this would inappropriately politicize the scientific process and result in wasted investment (my tax dollars) of projects already underway. You should make every effort to keep politics OUT of scientific decision making and instead make every effort to ensure that an objective, replicable scientific process drives decision making in terms of expenditures / investment of our public resources / tax dollars.

Description: I oppose the politicization of federal financial assistance rules and decision making.

90. Manoja kumar mallick

I am responding: On behalf of myself

Name: Manoja kumar mallick

Name of Organization: Mariyan foundation

Type of Organization: Professional Organization/Association

Role: Member of the Public

Date Submitted: 7/14/2026

1. Rigor and Reproducibility: I am trying to mr bill gates sir before 11 year help fo poor hungry people but no respond please help me

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

91. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/15/2026

1. Rigor and Reproducibility: Strongly agree.

2. Data, Software, and Model Sharing: Strongly agree.

Naive suggestion: For every high-quality software published by the PI, give them 0.5% percentile off towards grant review?

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

92. Al George

I am responding: On behalf of myself

Name: Al George

Name of Organization: Northwestern University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/15/2026

1. Rigor and Reproducibility: This is an important topic. It is vital to train reviewers of NIH grants to appreciate what this means. Sadly, we don't train reviewers in core aspects of rigor and reproducibility or on best practices for evaluating science. This seems like a major obstacle to implementing this policy. A second major obstacle relates to the recent proposal by OMB to allow political appointees without scientific training to determine which grants are funded or terminated. If the OMB rule goes forward, then peer-review is weakened and efforts to promote rigor and reproducibility at the grant review stage will be neutered.

2. Data, Software, and Model Sharing: no comment

3. Training and Mentorship: no comment

4. Collaboration: Collaboration promotes cross-disciplinary science and is vital to our research enterprise. Every effort should be made to promote and not discourage collaboration. Two recent policy changes or proposed changes will discourage collaboration: 1) limits to the number of grant submissions per PI per year is already discouraging MPI grants (some PIs hit the limit because of MPI grants and are discouraged from collaborating on new projects). 2) the proposed cap on simultaneously held grants will discourage MPI grant applications UNLESS an MPI grant is not counted as ONE grant for each PI (a fractional counting makes the most sense). These unintended consequences of these two policies are not being offset by other initiatives to promote collaboration.

5. Entrepreneurship and Translation: no comment

6. Foundational Scientific Exploration: Promoting risking research was the original intent of the R21 program but this nearly abandoned at this point. There was never enough money or time in an R21 to accomplish a paradiagm-shifting discovery. Breakthrough research is expensive research and the current NIH budget fosters conservative non-risky (safe) science. If we want more breakthroughs, we need to pay for the risky science that lead to them. Limiting grant applications and capping numbers of grants, cutting budgets will have the opposite effect.

7. Public Impact: There should be a stronger effort to inform the public about the benefits of biomedical research including making connections between seemingly obscure basic science findings to medical breakthroughs, and highlighting the long time horizon for translating basic research into advances in health and wellness. The NIH should be the leader in doing this, but it has not been the mission. While

this may sound like public relations, it is very important that NIH educates the public at large about why we do research and its long-term (not short term) benefits. We should also make sure the public knows that China is soon to become the leader in science and taxpayers in this country will have to accept that new drugs, new treatments, and other advances in healthcare will not be coming from the U.S.

8. Other Areas: no comment

93. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/15/2026

1. Rigor and Reproducibility: To strengthen rigor and reproducibility, NIH should incentivize independent replication, validation studies, transparent reporting, and the use of standardized protocols. Investigators who develop robust and reproducible research platforms, such as organoid models validated across multiple laboratories or cell lines, should receive recognition comparable to traditional high-impact publications.

2. Data, Software, and Model Sharing: Data, software, model systems, and standardized protocols should be recognized as valuable scientific products. High-quality datasets, computational tools, and experimental platforms that are widely adopted by the research community often have a lasting impact beyond individual publications and should be considered in research evaluation.

3. Training and Mentorship: NIH should also develop better measures of mentorship based on trainee outcomes, independence, and career development rather than simply the number of trainees. Similarly, evaluation of collaboration should move beyond first- or senior-author publications to recognize meaningful contributions to team science, including technology development, methodology, data generation, project leadership, and shared research resources.

4. Collaboration: Researchers should receive recognition for:
conceptualization;
methodology development;
data generation;
analysis;
resource creation;
clinical translation;
project management.

5. Entrepreneurship and Translation: NIH should recognize:
invention disclosures;
patents;
licensing;
startup creation;
clinical trials initiated;
regulatory milestones;

manufacturing platforms;
technology adoption.

6. Foundational Scientific Exploration: The current system sometimes discourages ambitious science because failure negatively affects funding.

NIH should create mechanisms where investigators are rewarded for:

developing new technologies;
exploring unconventional hypotheses;
creating enabling platforms.

Possible indicators:

creation of new research fields;
adoption of new methodologies;
follow-on discoveries by others.

7. Public Impact: Health impact
contribution to clinical guidelines;
improved diagnosis;
therapeutic development;
disease prevention.
Societal impact
patient engagement;
public accessibility;
education;
policy influence.
Economic impact
biotechnology development;
workforce development;
manufacturing capacity.

8. Other Areas: A major unintended consequence is metric overload. If NIH creates 50 new metrics, researchers will spend more time documenting science than doing science. NIH should prioritize a small number of meaningful indicators. Most importantly, these new indicators should be integrated into grant review, funding decisions, and investigator evaluation. Without corresponding changes to reviewer guidance and study section expertise, particularly for New Approach Methodologies (NAMs) and other emerging technologies, these efforts may not meaningfully change the current research culture or fully support innovative, high-impact biomedical science.

Description: I support NIH's effort to broaden the definition of scientific success beyond publication counts and citation metrics. Modern biomedical research increasingly relies on team science, New Approach Methodologies (NAMs), shared resources, and translational platforms that are not adequately recognized under current evaluation systems. However, these new indicators will have meaningful impact only if they are incorporated into grant review, funding decisions, and investigator evaluation.

Otherwise, they risk becoming additional reporting requirements without changing scientific culture. In particular, if NIH seeks to promote NAMs and other innovative technologies, reviewer training and study section composition should include appropriate expertise to ensure these approaches are evaluated fairly rather than judged primarily against traditional methods.

94. Hafizullah Safi

I am responding: On behalf of myself

Name: Hafizullah Safi

Name of Organization: Safi Natural Care Clinic

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 7/16/2026

1. Rigor and Reproducibility: Scientific rigor and reproducibility form the foundation of credible biomedical research. Regardless of scientific discipline or methodology, research findings should be sufficiently transparent and well documented to enable independent researchers to reproduce analyses, verify results, and build upon previous discoveries. Reproducibility not only strengthens confidence in scientific findings but also improves the efficiency of research investments by reducing unnecessary duplication, identifying methodological limitations, and accelerating cumulative scientific progress.
Current academic reward systems often prioritize the novelty of scientific findings over the quality and reproducibility of research methods. Consequently, researchers may receive greater recognition for publishing new discoveries than for validating existing findings, conducting replication studies, improving analytical methods, or producing high-quality datasets that support future research. Although replication studies and negative findings are essential components of the scientific process, they are frequently undervalued in traditional research assessment systems.
NIH has taken important steps through its Replication and Reproducibility Initiative to strengthen scientific integrity. Continued progress, however, will require evaluation systems that actively recognize researchers who consistently demonstrate rigorous scientific practices throughout the research lifecycle. Examples of activities that should receive greater recognition include:
• Publishing detailed study protocols and statistical analysis plans before data analysis.
• Making analytical code, algorithms, and computational workflows openly available whenever appropriate.
• Sharing de-identified research datasets and associated metadata in accordance with ethical and legal requirements.
• Conducting independent validation and replication studies.
• Reporting negative or inconclusive findings that contribute to scientific understanding and reduce publication bias.
• Using standardized reporting guidelines and internationally recognized research standards.
• Implementing transparent quality assurance and quality control procedures throughout data collection and analysis.
• Participating in collaborative efforts to reproduce important scientific findings across different populations and settings.
NIH may also consider incorporating measurable indicators of research rigor into funding and career evaluation processes. Examples include the availability of reproducible analytical workflows, documented data quality procedures, public access to research protocols, successful independent

validation of findings, adherence to FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, and demonstrated compliance with established reporting standards.
Importantly, promoting reproducibility should not be viewed as an administrative requirement but rather as an investment in scientific quality. Rewarding researchers who prioritize methodological excellence, transparency, and validation will strengthen public trust, improve the reliability of biomedical evidence, and increase the long-term value of publicly funded research. A research culture that values reproducibility equally with innovation will produce discoveries that are more robust, more transferable across populations, and more likely to generate lasting scientific and public health impact.

2. Data, Software, and Model Sharing: The value of biomedical research extends well beyond scientific publications. Increasingly, the greatest impact of research is realized through the sharing of high-quality datasets, analytical methods, software, computational models, and research tools that enable other investigators to validate findings, generate new hypotheses, and accelerate scientific discovery. Open and responsible sharing of scientific resources enhances transparency, improves reproducibility, reduces duplication of effort, and maximizes the return on public investment in research.
The COVID-19 pandemic clearly demonstrated the importance of rapid and transparent data sharing. International collaboration through surveillance networks, genomic databases, and open scientific platforms enabled researchers worldwide to monitor pathogen evolution, assess disease transmission, develop diagnostics, evaluate interventions, and accelerate vaccine development. These achievements highlighted that scientific progress depends not only on producing data, but also on making data accessible, interoperable, and reusable by the broader research community.
Despite these benefits, many researchers receive limited academic recognition for developing, curating, maintaining, and sharing valuable scientific datasets or software. Considerable effort is often required to establish surveillance systems, harmonize data standards, ensure data quality, document metadata, develop analytical tools, and maintain long-term databases. These contributions frequently require years of sustained investment but are rarely reflected in conventional measures of scientific productivity.
NIH should recognize data stewardship and scientific resource development as core research contributions. Investigators who make high-quality datasets, software packages, computational pipelines, and predictive models openly available should receive formal recognition comparable to traditional scholarly outputs. Such recognition would encourage broader participation in open science while improving the reproducibility and efficiency of biomedical research.
Future evaluation frameworks could include measurable indicators such as:
• Timely sharing of de-identified research data in accordance with ethical, legal, and regulatory requirements.
• Compliance with FAIR (Findable, Accessible, Interoperable, and Reusable) data principles.
• Development and maintenance of high-quality scientific databases, registries, surveillance systems, and biorepositories.
• Public availability of analytical software, statistical code, machine learning algorithms, and computational workflows.
• Documentation quality, metadata completeness, and standardized data dictionaries.
• Frequency of dataset reuse by independent investigators.
• Integration of shared datasets into national or international scientific platforms.
• Long-term sustainability and accessibility of shared scientific resources.
NIH should also encourage the citation of datasets, software, and digital research products as independent scholarly outputs. Standardized mechanisms for assigning persistent identifiers, such as

Digital Object Identifiers (DOIs), can improve attribution and enable these contributions to be tracked and evaluated alongside traditional publications.

Importantly, data sharing should be implemented responsibly. Policies must continue to safeguard participant privacy, protect confidential information, respect national regulations, and recognize the rights of data-generating institutions and investigators. In particular, researchers working in low- and middle-income countries should receive appropriate authorship, acknowledgment, and opportunities for scientific leadership when their data contribute to international collaborative research.

By recognizing data, software, and model sharing as essential scientific outputs, NIH can foster a research culture that values openness, collaboration, transparency, and long-term scientific utility. Such an approach will accelerate biomedical discovery while strengthening public trust and maximizing the societal benefits of federally funded research.

3. Training and Mentorship: The long-term success of biomedical research depends not only on scientific discoveries but also on the development of a skilled, diverse, and sustainable research workforce. Training and mentorship are fundamental investments that strengthen scientific capacity, promote innovation, improve research quality, and ensure the continuity of scientific excellence across generations. However, these contributions are often underrepresented in current systems used to evaluate scientific impact.

Effective mentorship extends beyond supervising research projects. High-quality mentors cultivate scientific integrity, critical thinking, leadership, collaboration, communication skills, and responsible research practices. They provide guidance that enables trainees to become independent investigators while fostering inclusive and supportive research environments. Similarly, structured training programs that enhance technical competencies, research methodology, data analysis, laboratory quality management, implementation science, and scientific writing contribute substantially to strengthening the biomedical research enterprise.

Current evaluation systems frequently recognize publications and grant income while giving comparatively little credit to sustained mentorship and workforce development. As a result, researchers who dedicate significant time to training early-career scientists, building institutional capacity, or strengthening research infrastructure may receive limited recognition despite their substantial long-term impact.

NIH should incorporate measurable indicators that recognize excellence in training and mentorship. Potential indicators include:

- Successful career progression of trainees into independent research or leadership positions.
- Contributions to multidisciplinary training programs and workforce development initiatives.
- Development of educational resources, curricula, and professional training materials.
- Mentorship of early-career investigators, particularly those from underrepresented or resource-limited settings.
- Establishment of sustainable institutional research capacity.
- Evidence of long-term mentoring relationships and continued professional development of trainees.
- Contributions to strengthening laboratory systems, surveillance capacity, research ethics, and implementation science.
- Recognition received from trainees, institutions, or professional organizations for mentoring excellence.

Importantly, evaluation should emphasize the quality and long-term outcomes of mentorship rather than simply the number of trainees supervised. A single mentor who successfully develops future scientific

leaders may have a greater and more lasting impact than one who supervises many trainees without meaningful professional development.
Investment in mentorship also promotes scientific integrity and reproducibility. Experienced mentors play a critical role in teaching rigorous study design, ethical conduct of research, transparent reporting, responsible data management, and collaborative scientific practices. Strengthening these competencies early in researchers' careers contributes to a more reliable, trustworthy, and resilient biomedical research ecosystem.
Recognizing mentorship and workforce development as essential scientific contributions will encourage researchers to invest in future generations of scientists while strengthening global research capacity and ensuring the long-term sustainability of biomedical innovation.

4. Collaboration: Biomedical research has become increasingly collaborative, interdisciplinary, and international. Addressing today's most pressing health challenges including emerging infectious diseases, antimicrobial resistance, pandemic preparedness, precision medicine, and chronic diseases requires expertise that extends far beyond any single investigator, institution, or scientific discipline. Successful research increasingly depends on effective collaboration among clinicians, epidemiologists, laboratory scientists, data scientists, statisticians, social scientists, engineers, public health practitioners, policymakers, and community stakeholders.
Despite this transformation, research evaluation systems continue to emphasize individual accomplishments, often relying heavily on first or senior authorship, principal investigator status, and individual grant awards. Such approaches may overlook the substantial scientific contributions made by investigators whose expertise is essential to collaborative research but whose roles are less visible in traditional academic metrics. Data managers, surveillance specialists, laboratory scientists, statisticians, software developers, implementation researchers, and project coordinators frequently make indispensable contributions that are difficult to quantify under current evaluation systems.
NIH should further strengthen mechanisms for recognizing meaningful contributions to team science. Evaluation should move beyond publication authorship alone and consider the quality, significance, and impact of each individual's contribution throughout the research process. Modern research benefits when multidisciplinary teams work together to design studies, generate high-quality data, develop analytical methods, interpret findings, and translate evidence into practice.
Potential indicators for recognizing collaborative research may include:
• Participation in multidisciplinary and multi-institutional research projects.
• Contributions to international scientific collaborations and research networks.
• Leadership or coordination of collaborative research initiatives.
• Development of shared research infrastructure, surveillance systems, or laboratory networks.
• Contributions to multicenter studies, implementation research, and consortium-based projects.
• Evidence of equitable partnerships that promote shared leadership and capacity strengthening.
• Collaborative development of standardized protocols, data standards, and quality assurance systems.
• Contributions to cross-disciplinary innovation that integrates complementary areas of expertise.
Evaluation systems should also recognize that impactful collaboration extends beyond academic institutions. Partnerships involving government agencies, public health organizations, healthcare providers, industry, community organizations, patient groups, and international partners often play a critical role in translating scientific discoveries into meaningful health outcomes. These collaborations strengthen the relevance, applicability, and sustainability of biomedical research while increasing its societal impact.

Importantly, collaborative research should be evaluated based on both scientific productivity and the effectiveness of the partnership itself. Indicators such as sustained collaboration, equitable authorship practices, shared decision-making, successful knowledge exchange, and measurable improvements in research capacity may provide valuable insight into the long-term success of collaborative initiatives. By rewarding meaningful collaboration, NIH can encourage researchers to build diverse scientific teams, share expertise across disciplines, reduce unnecessary duplication of effort, and accelerate the translation of scientific discoveries into improved health outcomes.

5. Entrepreneurship and Translation: Scientific discovery achieves its greatest value when research findings are translated into interventions, technologies, policies, diagnostics, therapeutics, and public health practices that improve human health. While fundamental discovery remains essential, evaluation systems should also recognize the processes that move scientific knowledge beyond publication and into real-world application.

Translation is rarely a single event. It involves a continuum of activities including technology development, validation studies, implementation research, regulatory evaluation, commercialization, policy development, clinical adoption, and health system integration. Each stage requires substantial scientific expertise and often involves multidisciplinary collaboration across academia, healthcare, government, and industry. However, many of these activities receive relatively limited recognition within traditional academic evaluation frameworks.

NIH should broaden the definition of translational impact to encompass the full spectrum of activities that convert scientific knowledge into measurable public benefit. Successful translation should not be assessed solely by patents or commercial products but also by the adoption of evidence-based practices, improvements in healthcare delivery, strengthening of public health systems, and incorporation of research findings into national or international guidelines.

Potential indicators of translational impact include:

• Adoption of research findings into clinical practice guidelines or public health policies.
• Development and implementation of new diagnostic methods, surveillance tools, or healthcare technologies.
• Creation of software platforms, digital health applications, decision-support systems, or analytical tools used in healthcare settings.
• Commercialization of biomedical innovations that improve patient care.
• Establishment of public-private partnerships supporting technology development.
• Successful implementation of research findings within healthcare systems or public health programs.
• Demonstrated improvements in healthcare quality, efficiency, or accessibility resulting from research.
• Contribution of scientific evidence to regulatory decisions or national health strategies.

Entrepreneurship should likewise be recognized as an important component of biomedical innovation. Researchers who successfully transform scientific discoveries into scalable solutions frequently contribute to economic development, job creation, technology transfer, and improved healthcare delivery. These activities complement rather than compete with traditional academic research and should be evaluated within the broader context of NIH's mission to improve health.

Importantly, translational success should also include innovations that primarily benefit public health rather than commercial markets. Research that strengthens disease surveillance, enhances laboratory systems, improves outbreak detection, supports emergency preparedness, or informs health policy may generate substantial societal value even when no commercial product results.

Recognizing diverse pathways of translation will encourage researchers to pursue innovations that

address unmet health needs while ensuring that publicly funded biomedical research delivers meaningful and measurable benefits to patients, healthcare systems, and society.

6. Foundational Scientific Exploration: Foundational scientific research is the cornerstone of biomedical discovery. Many of the most significant advances in medicine including vaccines, molecular diagnostics, gene editing technologies, and immunotherapies originated from curiosity driven investigations whose practical applications were not immediately apparent. Likewise, high-risk, high-reward research often challenges existing paradigms, explores unconventional hypotheses, and generates transformative breakthroughs that reshape entire scientific fields.

Despite its importance, foundational research frequently faces challenges within current evaluation systems. Investigators are often encouraged to pursue projects with predictable outcomes and a high likelihood of producing publications and grant renewals within relatively short timeframes. This environment may unintentionally discourage innovative ideas that require longer periods of development, involve greater uncertainty, or have a higher probability of negative or inconclusive findings.

NIH has long recognized the importance of supporting innovative science through programs that encourage high-risk, high-reward research. Continued success will require evaluation frameworks that acknowledge scientific ambition, creativity, and methodological excellence, even when research outcomes are uncertain.

Future evaluation systems should recognize contributions such as:

- Development of novel scientific hypotheses and conceptual frameworks.
- Creation of innovative research methodologies and experimental approaches.
- Exploration of emerging scientific fields with transformative potential.
- Long-term investment in foundational datasets, surveillance systems, and research infrastructure.
- Publication of well-designed negative or inconclusive findings that refine scientific understanding.
- Generation of resources that enable future scientific discoveries beyond the original research objectives.
- Development of interdisciplinary approaches that bridge multiple scientific domains.

Importantly, evaluation should distinguish between unsuccessful research resulting from poor scientific quality and well-designed research that produces unexpected, null, or negative findings. The latter represents an essential component of scientific progress and often prevents unnecessary duplication of effort while informing future investigations.

NIH may also consider rewarding researchers who contribute to long-term scientific infrastructure that benefits the broader research community. Establishing surveillance systems, longitudinal cohort studies, biobanks, reference laboratories, data repositories, and standardized methodologies often requires years of sustained investment before yielding substantial scientific outputs. These foundational resources support countless future investigations and should be recognized as enduring scientific contributions.

By valuing scientific curiosity, methodological innovation, and responsible exploration of challenging research questions, NIH can foster an environment that encourages bold ideas while maintaining the highest standards of scientific rigor and integrity.

7. Public Impact: The ultimate purpose of biomedical research is to improve human health. While scientific publications remain an important means of disseminating knowledge, the true impact of research should also be measured by its contribution to improving health outcomes, informing policy, strengthening healthcare systems, enhancing preparedness, reducing disease burden, and increasing

public trust in science.
Traditional academic metrics primarily measure scientific influence within the research community. Citation counts and publication metrics provide useful indicators of scholarly recognition but do not necessarily reflect whether research has improved clinical practice, informed health policy, strengthened disease surveillance, enhanced healthcare delivery, or benefited patients and communities.
NIH should incorporate broader measures of societal and public health impact into research evaluation. Recognizing these contributions would encourage investigators to pursue research that addresses important health challenges while promoting stronger connections between scientific discovery and real-world implementation.
Examples of measurable public impact may include:
• Adoption of research findings into national or international clinical guidelines and public health policies.
• Improvements in disease prevention, diagnosis, treatment, or patient outcomes resulting from research.
• Strengthening of laboratory systems, surveillance networks, and public health infrastructure.
• Contributions to emergency preparedness and response for emerging infectious diseases and other public health threats.
• Development of evidence that informs vaccination strategies, screening programs, or disease control interventions.
• Improvements in healthcare accessibility, quality, efficiency, or equity.
• Increased public confidence in science through transparent communication and responsible dissemination of research findings.
• Demonstrated societal, economic, or health system benefits attributable to scientific research.
Evaluation should also recognize that public impact often requires sustained engagement with policymakers, healthcare providers, public health authorities, patients, and communities. Activities such as knowledge translation, implementation science, public communication, and evidence-informed policymaking play essential roles in ensuring that scientific discoveries generate meaningful societal benefits.
Importantly, the impact of research should be assessed over appropriate timeframes. Some scientific discoveries produce immediate benefits, while others may influence healthcare practice or public policy years after publication. Flexible evaluation frameworks that recognize both short-term implementation and long-term societal contributions will provide a more comprehensive assessment of scientific value.
Ultimately, scientific success should be measured not only by the number of publications produced, but by the extent to which research improves lives, strengthens health systems, advances equity, informs decision-making, and contributes to healthier and more resilient societies.

8. Other Areas: Toward a Balanced Framework for Measuring Scientific Impact:
As biomedical research continues to evolve, evaluation systems should similarly evolve toward a more comprehensive and balanced approach to recognizing scientific contributions. Rather than relying predominantly on individual publication-based metrics, future assessment frameworks should incorporate multiple complementary dimensions that collectively reflect the diverse ways researchers contribute to scientific advancement.
A balanced evaluation framework should recognize excellence across the following domains:
• Scientific Quality: rigorous study design, reproducibility, methodological transparency, and high-quality publications.

• Open Science: responsible sharing of data, software, analytical methods, and research resources.
• Collaboration: meaningful contributions to multidisciplinary, multi-institutional, and international team science.
• Capacity Building: mentorship, workforce development, laboratory strengthening, and research infrastructure.
• Innovation and Translation: development of new technologies, implementation of research findings, and translation into clinical or public health practice.
• Public Impact: measurable improvements in health outcomes, policy development, preparedness, healthcare delivery, and societal benefit.

These domains should be viewed as complementary rather than competing measures of scientific excellence. Different disciplines, institutions, and career stages contribute to biomedical research in diverse ways, and evaluation systems should provide sufficient flexibility to recognize these differences while maintaining consistent standards of quality and accountability.

Importantly, no single metric can adequately capture scientific impact. Overreliance on publication counts or citation-based indicators may unintentionally discourage collaboration, open science, mentorship, implementation research, and other activities that generate substantial long-term value. A multidimensional evaluation framework would better align incentives with NIH's mission of generating knowledge that enhances health, lengthens life, and reduces illness and disability.

By recognizing a broader range of scientific contributions, NIH can strengthen research integrity, encourage collaboration, promote innovation, improve reproducibility, and maximize the public benefit derived from biomedical research investments.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Measuring-Scientific-Impact-Beyond-Publications_Safi.docx

Description: Measuring Scientific Impact Beyond Publications

Measuring Scientific Impact Beyond Publications:

Recommendations for Rewarding Rigorous, Collaborative, and Public Health-Oriented Research

Author

Dr. Hafizullah Safi

Kabul, Afghanistan

Independent Submission to the NIH Request for Information (NOT-OD-26-087)

hafeez.safi@gmail.com

July 2026

Disclaimer:

The views expressed in this document are solely those of the author and do not necessarily represent the views, policies, or positions of the World Health Organization (WHO), the Ministry of Public Health of Afghanistan, or any other affiliated institution.

Executive Summary:

Measuring Scientific Impact Beyond Publications: Recommendations for Rewarding Rigorous, Collaborative, and Public Health-Oriented Research:

Scientific discovery has traditionally been evaluated using indicators such as publication counts, citation metrics, journal impact factors, and grant funding. While these measures remain important, they no longer fully reflect the diverse contributions that drive modern biomedical research. Today's scientific enterprise is increasingly collaborative, multidisciplinary, data-intensive, and implementation-oriented. Many of the most valuable scientific contributions such as developing surveillance systems, sharing high-quality data, strengthening laboratory capacity, mentoring researchers, informing public health policy, and improving health outcomes are not adequately recognized by conventional evaluation metrics.

Based on experience in infectious disease surveillance, pandemic preparedness, implementation research, and international public health collaboration, this response recommends expanding the definition of scientific impact to better reflect contributions that improve research quality and public health. Scientific excellence should be measured not only by the production of new knowledge, but also by the reproducibility, accessibility, implementation, and real-world impact of that knowledge.

This response proposes that NIH adopt a more balanced framework for evaluating scientific contributions by recognizing excellence across multiple complementary domains, including:

- Scientific rigor and reproducibility.
- Open sharing of data, software, protocols, and analytical methods.
- Capacity building through training and mentorship.
- Contributions to interdisciplinary and international team science.
- Translation of research into policy, clinical practice, and public health action.
- Support for foundational and high-risk scientific exploration; and
- Demonstrable improvements in population health, preparedness, and societal benefit.

The recommendations presented in this document emphasize practical and measurable indicators that can be applied across different research disciplines, career stages, and institutional settings. They are intended to complement, rather than replace, traditional academic metrics while encouraging behaviors that strengthen scientific integrity, accelerate discovery, promote collaboration, and maximize the return on public investment in biomedical research.

Ultimately, scientific success should be defined not only by what researchers publish, but by how their work advances knowledge, strengthens health systems, informs decision-making, improves public health, and benefits society.

Key Recommendations to NIH

To better align research evaluation with the realities of modern biomedical science, NIH may consider the following recommendations:

1. Move Beyond Publication-Centric Metrics

Complement traditional indicators such as publication counts, journal impact factors, and citation metrics with broader measures that capture the full spectrum of scientific contributions and real-world impact.

2. Strengthen Recognition of Scientific Rigor and Reproducibility

Reward researchers who demonstrate methodological excellence through transparent reporting, protocol registration, replication studies, validation of findings, and adherence to established reporting standards.

3. Recognize Open Science as a Core Scientific Contribution

Incentivize responsible sharing of research data, software, analytical code, protocols, and other scientific resources that improve transparency, reproducibility, and research efficiency.

4. Reward High-Quality Data Stewardship and Resource Sharing

Recognize the development, maintenance, and sharing of high-quality datasets, surveillance systems, biorepositories, computational tools, and other research infrastructure that enable future scientific discovery.

5. Promote and Reward Team Science

Develop evaluation approaches that recognize meaningful contributions to multidisciplinary, multi-institutional, and international collaborative research, regardless of traditional authorship position.

6. Value Training, Mentorship, and Capacity Building

Recognize researchers who contribute to developing the next generation of scientists through mentorship, workforce development, laboratory strengthening, and sustainable research capacity.

7. Expand Recognition of Translation and Implementation

Reward research that successfully informs public health policy, clinical practice, healthcare delivery, implementation science, technology development, and evidence-based decision-making.

8. Measure Public Health and Societal Impact

Incorporate indicators that evaluate how research contributes to improving health outcomes, strengthening health systems, enhancing preparedness, reducing disease burden, and increasing public trust in science.

9. Support Foundational and High-Risk Scientific Exploration

Continue to encourage innovative, curiosity-driven, and high-risk research by recognizing scientific quality, creativity, and long-term contributions, even when immediate outcomes are uncertain.

10. Adopt a Multidimensional Framework for Scientific Impact

Implement a balanced evaluation framework that integrates scientific quality, open science, collaboration, capacity building, translation, and public impact to provide a more comprehensive assessment of scientific excellence across career stages, disciplines, and institutional settings.

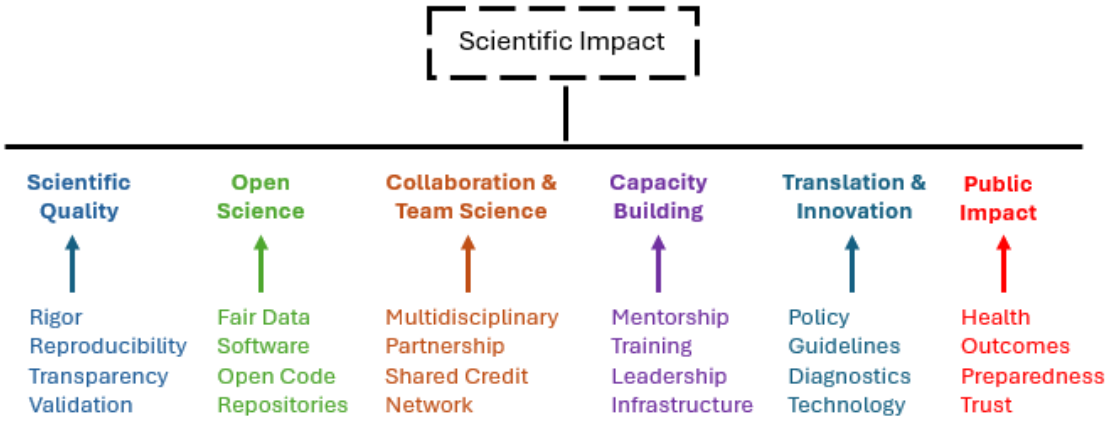
The future of biomedical research depends not only on generating new scientific knowledge, but also on ensuring that this knowledge is rigorous, reproducible, openly shared, collaboratively developed, effectively translated into practice, and ultimately improves human health. By adopting a more comprehensive and balanced approach to measuring scientific impact, NIH can strengthen research integrity, foster innovation, promote collaboration, and maximize the public value of biomedical research investments.

Table of Contents

Executive Summary: 2
Key Recommendations to NIH 3
Figure 1. A Conceptual Framework 6
Table 1. Suggested Indicators 6
Table 2. Proposed Balanced Scientific Impact Assessment Framework 8
Introduction 8
Section 1. Rigor and Reproducibility 9
Section 2. Data, Software, and Model Sharing 10
Section 3. Training and Mentorship 12
Section 4. Collaboration 13
Section 5. Entrepreneurship and Translation 15
Section 6. Foundational Scientific Exploration 16
Section 7. Public Impact 17
Section 8. Other Areas: Toward a Balanced Framework for Measuring Scientific Impact 19
Considerations 20
Conclusion 20
References: 21

The six domains presented below represent complementary dimensions of modern biomedical research. Together, they encompass the full research lifecycle—from rigorous knowledge generation and scientific quality to open science, collaboration, translation, implementation, and measurable public health impact.

Figure 1. A Conceptual Framework for Measuring Scientific Impact in Modern Biomedical Research.



Traditional publication-based metrics should be complemented by six interconnected domains that collectively capture the diverse contributions required to advance biomedical discovery and improve public health.

Scientific impact should be evaluated as a multidimensional construct rather than through any single metric.

Table 1. Suggested Indicators for Measuring Scientific Impact Across the Six Domains

Domain	Examples of Scientific Contributions	Suggested Measurable Indicators
1. Scientific Quality	Rigorous study design, reproducibility, methodological transparency, validation studies	Evidence of reproducible research practices; independent validation of findings; adherence to reporting guidelines (CONSORT, STROBE, PRISMA, etc.); protocol registration; quality assurance measures; publication of negative findings
2. Open Science	Data sharing, software development, analytical pipelines, research protocols	FAIR data compliance; publicly available datasets; software/code repositories; dataset DOI citations; number of external

		dataset reuses; availability of statistical code and protocols
3. Collaboration & Team Science	Multidisciplinary research, international partnerships, shared infrastructure	Number of multi-institutional projects; multidisciplinary publications; international collaborations; contribution statements (CRediT taxonomy); leadership of collaborative networks; long-term partnerships
4. Capacity Building	Mentorship, workforce development, laboratory strengthening, research infrastructure	Number of trainees successfully mentored; trainee career progression; training programs developed; laboratory accreditation achieved; surveillance systems strengthened; institutional research capacity established
5. Translation & Innovation	Technology development, implementation science, clinical or public health application	Research incorporated into clinical guidelines or public health policy; patents; diagnostic tools developed; software adopted; implementation of new interventions; technology transfer activities
6. Public Impact	Health improvement, policy influence, preparedness, societal benefit	Health outcomes improved; disease burden reduced; policy adoption; surveillance performance improved; healthcare quality enhanced; public trust initiatives; preparedness indicators; cost savings to health systems

Table 2. Proposed Balanced Scientific Impact Assessment Framework

Domain	Primary Evaluation Focus	Illustrative Relative Weight*
Scientific Quality	Rigor, reproducibility, methodological excellence	25%
Open Science	Data, software, protocols, transparency	15%
Collaboration & Team Science	Interdisciplinary and international partnerships	15%
Capacity Building	Mentorship, workforce development, infrastructure	15%
Translation & Innovation	Policy, clinical implementation, commercialization	15%
Public Impact	Health outcomes, preparedness, societal benefit	15%

The framework and indicators presented above provide a practical foundation for the recommendations discussed in the following sections. Each domain is explored in greater detail with suggested approaches for strengthening research evaluation and recognizing scientific contributions.

Introduction

The biomedical research landscape has changed dramatically over the past several decades. Scientific discovery has evolved from predominantly individual laboratory-based research into a highly collaborative, multidisciplinary, and globally connected enterprise. Advances in genomics, artificial intelligence, digital health, large-scale surveillance systems, and international research partnerships have transformed the way scientific knowledge is generated, validated, and translated into public health action. These changes have created new opportunities for accelerating biomedical discovery while also highlighting the need for research systems that prioritize scientific rigor, transparency, reproducibility, and collaboration.

Despite these advances, the methods used to evaluate scientific success have changed relatively little. Publications, citation counts, journal impact factors, and grant funding continue to serve as the primary indicators of scientific achievement. Although these metrics provide useful information regarding research productivity and scholarly influence, they often fail to capture many of the activities that are essential for advancing modern biomedical science. Contributions such as data sharing, development of research infrastructure, implementation of surveillance systems, mentorship, multidisciplinary

collaboration, capacity building, policy translation, and improvements in population health frequently receive limited recognition despite their substantial scientific and societal value.

The COVID-19 pandemic clearly demonstrated that impactful science extends far beyond scientific publications. Rapid international data sharing, open access to genomic sequences, collaborative research networks, coordinated laboratory systems, and transparent communication enabled the global scientific community to characterize a novel pathogen, develop diagnostic methods, evaluate interventions, and accelerate vaccine development within an unprecedented timeframe. These achievements were possible because researchers, public health institutions, governments, and international organizations worked together under principles of openness, collaboration, and shared scientific responsibility.

As biomedical research continues to evolve, evaluation systems should similarly evolve to recognize the broad range of contributions that advance scientific knowledge and improve public health. Measuring scientific impact should not only assess the generation of new discoveries but also the quality, reproducibility, accessibility, implementation, and real-world influence of those discoveries.

The following recommendations propose practical approaches that NIH may consider when developing future frameworks for measuring and rewarding scientific impact. These recommendations aim to complement traditional academic metrics by incorporating indicators that encourage rigorous, collaborative, transparent, and mission-driven biomedical research.

Section 1. Rigor and Reproducibility

Scientific rigor and reproducibility form the foundation of credible biomedical research. Regardless of scientific discipline or methodology, research findings should be sufficiently transparent and well documented to enable independent researchers to reproduce analyses, verify results, and build upon previous discoveries. Reproducibility not only strengthens confidence in scientific findings but also improves the efficiency of research investments by reducing unnecessary duplication, identifying methodological limitations, and accelerating cumulative scientific progress.

Current academic reward systems often prioritize the novelty of scientific findings over the quality and reproducibility of research methods. Consequently, researchers may receive greater recognition for publishing new discoveries than for validating existing findings, conducting replication studies, improving analytical methods, or producing high-quality datasets that support future research. Although replication studies and negative findings are essential components of the scientific process, they are frequently undervalued in traditional research assessment systems.

NIH has taken important steps through its Replication and Reproducibility Initiative to strengthen scientific integrity. Continued progress, however, will require evaluation

systems that actively recognize researchers who consistently demonstrate rigorous scientific practices throughout the research lifecycle.

Examples of activities that should receive greater recognition include:

- Publishing detailed study protocols and statistical analysis plans before data analysis.
- Making analytical code, algorithms, and computational workflows openly available whenever appropriate.
- Sharing de-identified research datasets and associated metadata in accordance with ethical and legal requirements.
- Conducting independent validation and replication studies.
- Reporting negative or inconclusive findings that contribute to scientific understanding and reduce publication bias.
- Using standardized reporting guidelines and internationally recognized research standards.
- Implementing transparent quality assurance and quality control procedures throughout data collection and analysis.
- Participating in collaborative efforts to reproduce important scientific findings across different populations and settings.

NIH may also consider incorporating measurable indicators of research rigor into funding and career evaluation processes. Examples include the availability of reproducible analytical workflows, documented data quality procedures, public access to research protocols, successful independent validation of findings, adherence to FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, and demonstrated compliance with established reporting standards.

Importantly, promoting reproducibility should not be viewed as an administrative requirement but rather as an investment in scientific quality. Rewarding researchers who prioritize methodological excellence, transparency, and validation will strengthen public trust, improve the reliability of biomedical evidence, and increase the long-term value of publicly funded research. A research culture that values reproducibility equally with innovation will produce discoveries that are more robust, more transferable across populations, and more likely to generate lasting scientific and public health impact.

Section 2. Data, Software, and Model Sharing

The value of biomedical research extends well beyond scientific publications. Increasingly, the greatest impact of research is realized through the sharing of high-quality datasets, analytical methods, software, computational models, and research tools that enable other

investigators to validate findings, generate new hypotheses, and accelerate scientific discovery. Open and responsible sharing of scientific resources enhances transparency, improves reproducibility, reduces duplication of effort, and maximizes the return on public investment in research.

The COVID-19 pandemic clearly demonstrated the importance of rapid and transparent data sharing. International collaboration through surveillance networks, genomic databases, and open scientific platforms enabled researchers worldwide to monitor pathogen evolution, assess disease transmission, develop diagnostics, evaluate interventions, and accelerate vaccine development. These achievements highlighted that scientific progress depends not only on producing data, but also on making data accessible, interoperable, and reusable by the broader research community.

Despite these benefits, many researchers receive limited academic recognition for developing, curating, maintaining, and sharing valuable scientific datasets or software. Considerable effort is often required to establish surveillance systems, harmonize data standards, ensure data quality, document metadata, develop analytical tools, and maintain long-term databases. These contributions frequently require years of sustained investment but are rarely reflected in conventional measures of scientific productivity.

NIH should recognize data stewardship and scientific resource development as core research contributions. Investigators who make high-quality datasets, software packages, computational pipelines, and predictive models openly available should receive formal recognition comparable to traditional scholarly outputs. Such recognition would encourage broader participation in open science while improving the reproducibility and efficiency of biomedical research.

Future evaluation frameworks could include measurable indicators such as:

- Timely sharing of de-identified research data in accordance with ethical, legal, and regulatory requirements.
- Compliance with FAIR (Findable, Accessible, Interoperable, and Reusable) data principles.
- Development and maintenance of high-quality scientific databases, registries, surveillance systems, and biorepositories.
- Public availability of analytical software, statistical code, machine learning algorithms, and computational workflows.
- Documentation quality, metadata completeness, and standardized data dictionaries.
- Frequency of dataset reuse by independent investigators.
- Integration of shared datasets into national or international scientific platforms.
- Long-term sustainability and accessibility of shared scientific resources.

NIH should also encourage the citation of datasets, software, and digital research products as independent scholarly outputs. Standardized mechanisms for assigning persistent identifiers, such as Digital Object Identifiers (DOIs), can improve attribution and enable these contributions to be tracked and evaluated alongside traditional publications.

Importantly, data sharing should be implemented responsibly. Policies must continue to safeguard participant privacy, protect confidential information, respect national regulations, and recognize the rights of data-generating institutions and investigators. In particular, researchers working in low- and middle-income countries should receive appropriate authorship, acknowledgment, and opportunities for scientific leadership when their data contribute to international collaborative research.

By recognizing data, software, and model sharing as essential scientific outputs, NIH can foster a research culture that values openness, collaboration, transparency, and long-term scientific utility. Such an approach will accelerate biomedical discovery while strengthening public trust and maximizing the societal benefits of federally funded research.

Section 3. Training and Mentorship

The long-term success of biomedical research depends not only on scientific discoveries but also on the development of a skilled, diverse, and sustainable research workforce. Training and mentorship are fundamental investments that strengthen scientific capacity, promote innovation, improve research quality, and ensure the continuity of scientific excellence across generations. However, these contributions are often underrepresented in current systems used to evaluate scientific impact.

Effective mentorship extends beyond supervising research projects. High-quality mentors cultivate scientific integrity, critical thinking, leadership, collaboration, communication skills, and responsible research practices. They provide guidance that enables trainees to become independent investigators while fostering inclusive and supportive research environments. Similarly, structured training programs that enhance technical competencies, research methodology, data analysis, laboratory quality management, implementation science, and scientific writing contribute substantially to strengthening the biomedical research enterprise.

Current evaluation systems frequently recognize publications and grant income while giving comparatively little credit to sustained mentorship and workforce development. As a result, researchers who dedicate significant time to training early-career scientists, building institutional capacity, or strengthening research infrastructure may receive limited recognition despite their substantial long-term impact.

NIH should incorporate measurable indicators that recognize excellence in training and mentorship. Potential indicators include:

- Successful career progression of trainees into independent research or leadership positions.
- Contributions to multidisciplinary training programs and workforce development initiatives.
- Development of educational resources, curricula, and professional training materials.
- Mentorship of early-career investigators, particularly those from underrepresented or resource-limited settings.
- Establishment of sustainable institutional research capacity.
- Evidence of long-term mentoring relationships and continued professional development of trainees.
- Contributions to strengthening laboratory systems, surveillance capacity, research ethics, and implementation science.
- Recognition received from trainees, institutions, or professional organizations for mentoring excellence.

Importantly, evaluation should emphasize the quality and long-term outcomes of mentorship rather than simply the number of trainees supervised. A single mentor who successfully develops future scientific leaders may have a greater and more lasting impact than one who supervises many trainees without meaningful professional development.

Investment in mentorship also promotes scientific integrity and reproducibility. Experienced mentors play a critical role in teaching rigorous study design, ethical conduct of research, transparent reporting, responsible data management, and collaborative scientific practices. Strengthening these competencies early in researchers' careers contributes to a more reliable, trustworthy, and resilient biomedical research ecosystem.

Recognizing mentorship and workforce development as essential scientific contributions will encourage researchers to invest in future generations of scientists while strengthening global research capacity and ensuring the long-term sustainability of biomedical innovation.

Section 4. Collaboration

Biomedical research has become increasingly collaborative, interdisciplinary, and international. Addressing today's most pressing health challenges including emerging infectious diseases, antimicrobial resistance, pandemic preparedness, precision medicine, and chronic diseases requires expertise that extends far beyond any single investigator, institution, or scientific discipline. Successful research increasingly depends on effective collaboration among clinicians, epidemiologists, laboratory scientists, data scientists, statisticians, social scientists, engineers, public health practitioners, policymakers, and community stakeholders.

Despite this transformation, research evaluation systems continue to emphasize individual accomplishments, often relying heavily on first or senior authorship, principal investigator status, and individual grant awards. Such approaches may overlook the substantial scientific contributions made by investigators whose expertise is essential to collaborative research but whose roles are less visible in traditional academic metrics. Data managers, surveillance specialists, laboratory scientists, statisticians, software developers, implementation researchers, and project coordinators frequently make indispensable contributions that are difficult to quantify under current evaluation systems.

NIH should further strengthen mechanisms for recognizing meaningful contributions to team science. Evaluation should move beyond publication authorship alone and consider the quality, significance, and impact of each individual's contribution throughout the research process. Modern research benefits when multidisciplinary teams work together to design studies, generate high-quality data, develop analytical methods, interpret findings, and translate evidence into practice.

Potential indicators for recognizing collaborative research may include:

- Participation in multidisciplinary and multi-institutional research projects.
- Contributions to international scientific collaborations and research networks.
- Leadership or coordination of collaborative research initiatives.
- Development of shared research infrastructure, surveillance systems, or laboratory networks.
- Contributions to multicenter studies, implementation research, and consortium-based projects.
- Evidence of equitable partnerships that promote shared leadership and capacity strengthening.
- Collaborative development of standardized protocols, data standards, and quality assurance systems.
- Contributions to cross-disciplinary innovation that integrates complementary areas of expertise.

Evaluation systems should also recognize that impactful collaboration extends beyond academic institutions. Partnerships involving government agencies, public health organizations, healthcare providers, industry, community organizations, patient groups, and international partners often play a critical role in translating scientific discoveries into meaningful health outcomes. These collaborations strengthen the relevance, applicability, and sustainability of biomedical research while increasing its societal impact.

Importantly, collaborative research should be evaluated based on both scientific productivity and the effectiveness of the partnership itself. Indicators such as sustained

collaboration, equitable authorship practices, shared decision-making, successful knowledge exchange, and measurable improvements in research capacity may provide valuable insight into the long-term success of collaborative initiatives.

By rewarding meaningful collaboration, NIH can encourage researchers to build diverse scientific teams, share expertise across disciplines, reduce unnecessary duplication of effort, and accelerate the translation of scientific discoveries into improved health outcomes.

Section 5. Entrepreneurship and Translation

Scientific discovery achieves its greatest value when research findings are translated into interventions, technologies, policies, diagnostics, therapeutics, and public health practices that improve human health. While fundamental discovery remains essential, evaluation systems should also recognize the processes that move scientific knowledge beyond publication and into real-world application.

Translation is rarely a single event. It involves a continuum of activities including technology development, validation studies, implementation research, regulatory evaluation, commercialization, policy development, clinical adoption, and health system integration. Each stage requires substantial scientific expertise and often involves multidisciplinary collaboration across academia, healthcare, government, and industry. However, many of these activities receive relatively limited recognition within traditional academic evaluation frameworks.

NIH should broaden the definition of translational impact to encompass the full spectrum of activities that convert scientific knowledge into measurable public benefit. Successful translation should not be assessed solely by patents or commercial products but also by the adoption of evidence-based practices, improvements in healthcare delivery, strengthening of public health systems, and incorporation of research findings into national or international guidelines.

Potential indicators of translational impact include:

- Adoption of research findings into clinical practice guidelines or public health policies.
- Development and implementation of new diagnostic methods, surveillance tools, or healthcare technologies.
- Creation of software platforms, digital health applications, decision-support systems, or analytical tools used in healthcare settings.
- Commercialization of biomedical innovations that improve patient care.
- Establishment of public-private partnerships supporting technology development.

- Successful implementation of research findings within healthcare systems or public health programs.
- Demonstrated improvements in healthcare quality, efficiency, or accessibility resulting from research.
- Contribution of scientific evidence to regulatory decisions or national health strategies.

Entrepreneurship should likewise be recognized as an important component of biomedical innovation. Researchers who successfully transform scientific discoveries into scalable solutions frequently contribute to economic development, job creation, technology transfer, and improved healthcare delivery. These activities complement rather than compete with traditional academic research and should be evaluated within the broader context of NIH's mission to improve health.

Importantly, translational success should also include innovations that primarily benefit public health rather than commercial markets. Research that strengthens disease surveillance, enhances laboratory systems, improves outbreak detection, supports emergency preparedness, or informs health policy may generate substantial societal value even when no commercial product results.

Recognizing diverse pathways of translation will encourage researchers to pursue innovations that address unmet health needs while ensuring that publicly funded biomedical research delivers meaningful and measurable benefits to patients, healthcare systems, and society.

Section 6. Foundational Scientific Exploration

Foundational scientific research is the cornerstone of biomedical discovery. Many of the most significant advances in medicine including vaccines, molecular diagnostics, gene editing technologies, and immunotherapies originated from curiosity driven investigations whose practical applications were not immediately apparent. Likewise, high-risk, high-reward research often challenges existing paradigms, explores unconventional hypotheses, and generates transformative breakthroughs that reshape entire scientific fields.

Despite its importance, foundational research frequently faces challenges within current evaluation systems. Investigators are often encouraged to pursue projects with predictable outcomes and a high likelihood of producing publications and grant renewals within relatively short timeframes. This environment may unintentionally discourage innovative ideas that require longer periods of development, involve greater uncertainty, or have a higher probability of negative or inconclusive findings.

NIH has long recognized the importance of supporting innovative science through programs that encourage high-risk, high-reward research. Continued success will require evaluation

frameworks that acknowledge scientific ambition, creativity, and methodological excellence, even when research outcomes are uncertain.

Future evaluation systems should recognize contributions such as:

- Development of novel scientific hypotheses and conceptual frameworks.
- Creation of innovative research methodologies and experimental approaches.
- Exploration of emerging scientific fields with transformative potential.
- Long-term investment in foundational datasets, surveillance systems, and research infrastructure.
- Publication of well-designed negative or inconclusive findings that refine scientific understanding.
- Generation of resources that enable future scientific discoveries beyond the original research objectives.
- Development of interdisciplinary approaches that bridge multiple scientific domains.

Importantly, evaluation should distinguish between unsuccessful research resulting from poor scientific quality and well-designed research that produces unexpected, null, or negative findings. The latter represents an essential component of scientific progress and often prevents unnecessary duplication of effort while informing future investigations.

NIH may also consider rewarding researchers who contribute to long-term scientific infrastructure that benefits the broader research community. Establishing surveillance systems, longitudinal cohort studies, biobanks, reference laboratories, data repositories, and standardized methodologies often requires years of sustained investment before yielding substantial scientific outputs. These foundational resources support countless future investigations and should be recognized as enduring scientific contributions.

By valuing scientific curiosity, methodological innovation, and responsible exploration of challenging research questions, NIH can foster an environment that encourages bold ideas while maintaining the highest standards of scientific rigor and integrity.

Section 7. Public Impact

The ultimate purpose of biomedical research is to improve human health. While scientific publications remain an important means of disseminating knowledge, the true impact of research should also be measured by its contribution to improving health outcomes, informing policy, strengthening healthcare systems, enhancing preparedness, reducing disease burden, and increasing public trust in science.

Traditional academic metrics primarily measure scientific influence within the research community. Citation counts and publication metrics provide useful indicators of scholarly recognition but do not necessarily reflect whether research has improved clinical practice, informed health policy, strengthened disease surveillance, enhanced healthcare delivery, or benefited patients and communities.

NIH should incorporate broader measures of societal and public health impact into research evaluation. Recognizing these contributions would encourage investigators to pursue research that addresses important health challenges while promoting stronger connections between scientific discovery and real-world implementation.

Examples of measurable public impact may include:

- Adoption of research findings into national or international clinical guidelines and public health policies.
- Improvements in disease prevention, diagnosis, treatment, or patient outcomes resulting from research.
- Strengthening of laboratory systems, surveillance networks, and public health infrastructure.
- Contributions to emergency preparedness and response for emerging infectious diseases and other public health threats.
- Development of evidence that informs vaccination strategies, screening programs, or disease control interventions.
- Improvements in healthcare accessibility, quality, efficiency, or equity.
- Increased public confidence in science through transparent communication and responsible dissemination of research findings.
- Demonstrated societal, economic, or health system benefits attributable to scientific research.

Evaluation should also recognize that public impact often requires sustained engagement with policymakers, healthcare providers, public health authorities, patients, and communities. Activities such as knowledge translation, implementation science, public communication, and evidence-informed policymaking play essential roles in ensuring that scientific discoveries generate meaningful societal benefits.

Importantly, the impact of research should be assessed over appropriate timeframes. Some scientific discoveries produce immediate benefits, while others may influence healthcare practice or public policy years after publication. Flexible evaluation frameworks that recognize both short-term implementation and long-term societal contributions will provide a more comprehensive assessment of scientific value.

Ultimately, scientific success should be measured not only by the number of publications produced, but by the extent to which research improves lives, strengthens health systems, advances equity, informs decision-making, and contributes to healthier and more resilient societies.

Section 8. Other Areas: Toward a Balanced Framework for Measuring Scientific Impact

As biomedical research continues to evolve, evaluation systems should similarly evolve toward a more comprehensive and balanced approach to recognizing scientific contributions. Rather than relying predominantly on individual publication-based metrics, future assessment frameworks should incorporate multiple complementary dimensions that collectively reflect the diverse ways researchers contribute to scientific advancement.

A balanced evaluation framework should recognize excellence across the following domains:

- **Scientific Quality:** rigorous study design, reproducibility, methodological transparency, and high-quality publications.
- **Open Science:** responsible sharing of data, software, analytical methods, and research resources.
- **Collaboration:** meaningful contributions to multidisciplinary, multi-institutional, and international team science.
- **Capacity Building:** mentorship, workforce development, laboratory strengthening, and research infrastructure.
- **Innovation and Translation:** development of new technologies, implementation of research findings, and translation into clinical or public health practice.
- **Public Impact:** measurable improvements in health outcomes, policy development, preparedness, healthcare delivery, and societal benefit.

These domains should be viewed as complementary rather than competing measures of scientific excellence. Different disciplines, institutions, and career stages contribute to biomedical research in diverse ways, and evaluation systems should provide sufficient flexibility to recognize these differences while maintaining consistent standards of quality and accountability.

Importantly, no single metric can adequately capture scientific impact. Overreliance on publication counts or citation-based indicators may unintentionally discourage collaboration, open science, mentorship, implementation research, and other activities that generate substantial long-term value. A multidimensional evaluation framework would better align incentives with NIH's mission of generating knowledge that enhances health, lengthens life, and reduces illness and disability.

By recognizing a broader range of scientific contributions, NIH can strengthen research integrity, encourage collaboration, promote innovation, improve reproducibility, and maximize the public benefit derived from biomedical research investments.

Considerations

The recommendations presented in this document are intended to provide general principles for strengthening research evaluation across diverse biomedical disciplines. The relative importance of individual indicators will vary according to research field, career stage, institutional mission, and funding mechanism. Accordingly, a flexible and context-specific implementation approach will be essential.

Conclusion

Biomedical research has entered an era characterized by unprecedented collaboration, technological innovation, data generation, and global scientific connectivity. While traditional measures of research productivity remain valuable, they should be complemented by evaluation approaches that more accurately reflect the diverse activities required to generate trustworthy scientific knowledge and translate discoveries into meaningful improvements in health.

The recommendations presented in this response are intended to support NIH's ongoing efforts to strengthen rigor, reproducibility, transparency, collaboration, and public trust in biomedical research. Expanding evaluation frameworks to recognize data sharing, team science, mentorship, implementation, innovation, and public health impact will encourage behaviors that improve both the quality and the societal value of scientific research.

Ultimately, the success of biomedical research should be measured not solely by the number of publications produced, but by the quality, reproducibility, accessibility, implementation, and lasting impact of scientific knowledge. A more balanced and inclusive approach to research assessment will better recognize the full spectrum of contributions that advance discovery, improve health, and strengthen the global biomedical research enterprise.

References

Declaration on Research Assessment (DORA). (2012). *San Francisco Declaration on Research Assessment (DORA).* https://sfdora.org/

Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? *How the Contributor Roles Taxonomy (CRediT) is helping the shift from authorship to contributorship. Learned Publishing, 32*(1), 71–74. https://doi.org/10.1002/leap.1210

National Institutes of Health. (2026). *Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087).* U.S. Department of Health and Human Services. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

National Institutes of Health. (2024). *NIH Replication and Reproducibility Initiative.* U.S. Department of Health and Human Services. https://www.nih.gov/

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c869. https://doi.org/10.1136/bmj.c869

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *PLoS Medicine, 4*(10), e296. https://doi.org/10.1371/journal.pmed.0040296

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429–431. https://doi.org/10.1038/520429a

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., Coriat, A. M., Foeger, N., & Dirnagl, U. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLOS Biology, 18*(7), e3000737. https://doi.org/10.1371/journal.pbio.3000737

World Health Organization. (2021). *Global strategy on digital health 2020–2025.* World Health Organization. https://www.who.int/publications/i/item/9789240020924

National Academies of Sciences, Engineering, and Medicine. (2017). *Fostering integrity in research.* National Academies Press. https://doi.org/10.17226/21896

95. JP Flores

I am responding: On behalf of myself

Name: JP Flores

Name of Organization: Case Western Reserve University

Type of Organization: Academic Institution

Role: Research Trainee/Fellow

Date Submitted: 7/16/2026

1. Rigor and Reproducibility: Scientific rigor should be recognized as a primary research output rather than a compliance requirement. While biomedical research has traditionally rewarded novelty, equally important activities, including replication studies, validation experiments, publication of negative or null findings, benchmarking analyses, and methodological refinement, remain undervalued despite their essential role in building a trustworthy scientific literature.

NIH should develop measurable indicators that recognize contributions to reproducibility, including publication of independent replication studies, release of reproducible computational workflows, transparent reporting of methods and quality-control metrics, reanalysis of publicly available datasets, and development of benchmark datasets or community standards. Well-designed studies that produce negative or null results should be recognized as valuable scientific contributions because they prevent unnecessary duplication, refine biological hypotheses, and improve future experimental design.

To further incentivize rigorous science, NIH should consider dedicated funding mechanisms for replication studies, administrative supplements to validate influential findings, and grant review criteria that explicitly reward reproducibility efforts. Expectations should remain flexible across disciplines and career stages while recognizing that rigorous science is an ongoing process of refining knowledge rather than simply producing novel discoveries.

2. Data, Software, and Model Sharing: Data, software, computational workflows, and predictive models have become foundational products of modern biomedical research. Yet these resources often receive substantially less recognition than traditional publications despite enabling thousands of downstream discoveries.

NIH should explicitly recognize open-source software, databases, computational workflows, interoperable standards, and publicly accessible datasets as primary scholarly outputs. Indicators of impact should include software adoption, community contributions, dataset reuse, workflow implementation, API usage, long-term maintenance, and compliance with FAIR (Findable, Accessible, Interoperable, Reusable) data principles rather than publication citations alone.

Importantly, NIH should establish dedicated funding mechanisms to support the development, maintenance, and long-term sustainability of open-source scientific software and infrastructure. Many of

the most widely used computational tools in biomedical research are developed and maintained by small research groups with limited long-term support despite serving thousands of investigators worldwide. Sustaining these resources should be recognized as a valuable scientific contribution, not merely technical service.

3. Training and Mentorship: The future strength of the biomedical research enterprise depends not only on producing excellent science but also on preparing scientists for the increasingly diverse roles they play in society. Modern scientists serve as mentors, educators, communicators, entrepreneurs, software developers, policy advisors, and community partners in addition to conducting research.

NIH should broaden how mentorship and training are evaluated by recognizing contributions that prepare trainees for this evolving landscape. Potential indicators include trainee career outcomes across academia, industry, government, education, policy, entrepreneurship, nonprofit organizations, and clinical practice; evidence-based mentoring practices; trainee independence; structured mentorship programs; and mentoring climate.

Training programs should also incentivize skills that are increasingly essential to modern science but are often treated as optional. These include science communication, science policy, community engagement, entrepreneurship, open science practices, data stewardship, peer review, and interdisciplinary collaboration. Investigators who train students in transparent scientific practices, including preprinting, reproducible computational workflows, responsible data sharing, accessible communication, and constructive peer review, should receive formal recognition for these contributions.

Scientific excellence should include preparing the next generation of scientists not only to conduct outstanding research but also to effectively communicate, translate, and steward that research for the benefit of society.

4. Collaboration: Biomedical research increasingly relies on interdisciplinary teams that span institutions, sectors, and scientific disciplines. However, many evaluation systems continue to emphasize individual accomplishments rather than collaborative contributions.

NIH should recognize meaningful contributions to team science regardless of publication authorship position by considering leadership of collaborative projects, development of shared infrastructure, multi-investigator grants, co-mentorship, community resource development, and documented contributions through narrative team science statements or contributor role taxonomies.

In addition, NIH should expand funding mechanisms that intentionally build scientific capacity through collaborations with Minority Serving Institutions (MSIs), Historically Black Colleges and Universities (HBCUs), Tribal Colleges and Universities (TCUs), community colleges, primarily undergraduate institutions, and emerging research institutions. These partnerships should be structured as reciprocal scientific collaborations that build long-term research capacity rather than short-term subcontracting arrangements.

Such investments would not only strengthen scientific discovery but also broaden participation in biomedical research while expanding research capacity throughout the United States.

5. Entrepreneurship and Translation: Translation should be viewed as a bidirectional process rather than a one-way pathway from bench to bedside. Basic scientists should have greater opportunities to move discoveries toward clinical implementation, while clinicians, patients, engineers, public health practitioners, policymakers, and industry partners should have meaningful opportunities to inform fundamental biological discovery.

NIH should recognize diverse forms of translation, including clinical implementation, public health impact, policy translation, development of diagnostics and research tools, open-source technologies, nonprofit partnerships, educational innovations, and commercialization where appropriate.

Expanding funding mechanisms that intentionally connect basic science laboratories with clinicians, public health practitioners, patient organizations, entrepreneurs, and industry partners would strengthen the entire translational ecosystem. These partnerships should be viewed as complementary approaches for maximizing the public benefit of biomedical research rather than competing models of scientific impact.

6. Foundational Scientific Exploration: Many of the most transformative advances in biomedical research emerge from high-risk, high-reward investigations that require sustained investment and may not produce immediate publications or translational outcomes. Current incentive structures often favor predictable projects with short-term deliverables over ambitious questions capable of fundamentally changing scientific understanding.

NIH should continue expanding support for foundational scientific exploration by recognizing long-term infrastructure development, technology innovation, biological resource generation, atlas-building efforts, methodological advances, and exploratory research that creates new scientific directions.

Review criteria should emphasize scientific significance, originality, and long-term potential rather than short-term publication metrics alone. Investigators should not be penalized for pursuing ambitious projects whose greatest impacts may emerge years after the initial funding period.

7. Public Impact: The ultimate purpose of biomedical research is to improve human health and benefit society. Accordingly, public impact should become a core dimension of scientific excellence.

Public impact should extend beyond counting outreach activities or media appearances. Instead, NIH should recognize sustained, reciprocal relationships between scientists and the communities they serve.

Potential indicators include long-term partnerships with patient advocacy organizations, community advisory boards, collaborations with K-12 schools and community colleges, publicly accessible educational resources, contributions to evidence-informed policymaking, citizen science initiatives, implementation of research findings into clinical practice or public health, and measurable efforts to strengthen public trust in science.

NIH should also recognize contributions that improve the accessibility and transparency of science itself, including open educational resources, accessible software, publicly available datasets, preprints, transparent communication of findings, and efforts to make biomedical knowledge broadly

understandable and reusable.

The goal should not simply be increasing public awareness of science but strengthening enduring relationships between scientific institutions and the communities they exist to serve.

8. Other Areas: One area deserving additional consideration is modernization of scientific publishing itself. The current publication system remains one of the strongest drivers of scientific incentives, yet many activities essential to scientific quality remain largely invisible during evaluation.

NIH has an opportunity to encourage a transition toward more open, transparent, and efficient models of scientific communication by explicitly recognizing preprints, open peer review, and emerging Publish, Review, Curate (PRC) models. These approaches separate dissemination, evaluation, and curation into distinct scholarly activities, allowing discoveries to become immediately accessible while expert review and long-term curation continue transparently.

NIH should consider allowing investigators to document preprints, citable peer reviews, open review activities, and community curation efforts as evidence of scholarly impact. Likewise, reviewer contributions, editorial leadership, and development of community standards should be recognized as meaningful scientific outputs rather than invisible service activities.

Finally, NIH should consider adopting a broader Scientific Impact Portfolio that complements traditional publication metrics. Such a portfolio would allow investigators to document contributions across scientific discovery, reproducibility, software and data infrastructure, mentorship, collaboration, translation, public engagement, scientific citizenship, and open science. This approach would better reflect the collaborative, interdisciplinary, and mission-driven nature of modern biomedical research while creating incentives that strengthen the scientific enterprise as a whole rather than rewarding only individual productivity.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/RFI-on-Measuring-and-Rewarding-Scientific-Impact-_JPF.docx

Description: Word document version of my responses

Response to NIH Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Rigor and Reproducibility

Scientific rigor should be recognized as a primary research output rather than a compliance requirement. While biomedical research has traditionally rewarded novelty, equally important activities, including replication studies, validation experiments, publication of negative or null findings, benchmarking analyses, and methodological refinement, remain undervalued despite their essential role in building a trustworthy scientific literature.

NIH should develop measurable indicators that recognize contributions to reproducibility, including publication of independent replication studies, release of reproducible computational workflows, transparent reporting of methods and quality-control metrics, reanalysis of publicly available datasets, and development of benchmark datasets or community standards. Well-designed studies that produce negative or null results should be recognized as valuable scientific contributions because they prevent unnecessary duplication, refine biological hypotheses, and improve future experimental design.

To further incentivize rigorous science, NIH should consider dedicated funding mechanisms for replication studies, administrative supplements to validate influential findings, and grant review criteria that explicitly reward reproducibility efforts. Expectations should remain flexible across disciplines and career stages while recognizing that rigorous science is an ongoing process of refining knowledge rather than simply producing novel discoveries.

Data, Software, and Model Sharing

Data, software, computational workflows, and predictive models have become foundational products of modern biomedical research. Yet these resources often receive substantially less recognition than traditional publications despite enabling thousands of downstream discoveries.

NIH should explicitly recognize open-source software, databases, computational workflows, interoperable standards, and publicly accessible datasets as primary scholarly outputs. Indicators of impact should include software adoption, community contributions, dataset reuse, workflow implementation, API usage, long-term maintenance, and compliance with FAIR (Findable, Accessible, Interoperable, Reusable) data principles rather than publication citations alone.

Importantly, NIH should establish dedicated funding mechanisms to support the development, maintenance, and long-term sustainability of open-source scientific software and infrastructure. Many of the most widely used computational tools in biomedical research are developed and maintained by small research groups with

limited long-term support despite serving thousands of investigators worldwide. Sustaining these resources should be recognized as a valuable scientific contribution, not merely technical service.

Training and Mentorship

The future strength of the biomedical research enterprise depends not only on producing excellent science but also on preparing scientists for the increasingly diverse roles they play in society. Modern scientists serve as mentors, educators, communicators, entrepreneurs, software developers, policy advisors, and community partners in addition to conducting research.

NIH should broaden how mentorship and training are evaluated by recognizing contributions that prepare trainees for this evolving landscape. Potential indicators include trainee career outcomes across academia, industry, government, education, policy, entrepreneurship, nonprofit organizations, and clinical practice; evidence-based mentoring practices; trainee independence; structured mentorship programs; and mentoring climate.

Training programs should also incentivize skills that are increasingly essential to modern science but are often treated as optional. These include science communication, science policy, community engagement, entrepreneurship, open science practices, data stewardship, peer review, and interdisciplinary collaboration. Investigators who train students in transparent scientific practices, including preprinting, reproducible computational workflows, responsible data sharing, accessible communication, and constructive peer review, should receive formal recognition for these contributions.

Scientific excellence should include preparing the next generation of scientists not only to conduct outstanding research but also to effectively communicate, translate, and steward that research for the benefit of society.

Collaboration

Biomedical research increasingly relies on interdisciplinary teams that span institutions, sectors, and scientific disciplines. However, many evaluation systems continue to emphasize individual accomplishments rather than collaborative contributions.

NIH should recognize meaningful contributions to team science regardless of publication authorship position by considering leadership of collaborative projects, development of shared infrastructure, multi-investigator grants, co-mentorship, community resource development, and documented contributions through narrative team science statements or contributor role taxonomies.

In addition, NIH should expand funding mechanisms that intentionally build scientific capacity through collaborations with Minority Serving Institutions (MSIs), Historically

Black Colleges and Universities (HBCUs), Tribal Colleges and Universities (TCUs), community colleges, primarily undergraduate institutions, and emerging research institutions. These partnerships should be structured as reciprocal scientific collaborations that build long-term research capacity rather than short-term subcontracting arrangements.

Such investments would not only strengthen scientific discovery but also broaden participation in biomedical research while expanding research capacity throughout the United States.

Entrepreneurship and Translation

Translation should be viewed as a bidirectional process rather than a one-way pathway from bench to bedside. Basic scientists should have greater opportunities to move discoveries toward clinical implementation, while clinicians, patients, engineers, public health practitioners, policymakers, and industry partners should have meaningful opportunities to inform fundamental biological discovery.

NIH should recognize diverse forms of translation, including clinical implementation, public health impact, policy translation, development of diagnostics and research tools, open-source technologies, nonprofit partnerships, educational innovations, and commercialization where appropriate.

Expanding funding mechanisms that intentionally connect basic science laboratories with clinicians, public health practitioners, patient organizations, entrepreneurs, and industry partners would strengthen the entire translational ecosystem. These partnerships should be viewed as complementary approaches for maximizing the public benefit of biomedical research rather than competing models of scientific impact.

Foundational Scientific Exploration

Many of the most transformative advances in biomedical research emerge from high-risk, high-reward investigations that require sustained investment and may not produce immediate publications or translational outcomes. Current incentive structures often favor predictable projects with short-term deliverables over ambitious questions capable of fundamentally changing scientific understanding.

NIH should continue expanding support for foundational scientific exploration by recognizing long-term infrastructure development, technology innovation, biological resource generation, atlas-building efforts, methodological advances, and exploratory research that creates new scientific directions.

Review criteria should emphasize scientific significance, originality, and long-term potential rather than short-term publication metrics alone. Investigators should not be penalized for pursuing ambitious projects whose greatest impacts may emerge years after the initial funding period.

Public Impact

The ultimate purpose of biomedical research is to improve human health and benefit society. Accordingly, public impact should become a core dimension of scientific excellence.

Public impact should extend beyond counting outreach activities or media appearances. Instead, NIH should recognize sustained, reciprocal relationships between scientists and the communities they serve.

Potential indicators include long-term partnerships with patient advocacy organizations, community advisory boards, collaborations with K-12 schools and community colleges, publicly accessible educational resources, contributions to evidence-informed policymaking, citizen science initiatives, implementation of research findings into clinical practice or public health, and measurable efforts to strengthen public trust in science.

NIH should also recognize contributions that improve the accessibility and transparency of science itself, including open educational resources, accessible software, publicly available datasets, preprints, transparent communication of findings, and efforts to make biomedical knowledge broadly understandable and reusable.

The goal should not simply be increasing public awareness of science but strengthening enduring relationships between scientific institutions and the communities they exist to serve.

Other Areas

One area deserving additional consideration is modernization of scientific publishing itself. The current publication system remains one of the strongest drivers of scientific incentives, yet many activities essential to scientific quality remain largely invisible during evaluation.

NIH has an opportunity to encourage a transition toward more open, transparent, and efficient models of scientific communication by explicitly recognizing preprints, open peer review, and emerging Publish, Review, Curate (PRC) models. These approaches separate dissemination, evaluation, and curation into distinct scholarly activities, allowing discoveries to become immediately accessible while expert review and long-term curation continue transparently.

NIH should consider allowing investigators to document preprints, citable peer reviews, open review activities, and community curation efforts as evidence of scholarly impact. Likewise, reviewer contributions, editorial leadership, and development of community standards should be recognized as meaningful scientific outputs rather than invisible service activities.

Finally, NIH should consider adopting a broader Scientific Impact Portfolio that complements traditional publication metrics. Such a portfolio would allow investigators to document contributions across scientific discovery, reproducibility, software and data infrastructure, mentorship, collaboration, translation, public engagement, scientific citizenship, and open science. This approach would better reflect the collaborative, interdisciplinary, and mission-driven nature of modern biomedical research while creating incentives that strengthen the scientific enterprise as a whole rather than rewarding only individual productivity.

Sincerely,

JP Flores, PhD

96. Peter Deak

I am responding: On behalf of myself

Name: Peter Deak

Name of Organization: Drexel University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/16/2026

1. Rigor and Reproducibility: The requires a culture change that unfortunately is an uphill battle. From my experience, "rigor" is not well defined as a concept and as a result instead gets wielded as a weapon to tank proposals rather than as a guide for reproducible research. For a reviewer on an NIH panel, rigor can mean sticking to analysis conventions in the field or it can mean thoughtfully considering alternative hypotheses. I think there needs to be a review of what "rigor" actually means for scientists and a way to separate superficial methods of analysis or data presentation from substantive critiques of lack of consideration of alternative hypotheses or flaws in experimental design. Then we can move onto reproducibility.

For the rigor problem, this should involve a core definition of "rigor" and what this means to science. They, very importantly, this requires active monitoring from NIH staff when critiques of rigor are inappropriately applied.

Next for reproducibility, this first requires a standardization of methods reporting in publications. Next, financial incentives for other groups to reproduce or refute findings from a previous study. This would be as simple as an RFA that specifically requests reproducibility studies or a dedicated NIH internal laboratory.

2. Data, Software, and Model Sharing: This is easy as well, just make complete accessibility of all datasets a requirement of all NIH grants. Start canceling grants that do not comply. To be honest, I think this has not been happening not due to fraud but because it takes work.

3. Training and Mentorship: I would contend that T-training programs are not a good use of money. They are commonly given to well-funded institutions and are sort of an afterthought for most investigators, rather a conduit to fund a student instead of a mechanism for training. Most training occurs in the 1:1 relationship between mentor/mentee. So, start rewarding good individual mentors. Look for track records of successful students, relative to the funding a mentor has received. Surveying trainees are also a critical piece in here. Beyond the scope of the NIH, but we need to, as a scientific community, start rewarding effective mentorship and shaming poor mentorship.

4. Collaboration: I think this is already relatively well done, especially if the scientists lead the endeavors. I think the only minor concern is the "silent" collaborator, one who is an important name but doesn't

actually contribute to the research. If this is seen in yearly progress reports, I think removing them from the grant is a good option.

5. Entrepreneurship and Translation: I think this is honestly outside the NIH purview. I think the NIH needs more programs collaborating with industry to move translation.

6. Foundational Scientific Exploration: This is probably the area with the most room for growth and is fundamentally a problem of the broken peer-review system, which inherently prioritizes middle-of-the-road, moderate impact science. This is due to the 3 reviewer paradigm, where if one reviewer doesn't understand or believe in a high-risk idea, the project is doomed, even if the other two are enthusiastic. I have seen this first-hand dozens of time, both in my grants and others, where a flimsy "rigor" critique bleeds into the factor 1 score and ruins the grant. This is where the SROs need to have more autonomy to push back on reviews that mis-categorize their review critiques. Additionally, there is also a problem with well-funded institutes getting a passive benefit of the doubt and an inherently lower bar for rigor due to the reputation of the institution or investigator. This is a tricky problem, as it is more likely that a well established PI will have access to more resources and will get some productive science out of it. So in a sense, this bias is justified, but also leads to the same institutions with the same mindsets getting the majority of the funding, which can lead to a lack of new ideas and perspectives.

My suggestions on how to fix:
1. Make an initial letter of intent round to only focus on the innovative idea rather than "rigor".
2. Make this initial stage double blind to remove bias, only focus on the ideas
3. Allow applicants to review the reviewers on their review to get more feedback and train or drop poor performing reviewers. This needs to be done very carefully and considered in context with reviewer scores of applicants.
4. If possible, allow for "outlier" scores to be dropped
5. More high-risk/high reward funding proposals. We also need to drop the idea that all funding proposals requires (either implicitly or explicitly) preliminary data. It is an open secret that R21 actually require some preliminary data in order to be funded.

7. Public Impact: This is a hard one too. This is inherently against the headwinds of our poor political climate, which public support for biomedical research. I think framing the NIH as responsible for tax dollars and being open about how the money is spent is important. I think public access to all financial data and I would argue even peer review records would be helpful. If scientists want public funding, all their research efforts are then open to public scrutiny. I personally think that public shaming of bad actors in the biomedical research world is an underused strategy that could improve public trust.

8. Other Areas: I hate the funding application limit of 6 per year. This needs to be updated to factor in number of proposals currently supported. If a PI (like me) is set to completely run out of money, there needs to be a consideration of that and an increase in the limit.

Description: Consequences of poor definitation of rigor, institutional biases are driving poor peer review results leading to some stagnation and lack of public trust

97. Chunyu Liu

I am responding: On behalf of myself

Name: Chunyu Liu

Name of Organization: SUNY Upstate Medical University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/17/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

98. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/17/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: NIH theoretically has the ability to mine applications with the purpose of identifying common research resources and then help connect applicants to these resources relevant to their project. Applicants already doing this could be incentivized by RFAs that specifically call for projects that would facilitate data, software, and model sharing.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

99. Hossein Ravanbakhsh

I am responding: On behalf of myself

Name: Hossein Ravanbakhsh

Name of Organization: The University of Akron

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/17/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: As an Early Stage Investigator, I appreciate NIH's efforts to improve application quality and review efficiency. However, a fixed limit on the number of proposals per year (6 applications annually) may have unintended consequences for ESIs. Early investigators often require multiple funding attempts while establishing independent research programs, generating preliminary data, and developing grantsmanship skills. Unlike established investigators, ESIs generally have fewer alternative funding sources and smaller research infrastructures. Restricting the number of applications may therefore reduce innovation and slow the development of the next generation of NIH-funded investigators. I encourage NIH to consider exempting ESIs from the application cap or adopting a higher limit specifically for ESIs. Alternatively, NIH could exempt ESI-specific funding mechanisms (e.g., DP2, R21 Trailblazer, MIRA ESI, and R16 SuRE-First) from the proposal cap, as these programs are specifically intended to support early-career investigators.

100. David Rocha

I am responding: On behalf of myself

Name: David Rocha

Name of Organization:

Type of Organization: Not Applicable

Role: Member of the Public

Date Submitted: 7/17/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: All NIH institutes and centers need to converge into a computable NIH ecosystem so that terminology, knowledge, services, and information is machine readable and executable across all organizational boundaries.

The NLM's regional medical libraries need to move to FHIR APIs at scale so that the NLM's regional network can provide education, assistance, and utilize terminology services which would allow healthcare organizations to adopt rxnorm, LOINC, SNOMED CT, and FHIR at scale, and other interoperability standards (including the NIH related standards such as GA4GH, OMOP, HPO, Orphanet) at scale. This will scale EBM on FHIR, Vulcan, between the academic, research, and clinical communities.

101. Neehar Kundurti

I am responding: On behalf of myself

Name: Neehar Kundurti

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/17/2026

1. Rigor and Reproducibility: AI use disclosure: Apple Intelligence "Writing Tools" used to proofread the document's word and character count and tighten the prose.

Across the domains, a single theme recurs: as artificial intelligence (AI) tools enter both the conduct of research and the training of researchers, the indicators NIH uses to measure success should reward verifiable, well-documented, reproducible contributions and the human competencies that produce them, rather than raw output. The recommendations below emphasize measurable indicators, implementation feasibility, and the unintended consequences worth anticipating.

Reproducibility is foundational to rigor, and AI tools cut both ways. Used well, they can make analyses more reproducible through their scripts, documented pipelines, and structured artifacts that can be re-run. Used uncritically, they introduce new and distinctive failure modes that undermine reproducibility with fabricated citations, hallucinated findings presented fluently, silent dependence on training data that may not match the study population, and outputs that vary with imperceptible changes in prompting. New incentives and infrastructure for replication should therefore explicitly account for AI-assisted work.

Recommended measurable indicators:
1. Documented verification of AI-assisted outputs
Where AI tools contribute to analysis, code, or synthesis, reward documented evidence that the output was independently verified i.e. citations checked against primary sources, analyses independently reproduced, generated code validated against known inputs. Absence of this is itself informative. Something as benign as Apple Intelligence's "Writing Tools" used to reshape a manuscript to bring it within a word limit needs to be verified and disclosed as being done as such. The discipline of verification is a reproducibility skill one must hone in an AI-assisted workflow.
2. Reproducibility of the artifact, not just the result
Research utilizing the artifacts must disclose the artifact unless documentation of proprietary use. NIH must favor research products that are sent with the means to reproduce them including the data, code, environment, and a clear account of any AI assistance. A reproducible analysis script tied to its data and a figure traceable to its source values are assessable, re-runnable units of rigor.
3. Replication and validation studies as rewarded outputs
Replication and validation are currently under-incentivized relative to novel findings. Recognizing them

as fundable, creditable contributions (including validation of AI-based tools and models across populations and settings) addresses the distribution-shift failure mode that makes many AI findings non-reproducible across sites.
4. Transparent AI-use disclosure tied to the record
Disclosure of which tool was used, by whom, when, for what purpose, and how the output was verified should travel with the research product. This is a low-cost indicator that supports auditing and reproduction and aligns with existing publisher and funder norms.

Tool evolution lag
There is a structural mismatch between the pace of AI development and the timelines of biomedical publishing and policy. The interval between submission and publication of a paper, with the longer interval over which a funding policy stays in force, is measured in months to years, while the capabilities, failure modes, and even the identities of widely used AI tools change on a far shorter cycle i.e days to weeks. A specific tool named in a paper, a specific model behavior described in a methods section, or a specific prohibition written into a policy can be out of date before the work reaches its audience. This has two consequences for how NIH should frame reproducibility indicators:
1. Appraise the verification process, not the named tool
Because the tools change faster than the literature that describes them, durable indicators should reward the generalizable competency i.e. did the researcher critically appraise and verify the AI output, document the basis for trusting it, and check whether it transfers to their setting, rather than endorsing or prohibiting particular named products that will be superseded. The appraisal skills are stable across tool generations but the tools are not.
2. Specify behavior and intent, not brand
Disclosure and policy language should describe the function and the verification expected (“a generative model was used to draft X and all output was verified against primary sources”) rather than freezing a brand name or version into a permanent record, so that the requirement remains meaningful as the underlying technology turns over.
3. Build in periodic re-evaluation
Any standard that touches AI should carry an explicit re-evaluation cadence (a 6-to-12-month cycle is a reasonable starting point, but as tools evolve, that may need to occur on a much shorter timeline) so that indicators and guidance are refreshed as tools evolve, rather than ossifying around a single moment in the technology’s development.

Implementation Feasibility
Verification records, disclosure statements, and reproducibility packages (data plus code plus environment) reuse practices that rigorous labs already follow; the change is making them an explicit, rewarded part of how success is measured. Re-evaluation cadences are administratively light (a scheduled review rather than new infrastructure). The heaviest lift is cultural: shifting credit from volume of output toward verifiable, reproducible, well-documented contribution.

Unintended consequences to anticipate
1. Verification documentation can become a box-checking ritual; spot-checking a sample of verification records, rather than trusting attestation alone, preserves its meaning.
2. Over-specifying reproducibility packaging may disadvantage fields where data cannot be shared (for

example, work involving protected health information); requirements should be calibrated to what each discipline can feasibly and ethically share.
3. A prohibition-based stance toward specific AI tools will age poorly and may push use underground; a verification- and disclosure-based stance is more durable and more honest about how research is now done.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: NIH notes that advancing its mission requires a stable, growing biomedical workforce and that additional indicators are needed to assess the impact of training and mentorship. We suggest that, in the current moment, a defensible indicator set must account for how AI tools are reshaping what trainees learn to do and what they may fail to learn. The risk is not hypothetical: when AI assistance is available for literature synthesis, drafting, coding, and analysis, trainees can produce competent-looking output without developing the underlying competence. We have found it useful to distinguish de-skilling (erosion of an established skill in an experienced researcher) from never-skilling (failure to develop a foundational skill in the first place, because a tool supplied the answer before the trainee built the capacity). Training metrics that count only outputs are blind to this distinction.

Recommended measurable indicators
1. Competency attainment, not output counts
Track demonstrated research competencies attained per trainee against a defined framework (for example, study design, data management, statistical reasoning, responsible conduct, and critical appraisal of AI-assisted outputs), assessed through mentored observation rather than publication or citation tallies. Competency frameworks already exist or are emerging across disciplines and can be adapted; the measurable unit is the verified competency, not the paper.
2. AI-literacy as an explicit, assessed competency
Add critical appraisal of AI tools such as recognition of fabricated citations, hallucinated findings, training-data drift, and distribution shift, to required training, and assess it directly. A trainee who can identify why an AI-generated reference is fabricated, or when a model's performance will not transfer to a new population, is demonstrating a reproducibility-relevant skill that should be recognized.
3. Verification discipline as a trackable behavior
Where AI assistance is used in trainee work, the verification step (independent reproduction of an analysis, source-checking of every AI-suggested citation, validation of AI-generated code against known inputs) is itself the teachable, assessable competency. Programs can record whether verification was performed and documented, giving NIH a process indicator that is feasible to collect and directly tied to research integrity.
4. Mentorship effort made visible
Protected mentorship time is supported by existing policy but is under-measured. A simple, feasible indicator is documented mentored-effort hours per trainee (the actual effort in hours a mentor devotes to a trainee's development) captured through the same effort-reporting infrastructure institutions already maintain. This converts an invisible contribution into a quantity NIH can assess, and it rewards mentors whose value does not show up in their own publication record.

Implementation and feasibility
Effort-based and competency-based indicators are feasible because they reuse infrastructure institutions

already operate: effort certification systems, milestone and evaluation records, and individual development plans. The marginal cost is structuring these records so that mentored effort and competency attainment can be summarized, not inventing a new reporting system. Structured, auditable records of mentored effort and competency milestones can be maintained in standard project-tracking tools and exported for reporting; the discipline is consistency of entry, not new technology.

Unintended consequences to anticipate
1. Any competency checklist can be gamed if attainment is self-attested without verification; mentored observation or work-product review should anchor the assessment.
2. Counting mentored hours could incentivize quantity of contact over quality; pairing the hours indicator with trainee-outcome and competency measures mitigates this.
3. Mandating AI-literacy training without supplying current teaching material risks the training aging faster than it can be delivered (see Tool Evolution Lag in Rigor and Reproducibility)

4. Collaboration: NIH seeks new approaches for assessing and recognizing contributions to team-based science as research grows more interdisciplinary. The dominant current metrics of authorship position and citation counts capture team contribution poorly and are especially ill-suited to the kinds of contribution that make team science reproducible i.e. data stewardship, methods and software development, operational coordination, and the verification work that lets one team build on another's results. AI tools sharpen this problem, because a contributor can now generate a large volume of text or code quickly, and output volume is an even weaker proxy for contribution than it was before.

Recommended measurable indicators
1. Structured contribution taxonomies
Adopt and reward use of a structured contributor-role taxonomy (such as CRediT) so that data curation, software, methodology, validation, and project administration are recognized alongside writing. The measurable indicator is the presence and granularity of contribution statements, which NIH can request and assess. This directly surfaces the reproducibility-supporting roles that authorship order hides.
2. Recognition for shared, reusable research infrastructure
Credit the creation of shared resources the whole team and field reuse: documented datasets, validated analysis pipelines, reusable templates and schemas for data collection, and operational tooling. A practical indicator is documented adoption of a shared resource by other projects or teams, which rewards the contributor who built the thing everyone depends on rather than only the contributor who published first.
3. Verifiable, attributable artifacts over prose volume
As AI lowers the cost of producing prose, contribution is better assessed through discrete verifiable artifacts such as a validated dataset, a reproducible analysis script, a documented data-collection schema, a figure traceable to its source data, each attributable to a named contributor and independently checkable. These artifacts are more meaningful units of team contribution than word count and are inherently reproducibility-friendly because they can be re-run and re-checked.
4. Disclosure of AI assistance as a team norm
Recognize teams that maintain clear, consistent disclosure of where and how AI tools contributed to the work. Transparent disclosure is a marker of a well-governed collaboration and supports downstream reproducibility; its presence and consistency are assessable.

Implementation feasibility
CRediT and similar taxonomies are already in use at many journals and impose minimal burden. Crediting reusable infrastructure requires a lightweight way to register and cite resources (persistent identifiers for datasets, code, and protocols already exist). Attributing verifiable artifacts to contributors is feasible wherever version control or structured project records are kept, which is increasingly standard.

Unintended consequences to anticipate
1. Contribution taxonomies can be inflated (everyone credited for everything); periodic audit of a sample of contribution statements keeps them meaningful.
2. Rewarding reusable infrastructure could bias effort toward tool-building over discovery if over-weighted; it should complement, not replace, scientific-contribution measures.
3. Artifact-based credit favors fields and teams with mature data and code practices; NIH should calibrate expectations across disciplines and institution types so that less-resourced groups are not penalized.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

102. Sean Whelan

I am responding: On behalf of myself

Name: Sean Whelan

Name of Organization: Washington University in St.Louis

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/18/2026

1. Rigor and Reproducibility: Use of NIH $ to support replication of studies is not a valid use of taxpayer funds. The existing system already ensures rigorous and reproducible science is that which is advanced. If something isn't replicable it isn't pursued.

2. Data, Software, and Model Sharing: Excellent. NIH has already invested in data sharing policies. Funds should be included in grants to ensure compliance with FAIR standards.

3. Training and Mentorship: Short term metrics here are largely meaningless. The impact of a mentor is measured over many years if not actually decades and not subject to the vagaries or whims of any short term assessment
.

4. Collaboration: Abandon your plans to penalize investigators who have many shared grants. This is what your misguided notion of capping grants will do. Incentivize collaborative grants.

5. Entrepreneurship and Translation: This is a goal of all science. It is naive to think that one can predict what basic science will ultimately be translational. For example…. studying phage defense mechanisms…(not supported by NIH). Many examples of work funded that was not immediately apparently translational that ultimately was (reverse genetics of vsv being an example)

6. Foundational Scientific Exploration: Absolutely. This is why peer review with a lack of political interference is so critical. It's also why not wasting $ funding replication work is also important. What works gets picked up, and followed.

7. Public Impact: Remove political oversight. And invest in science communication regardless of whether that is politically uncomfortable. Science is the pursuit of truth.

8. Other Areas: N/A

103. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Research Trainee/Fellow

Date Submitted: 7/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Response_NOT_OD_26_087.pdf

Description: Response to NOT_OD_26_087

Responding to NOT-OD-26-087

Dear NIH Office of Science Policy,

With respect to neurobiology laboratories, specific, measurable indicators and incentives that reflect the collaborative, rigorous, and mission-driven laboratories in modern biomedical science should likely reflect the small N and iterative experimentation and exploration that results from this type of research. Metrics like pre-registration are not typically performed for this type of research, so creating a multi-point metric that measures the level of technical rigor should look differently for neurobiology labs compared to labs that use large datasets of human data.

However, technical rigor is not the same as scientific rigor. And a publication can be extremely technically rigorous, with every statistic listed and every virus catalogued, but the paper could be biased on where or how the experiment was focused and be missing the bigger picture of the biology. So how do we measure scientific judgement?

Here are a handful of things to think about:

1. Alternative hypotheses Testing Index

Was more than one plausible competing hypothesis experimentally addressed? Here are examples: were multiple brain regions shown after tract tracing and did the author give a reason grounded in the evidence and/or literature for following up on the one chosen? Were neighboring or connected structures tested? Were collateral projections tested? Were multiple cell types tested or surveyed? Off-target behavioral pharmacology ruled out? And especially for pain research, when the sensory system is tested, were motivational-affective versus motor effects isolated?

2. Biological Fidelity Index

The NIH has traditionally emphasized hypothesis-driven work, but major advances have come from exploratory or discovery-driven research. Informative datasets like spatial transcriptomic projection mapping, longitudinal calcium imaging across disease progression, along with mechanism testing would make a robust biological fidelity index. Other elements could be is anatomical axis (e.g. rostrocaudal axis) sampled rather than a single coronal section? An example of why this is important is the hyperfocus in the posterior ventrolateral periaqueductal gray for pain when that brain region is enormous and has a large rostrocaudal axis. Is sufficient temporal resolution considered to detect state transitions? Are multiple cell types or circuit components considered when the biology is known to be heterogenous? The NIH already requires inclusion of sex as a biological variable, so this extension to incorporate more biological fidelity follows in that spirit.

3. Mechanistic Depth and Multiple Levels of Evidence

This can be a metric that quantifies the level of evidence accumulated over a career. Did a scientist incorporate multiple neuroanatomical regions, multiple techniques measuring physiology, measuring necessity and sufficiency, temporal dynamics, upstream effectors and downstream targets, multiple types of behaviors specificity? And what is the breadth of this within a paper or within a timespan?

4. Literature Integration Score

What I find frustrating is that most neuroscientists I meet are obsessed with keeping up with the Joneses with respect to reading and then pursuing projects that are trendy, and all the work coming out seems to be different flavors of the same thing. Suddenly, everyone is using t-SNE plots on everything, even micro-behaviors, when a plot with pretty colors and a seeming complexity doesn't really tell you anything new about what the animal is doing. What I see more rarely are scientists that are comfortable digging up older lines of research that were abandoned and using modern tools to explore the finding. In addition, leaps of logic and baseless definitions are sometimes allowed in publications and caution should be used when interpreting the results when the literature does not support the basis for the study. To me, literature from 40 or 50 years ago was often on the cusp of finding important biology, but we were limited by the tools available. There are so many good studies from literature of the 1970s through 1990s that were simply set aside because of a lack of tools. Novel ideas are great and I applaud them, but if you dig down, those novel ideas are often based on findings from researchers of previous generations. I would like to see more competing lines of study followed through using modern techniques.

5. Effort toward falsifying the hypothesis

Outstanding science often comes from time spent trying to disprove the central hypothesis, yet this effort is invisible with today's metrics. What experiments were performed specifically because they could have disproven the central hypothesis? Attempts at refutation are central to strong scientific effort.

6. Conceptual Longevity

At two, 5, 10, 15 and 20 years later, are scientists still building on the framework produced? This cannot evaluate early work but should be considered for retrospective assessments.

7. Uncertainty Reduction

Instead of productivity in terms of paper count, what could be scored is whether competing hypotheses were tested, whether assumptions were explicitly stated, whether negative results were reported that strengthened or reshaped hypotheses, whether conclusions matched evidence, whether limitations were discussed quantitatively, and whether other labs were able to identify remaining uncertainty extending the findings.

Given these lines of thinking, here is a putative Scientific Scorecard that combines technical and scientific rigor:

Metric	Example Target
Publications with complete methods	>95%
Statistics cataloged with effect sizes, analysis code and datasets made available for replication	>80%
Independent replication within laboratory (e.g. independent cohorts, different experimenters, etc)	>75% of major findings
Blinded or automated behavioral scoring	>90% where applicable

Histological verification and exclusion criteria reported	100%
RRIDs and reagents reported	100%
Alternative hypotheses experimentally tested	≥2 major competing explanations addressed per principal conclusion, where appropriate
Scientific completeness (mechanistic depth)	Major studies integrate multiple complementary lines of evidence (e.g., anatomy, physiology, behavior, causal manipulation, temporal dynamics)
Biological fidelity of experimental models	Experimental design reflects clinically or biologically relevant spatial, temporal, and physiological scales (e.g., chronic disease duration, anatomical organization, longitudinal measurements)
Assumption testing	Regular publication of studies that explicitly evaluate widely accepted methodological or conceptual assumptions in the field
Shared resources adopted externally	Increasing over time
Trainee mentoring evaluations	Above NIH benchmark
External collaborations	Multiple active projects

On the topic of training and mentorship, why does the NIH not have exit questionnaires for the NIH supported trainees with respect to all mentors encountered at the respective institution? I feel especially if the trainee left early, you should ask. I understand that these should be considered in the light of a stressful training environment, but I genuinely wonder why the NIH does not simply ask the trainees for feedback? One negative feedback should not weigh heavily on a faculty member, but a pattern may arise over time. The categories could be varied, such as how much do you feel this mentor advanced your scientific rigor, your ability to identify proper experimental controls, your grasp on the scientific literature? These kinds of letters of support are always a tricky thing in science, but if it goes one direction (the trainee must get letters of recommendation), then why not in the other (the mentor must get letters of mentorship evaluation). The trainees are trying to become part of the scientific workforce. Why not ask them about their experiences? You could also create a scorecard for institutions based on trainee exist evaluations that are delivered not to their institutional email addresses but their personal email addresses.

On the topic of entrepreneurship and translation – personally I am a behavioral neuroscientist and I want to warn against too much percent weight of a score toward commercialization and entrepreneurship. My goal as a scientist is to share data and build the scientific knowledge base. If you want me to commercialize what I produce, this goes against the sharing spirit. I understand for tools this is different but I do think that documenting and sharing a tool can equally achieve integration of scientific knowledge into health advances. What I see is a push for professors to start commercial businesses and develop tools that can be patented. This to me goes against the spirit of sharing knowledge.

On the topic of public impact, one of the major elements I want to point out is a recent push from the public against biomedical research using animal models. It is of course my duty as a scientist to reduce and replace the use of animal models whenever necessary, and to always

consider the health and welfare of the animal when designing experiments. However, there seems to be a lack of understanding that in order to address human suffering and loss of quality of life, we must address the biopsychological aspects of behavior. To do this, we use animal models because this cannot be done in a chip, and mechanistic biological findings cannot be achieved through neural imaging. We have made massive progress toward understanding how disease works through use of animal models. I want to reiterate that many scientists have a deep reverence for the models they use and do not take the matter lightly. But in order to make progress in how the brain gets stuck in chronic pain and why depression and cognitive deficits correlate with chronic pain we need to use animal models. And I think the NIH itself can help the public understand that the funding dollars going toward research using animal models has made a real impact on healthcare knowledge. This research is difficult to perform. The neurobiologists spend years learning how to design and perform experiments to study disease. We regularly discuss ethics and alternatives around our experiments. But we genuinely want to help solve healthcare problems and that's why we do what we do. I think the NIH could help publicize this sentiment to help build public trust and understanding of the societal benefit that is being funded.

Thank you for your time.

104. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/20/2026

1. Rigor and Reproducibility: Reward a lab's replication track record, pre-registration, reporting-standard and reagent-authentication compliance; fund replication directly. Guard: credit well-powered attempts regardless of outcome; don't penalize genuine exploration.

2. Data, Software, and Model Sharing: Reward verified FAIR compliance and downstream reuse (dataset/software citations, dependent works), not mere deposit. Guard: weight reuse over raw volume; fund curation for small labs; controlled-access credit for sensitive data.

3. Training and Mentorship: Reward longitudinal trainee outcomes, independence, and mentees' own rigor. Guard: value diverse (non-academic) paths; normalize for placement rates; centralize tracking so small/MSIs aren't penalized.

4. Collaboration: Reward specific roles via the CRediT taxonomy and enabling "invisible" work. Guard: make credit role-based, not authorship-count-based; don't penalize sole-investigator fields.

5. Entrepreneurship and Translation: Reward downstream translation (patents, licenses, INDs, guideline adoption) on realistic time lags, plus credit for enabling others' translation. Guard: keep as one axis; don't starve basic science or judge ECRs on products.

6. Foundational Scientific Exploration: Reward productive failure (rigorous null results), long-horizon impact, and assess high-risk work at portfolio level. Guard: de-weight productivity counts here — metrics can kill the risk-taking they measure.

7. Public Impact: Reward contribution to health outcomes, reach/equity, and transparency, via peer-assessed impact narratives (REF-style). Guard: require evidence; sample rather than universally mandate; support smaller institutions.

8. Other Areas: Replace single scores with a role-appropriate dashboard — a small basket of mission-aligned indicators matched to role and career stage, not the h-index or JIF. Design against Goodhart's law — pair hard-to-game quantitative indicators with peer-assessed narrative; audit; retire captured metrics. Reward reliability directly — fund and credit replication, null results, data/software reuse, and mentorship; this is where the reproducibility dividend is recovered. Differentiate by career stage, discipline, and institution type — one uniform standard predictably disadvantages ECRs, basic scientists, and smaller/minority-serving institutions. Pilot, evaluate, and sunset — treat each new metric as a hypothesis; measure whether it changes behavior and improves reliability; retire it if it backfires.

105. Fabio Renato Manzolli Leite

I am responding: On behalf of myself

Name: Fabio Renato Manzolli Leite

Name of Organization: University of Utah School of Dentistry

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/20/2026

1. Rigor and Reproducibility: Current incentives in biomedical research continue to reward scientific productivity more consistently than scientific reliability. Publications, citations, and novelty remain the dominant measures of success, while the activities that increase confidence in scientific knowledge—replication, methodological validation, transparent reporting, and rigorous negative findings—receive comparatively little recognition. NIH has a unique opportunity to change this by rewarding activities that strengthen confidence in biomedical evidence, not only those that generate new discoveries.
Rigor should be evaluated through a balanced portfolio of indicators rather than a single metric. These include transparent study design and reporting, appropriate protocol registration, independent replication or validation, availability of protocols and analytical code, adherence to discipline-specific reporting standards, and evidence that research outputs can be independently verified or meaningfully reused. NIH should also recognize contributions that improve the reliability of an entire field, including standardized methods, validated outcome measures, reporting guidelines, and multicenter harmonization initiatives.
Implementation should emphasize scientific principles rather than compliance checklists. Expectations should be appropriate to the research question, study design, discipline, career stage, and available resources. Likewise, causal inference should be judged by whether the chosen methodology appropriately addresses the scientific question rather than by a rigid hierarchy of study designs.
Ultimately, NIH should reward activities that increase confidence in scientific knowledge. Doing so would shift the research culture from maximizing scientific output to maximizing scientific reliability—the foundation on which meaningful discoveries and improvements in human health ultimately depend.

2. Data, Software, and Model Sharing: The value of data, software, and model sharing should be measured by their scientific reuse rather than by repository deposition alone. Making resources publicly available is an important first step, but it does not necessarily increase scientific impact if they cannot be readily understood, reproduced, or integrated into future research. NIH should reward the creation of reusable scientific assets rather than compliance with data-sharing requirements.
Evaluation should emphasize scientific utility through indicators such as adherence to FAIR principles, high-quality metadata and documentation, interoperable standards, reproducible analytical workflows, and evidence of independent reuse or validation. Development and long-term stewardship of widely adopted datasets, software, protocols, and other shared resources should be recognized as meaningful scientific contributions because they enable discoveries far beyond the original research project.
Implementation should remain flexible across disciplines and research settings. Expectations should

reflect differences in data sensitivity, participant privacy, intellectual property, and the types of resources generated by different fields. Controlled-access mechanisms should be supported when appropriate to balance scientific openness with ethical and regulatory responsibilities.
By rewarding scientific reuse rather than data sharing alone, NIH can maximize the value of public investment, strengthen collaboration, reduce unnecessary duplication of effort, and accelerate reliable biomedical discovery.

3. Training and Mentorship: Training and mentorship should be recognized as core scientific contributions because they determine the future strength of the biomedical research enterprise. Current evaluation systems often emphasize the number of trainees or training grants, yet these measures say little about the quality or long-term impact of mentorship. NIH should reward the development of people rather than simply the supervision of trainees.
Evaluation should focus on meaningful outcomes, including trainee career progression, scientific independence, multidisciplinary engagement, responsible research practices, leadership development, and preparation for diverse careers across academia, healthcare, industry, government, and other sectors of the biomedical workforce. Contributions that strengthen the broader research community—such as mentoring early-career investigators, fostering interdisciplinary collaboration, developing shared educational resources, and promoting international scientific exchange—should also be recognized as valuable scientific contributions.
Implementation should recognize that effective mentorship occurs under very different institutional and disciplinary contexts. Expectations should therefore reflect career stage, institutional mission, available resources, and the diversity of scientific career pathways rather than applying uniform benchmarks across the biomedical enterprise.
By rewarding meaningful mentorship and workforce development, NIH can strengthen the sustainability, diversity of perspectives, and long-term excellence of biomedical research while ensuring that future generations of scientists are prepared to address increasingly complex health challenges.

4. Collaboration: Modern biomedical research increasingly depends on interdisciplinary teams, yet research evaluation continues to rely heavily on traditional indicators such as principal investigator status and first or senior authorship. These measures often fail to recognize the diverse expertise required to produce rigorous, innovative, and translational science. NIH has an opportunity to shift from rewarding position within a project to recognizing meaningful scientific contribution.
Evaluation should emphasize how collaboration increases scientific value rather than simply counting collaborators or publications. Potential indicators include leadership of multidisciplinary initiatives, development of shared infrastructure or methodologies, successful coordination of multicenter studies, data stewardship, and sustained participation in collaborative networks that produce reproducible and impactful science. Standardized contribution statements, such as the CRediT taxonomy, should complement traditional metrics to better recognize individual scientific roles.
Implementation should encourage collaborations that bring together complementary expertise, technologies, patient populations, and perspectives that would otherwise be unavailable. International collaborations deserve particular consideration because they often provide unique scientific capabilities, longitudinal cohorts, specialized technologies, or access to diverse populations that cannot be replicated within a single institution or country. While appropriate oversight is essential, administrative and financial requirements should remain proportional so they do not unnecessarily discourage scientifically valuable partnerships.

By rewarding meaningful scientific contributions rather than authorship position or collaboration volume alone, NIH can foster a research culture that better reflects the realities of modern team science while accelerating innovation, reproducibility, and improvements in human health.

5. Entrepreneurship and Translation: Entrepreneurship and translation should be evaluated by their contribution to improving human health rather than by commercial success alone. While patents, licenses, startup companies, and industry partnerships are important pathways for translating biomedical discoveries, many of the greatest public health advances improve lives through changes in clinical practice, prevention, healthcare delivery, implementation, education, or widely adopted research tools. NIH should recognize this diversity of translational pathways.

Evaluation should emphasize measurable progress toward real-world impact, including adoption into clinical guidelines or healthcare systems, implementation of evidence-based interventions, development of validated diagnostics, software, or decision-support tools, regulatory milestones, commercialization when appropriate, and documented improvements in patient care or population health. Expectations should reflect the discipline and the pathway through which research delivers public benefit rather than applying a single model of translation.

Implementation should reward meaningful progress along the translational continuum rather than only end-stage commercialization. NIH should also avoid creating incentives that favor commercial potential over scientific or public health value, particularly in fields where the greatest impact is achieved through prevention, implementation, healthcare delivery, or reducing health disparities.

By recognizing multiple pathways through which discoveries improve health, NIH can foster an innovation ecosystem that rewards meaningful societal impact while ensuring that publicly funded research delivers lasting benefits to patients and communities.

6. Foundational Scientific Exploration: Foundational scientific exploration is essential because transformative discoveries rarely emerge from highly predictable research. Yet current evaluation systems often favor projects with extensive preliminary data, lower scientific uncertainty, and more predictable outcomes. While this may reduce funding risk, it can also discourage the bold ideas that have historically driven major advances in biomedical science.

NIH should recognize scientifically rigorous exploration as a valuable contribution in itself. Evaluation should reward originality of research questions, development of innovative methods or technologies, generation of foundational knowledge, and creation of new scientific directions, regardless of whether the original hypothesis is ultimately confirmed. Scientific uncertainty should not be viewed as methodological weakness.

Implementation should distinguish scientific risk from methodological rigor. High-risk research should be held to the same standards of study design, transparency, and reproducibility as any other research while recognizing that ambitious questions inherently carry a greater likelihood of unexpected or negative findings. Well-designed studies that challenge prevailing assumptions or eliminate unsupported hypotheses should be recognized as advancing science rather than representing scientific failure.

The goal of NIH investment should not be to minimize the risk of failure, but to maximize the opportunity for meaningful discovery through scientifically rigorous exploration.

7. Public Impact: The ultimate public impact of biomedical research should be measured by how it improves human health and well-being rather than by scientific visibility or commercial success alone. Publications, citations, patents, and media attention are useful indicators, but they are not the ultimate outcomes of publicly funded research. NIH should prioritize how research changes clinical practice,

healthcare delivery, public health, and the lives of patients and communities.
Evaluation should recognize multiple pathways to societal impact, including incorporation into clinical guidelines or health policy, implementation of evidence-based interventions, improvements in patient outcomes, development of accessible diagnostics, technologies, software, educational resources, and scientific tools that enable future discoveries. Public trust should also be recognized as an important outcome, built through rigorous, transparent, ethical, and responsible science.
Implementation should recognize that meaningful public impact is often cumulative, collaborative, and realized over many years. Expectations should therefore reflect differences in disciplines and translational pathways while maintaining a common focus on improving health and generating lasting societal benefit.
Ultimately, publicly funded research should be evaluated not only by the discoveries it generates, but by the extent to which those discoveries become accessible, trusted, adopted, and beneficial to society. By aligning research assessment with these outcomes, NIH can ensure that public investment delivers the greatest possible benefit to science, healthcare, and the American public.

8. Other Areas: Beyond the domains discussed above, NIH should explicitly recognize scientific stewardship as a core component of scientific impact. Many activities that strengthen the biomedical research enterprise—including developing community standards, maintaining shared resources, building collaborative networks, improving research integrity, contributing to peer review, harmonizing methodologies, and creating infrastructure for future discovery—remain largely invisible in traditional research assessment despite generating substantial long-term value.
More broadly, NIH should adopt a multidimensional evaluation framework that recognizes there is no single pathway to scientific excellence. Researchers contribute through discovery, methodological innovation, mentorship, collaboration, stewardship, translation, implementation, and scientific leadership. Rather than relying on single metrics or composite scores, NIH should evaluate a balanced portfolio of complementary contributions that collectively strengthen the biomedical research ecosystem and maximize the long-term value of public investment in biomedical research.

I commend NIH for undertaking this important initiative. The greatest opportunity is not simply to develop new metrics, but to redesign the incentives that shape biomedical research. Ultimately, research assessment should reward the behaviors that maximize the long-term value of publicly funded science rather than primarily measuring individual academic productivity. The incentives NIH establishes today will shape the culture of biomedical research for decades to come. By recognizing scientific contributions that improve the quality, reliability, translation, and societal impact of research, NIH can strengthen the biomedical research ecosystem while maximizing the public benefit of its investments.

106. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/22/2026

1. Rigor and Reproducibility: A panel can list out traditional interventions that have not been re-evaluated in the past 10-20years (e.g., routine blood tests in adults starting annually at age 18, daily or twice daily labs in inpatients, pap smears in women who have had the hpv vaccine) and then an RFA or dedicated funds where researchers can evaluate the impact of these traditional interventions.

2. Data, Software, and Model Sharing: This is one space where I am not sure whats the best approach. There is already so much data out there and I cannot imagine anyone out there wanting to run my data again to validate my results. But that may be specific to my work. I do believe any observational studies funded by the NIH should automatically include a separate funding announcement that can be applied for to validate the results once the authors make their datasets available. There are many bioinformatics/statisticians who could contribute to validating findings.

3. Training and Mentorship: The traditional indicators (publications, grants, clinical trials, appointments) are likely sufficient because there are so many unrealized gains from training that cannot be captured. For example, a clinician who is trained in research methodology is reading the literature and applying findings in their practice- something that directly impacts a patients life. There isnt any way this can be captured.

4. Collaboration: Team based science is already being enforced through study sections where peer reviewers request new study personnel to be added for their expertise. What is important is not pushing team based science as much because it creates an incentive to include members who may not be truly needed to accomplish the goals of the project but are included just to satisfy the study section reviewers.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: It is the study section peer review process that thwarts bold and innovative ideas. It may also be that those who sell "bold and innovative" ideas are just good at marketing and the ideas are not truly bold or innovative. Having served on many study sections- I am yet not sure if any of the grants I have supported have yielded a single discovery that has helped people live better or longer lives. Yet I hear so much excitement from my colleagues about how a particular idea is so "innovative"- perhaps innovation is over-rated and unnecessary for many grants. Sometimes you just need the grunt work to answer a basic question using standard methods- rigor and reproducibility work will never be innovative for example. Factor 1 already removed innovation (it really should only be about

significance)- yet we are getting comments like- there is no innovation. Maybe have "pitch" sessions for the IC directors where they invite 10-20% of applicants of grants that were competitive not discussed or not discussed (but whose ideas, if they work, could be game-changers and where projects are fatally flawed)- the IC directors can pick a few of these for funding and compare how they perform with the 1-3rd percentile grants. I wouldn't be surprised if they perform similarly using current metrics.

7. Public Impact: Its important that IC directors look at outcomes of grants (before funding) in a way that peer reviewers may not be looking. There are some very excellent grant writers out there who can sell ideas and there are reviewers who do similar things and so they get perfect scores. But when you look at their outcomes it becomes obvious there is no clear translation plan. For example, if someone is creating a public health messaging campaign to figure out whether it might land well to alter a certain behavior (for example increase medication adherence or increase acceptance of screening)- if the whole grant will create and test messages and then they publish their results - a tax payer may wonder how this actually changed the behavior in question. There must be a requirement to include a trial where the campaign is launched in the real world and the behaviors are tested to understand efficacy- otherwise these campaigns are created and never read a single person who could benefit. Similarly there are many basic science studies with no plan for translation even if they work- while it may be that some pharma company will pick it up for development it would help to have a standard pipeline to ensure the best and most promising discoveries in basic science are moved forward into drug development and phase 1 clinical trials.

8. Other Areas: AI could be leveraged to resurrect ideas from the past 5-10 years that researchers may have abandoned because study section peer reviewers just wouldn't score them high enough for funding. Applicants could be asked to reapply and then selection could be done by lottery. In general a small portion of grants that are not fatally flawed in some way (any unfunded grant can fall in this space) should be funded regardless of the peer review process via lottery (and perhaps after a pitch or video meeting between the PO and the applicant to ensure they can defend their ideas) and then outcomes of these lottery funded grants should be compared with the grants that are funded through the traditional process. The rock stars who keep getting grants have figured out a winning formula (they write well and have three peer-reviewers who think just like them). They don't necessarily move the science forward and they aren't necessarily creating ground-breaking discoveries. Ideas that are outside the realm of what peer-reviewers would think of funding are likely falling in the lower score areas but they may yield something. Another space to look is at the performance of the study sections- if there are metrics that come out from this process that would identify "impact"- we may find that certain study sections perform worse at identifying good grants. Those study sections likely need to be disbanded or restructured (or have more than one study section chair and train them really well in intervening when reviewers go down a path of reasoning that isnt helpful).

107. Joseph Braun

I am responding: On behalf of myself

Name: Joseph Braun

Name of Organization: Brown University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/22/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: This is important, but needs to have resources available to support it (e.g., money for staff, costs for platform maintenance).

3. Training and Mentorship: I believe that this is very undervalued in our evaluation of impact. I also believe that this needs to carefully consider one's career stage. Often early stage investigators are dinged for not having had prior trainees when they have a trainee submit a K or F grant, but someone has to be their first trainee!

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: I think it is important to recognize that not all scientists will be doing all of these things at all stages of their careers. Thus, any new method of evaluating them needs to consider career stage, field of study, and current scientific work.

108. Sandia biotech Inc.

I am responding: On behalf of an organization

Name: Michael Humbert

Name of Organization: Sandia biotech Inc.

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Organizational Official

Date Submitted: 7/22/2026

1. Rigor and Reproducibility: PROPOSAL: SOLVING THE REAGENT REPRODUCIBILITY CRISIS VIA SEQUENCE-DEFINED FLUORABODY® PROBES

1. THE TRUST PROBLEM IN CONTEMPORARY SCIENCE
The NIH must address the fact that reagent variability is the #1 cause of assay failure, driving 36% of all cases. Traditional fluorescent labeling fails due to inconsistent fluorophore stoichiometry, brightness, and spectral drift caused by chemical conjugation. Legacy fluorescent antibodies vary heavily in quality, forcing labs to spend $1K+ per reagent to re-validate what they cannot trust. This systemic problem contributes to:
a) ~$170B in annual global research waste.
b) ~50% of all preclinical research failing to reproduce.
c) Over 50% of commercial antibodies failing validation as sold.
d) Independent 2026 findings flagging 462 problematic validation images in major commercial catalogs, proving that marketing imagery cannot replace molecular validation.

2. THE PLATFORM SOLUTION: THE FLUORABODY® ENGINE
To restore reproducibility, the NIH should incentivize a transition to sequence-defined, animal-free reagents built at scale. The Sandia Biotech FluorAbody® Engine solves this by turning known antibody sequences directly into intrinsically fluorescent, sequence-defined scFv reagents.

By grafting target sequences onto an engineered scaffold fused to a fluorescent protein (with 8 colors available), the platform builds fluorescence directly into the antibody sequence itself. This guarantees predictable, controlled fluorophore loading and strict stoichiometry. True monoclonality removes polyclonal heterogeneity and off-target binding, delivering a reliable quantitative readout that stays trustworthy forever with zero lot-to-lot drift.

3. MEASURABLE POLICY INDICATORS
The NIH should implement a "Reagent Fidelity Metric" within peer review. Grant applications should receive preferential scoring under the "Approach" criterion if they commit to utilizing sequence-defined, recombinantly expressed biological detection engines over volatile, chemically conjugated legacy reagents.

2. Data, Software, and Model Sharing: PROPOSAL: ENHANCING TRANSPARENCY THROUGH SEQUENCE-LEVEL ASSAY DISCLOSURE

1. THE SHARING GAP IN BIOMEDICAL RESEARCH
True data sharing and assay reuse under current Data Management and Sharing (DMS) guidelines are blocked because legacy reagents are molecular "black boxes." When investigators share raw quantitative data, subsequent labs cannot compare or identically repeat the experiments because the exact physical properties, sequence identities, and labeling variations of the commercial probes are unrecorded. This inconsistent labeling undermines quantitative interpretation and breaks reproducibility across labs.

2. A DIRECT BRIDGE FROM DIGITAL SEQUENCE TO REAGENT
The future of data sharing relies on digital sequence databases, AI, and in-silico binder programs. The FluorAbody® Engine acts as a direct sequence-to-reagent workflow. The engine takes digital sequence inputs, extracts the paratope, and integrates it onto an engineered scaffold for recombinant expression. This converts digital sequences instantly into detection-ready, quantitative reporters. Because the blueprint is digital and sequence-defined, the physical identity of the probe is fully transparent and auditable.

3. STRATEGIES FOR MANDATORY REPORTING
The NIH should update mandatory DMS checklists to include a "Reagent Blueprint Table." PIs using fluorescent detection should be required to log the exact amino acid sequences or precise molecular definitions of their probes. Utilizing open-access sequence blueprints ensures any laboratory worldwide can express the identical physical tool, enabling lossless data reuse and flawless replication.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: PROPOSAL: ACCELERATING PRECLINICAL TRANSLATION VIA WORKFLOW COMPRESSION AND SCALABLE ENGINES

1. CLINICAL AND TRANSLATIONAL BOTTLENECKS
Translating academic breakthroughs into robust diagnostic or clinical products is severely bottlenecked by time waste and cost, contributing to a $28B annual U.S. R&D loss. Traditional immunofluorescence requires a multi-step workflow (fix, wash, block, primary antibody, wash, secondary antibody, wash) taking over 4 hours. This high-volume workflow amplifies human error and reagent variability, causing mismatched research and clinical outcomes.

2. OPERATIONAL IMPACT OF WORKFLOW COMPRESSION
Federal funding should prioritize entrepreneurial innovations that compress assay workflows without compromising quality. The FluorAbody® framework compresses detection into a rapid, single-step direct-staining protocol (fix, wash, stain, wash) completed in under 2 hours. This streamlined approach delivers:50% less assay time.43% fewer protocol steps.2x faster results.

In practice, this operational efficiency yields a sharp, highly localized, low-background signal that improves tissue penetration and accelerates biomarker validation (e.g., membrane HER2, EphA2, EGFR)

in regulated environments.

3. COMMERCIAL SCALABILITY AND ACQUISITION TARGETING
The NIH should utilize SBIR funding to support platform engines that scale seamlessly across entire product pipelines. The FluorAbody® Engine features a highly defensible model, including an active patent license from UNM Rainforest Innovations with no field-of-use restrictions, and a commercial portfolio of 71 unique validated targets across oncology, inflammation, and neurodegeneration. Transitioning small business incentives toward sequence-defined, recombinant platforms ensures that early-stage discoveries cleanly integrate into corporate automation catalogs, accelerating commercial diagnostic translation.

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

109. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Healthcare Institution

Role: Investigator/Researcher

Date Submitted: 7/23/2026

1. Rigor and Reproducibility: There is a laudable effort from NIH and publishers for science that is accurate, rigorous and reproducible. However, there an in-built bias as many scientists are known and well connected, sometimes they are able to publish okay studies in big journals. This practice should be somehow reduced/discouraged. NIH should consider funding proposals which are based on medium impact factor journals but with detailed analysis/robust data.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: Although collaboration is a big part of any scientific endeavor, its increasingly connected to "network" and "networking" for getting grant scored better and funded. May be NIH should ask every PI who is submitting a RO1 kind proposal to submit a list of talks they organized or faculties they invited to their home institute. None of these people should be in a grant review meeting for that host PI. Someone's NETWORK MUST NOT DRIVE FUNDING SUCEESS. I think NIH should limit number of R01 or large DoD grants a PI can have at a given time (feedback sought in another Request for Information) and PI MUST BE ASKED TO DECLARE A LIST OF SCIENTISTS/FACULTIES THEY INVITED/HOSTED at their home institute IN LAST 5 YEARS to reduce grant review bias/undeclared potential conflict.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

110. Dennis Drayna

I am responding: On behalf of myself

Name: Dennis Drayna

Name of Organization: NIH (Emeritus)

Type of Organization: Government Agency

Role: Investigator/Researcher

Date Submitted: 7/23/2026

1. Rigor and Reproducibility: You'll have to convince NIH study sections that giving money to a replication study is just as exciting and interesting as giving money to a new area of investigation.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: MD's wishing to do laboratory research face a fundamental and unchangeable hurdle, which is that any time away from the lab (for clinical work in particular) turns them into less than full time bench scientists. This puts them at a disadvantage for funding compared to those who devote full time to bench research. Lowering standards for the evaluation of funding applications for those who spend some of their time doing something other than bench research is not the answer to this problem. Personally, I believe one solution would be to remain insistent on the very highest, most competitive standards in bench research but offer a bonus system for applicants who bring patients or other clinical resources to the application.

4. Collaboration: When I started in research 50 years ago, I wanted to have (and often did have) as few authors on my papers as possible. Few authors on the paper other than me indicated that I did the heavy lifting and it raised my visibility and prestige within the scientific community. Now, journals tacitly want large (sometimes very large) authorship lists, as they believe papers with more authors equal more work and thus more importance and more impact. For example, I published some of my most important papers with 2 or 3 total authors on those papers. Today there are essentially zero papers in Cell, to choose an example, with 2 or 3 authors. This and other factors have now severely limited the career growth and development opportunities for young investigators, who now often labor under the intellectual and financial control of prominent scientists. This may be an efficient way of accomplishing one person's scientific vision and goals, but I believe we are squandering the most energetic and creative years of scientists' careers. Big Team science has advantages, but it should be recognized that this approach comes at a cost to the development of our future scientific workforce. The K99/R00 mechanism is one step toward solving this problem, but the real test is in the R01 world. Finding good people and getting them well started is what we need to do to maintain our world leadership in science.

5. Entrepreneurship and Translation: Put more money into SBIR funding, and solve the problem that arises when investors refuse to allow those funds to buy into the company and thus dilute their investment and control. To put it bluntly, the NIH's money invested in an enterprise is the same color as the investors' money, plain and simple.

6. Foundational Scientific Exploration: This is what made US science the best in the world. Remove political meddling and keep peer review as the primary gate keeper of funding. Make it a legal requirement that any NDA to the FDA must list the 25 most important basic research papers that led up to the development of that drug. And, insist that the funding source for those papers (it will overwhelmingly be the NIH) are an integral part of that list and are clearly visible to the public.

7. Public Impact: N/A

8. Other Areas: N/A

111. Frederick O. Cope, PhD MS MSc FACN

I am responding: On behalf of myself

Name: Frederick O. Cope, PhD MS MSc FACN

Name of Organization: Physis International LLC, A Drug Development Organization

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Organizational Official

Date Submitted: 7/23/2026

1. Rigor and Reproducibility: Perhaps this should have been a consideration long ago. In today's published science realm, I have never seen the published outcomes less rigorous and more suspect with regard to "real or unreal". The question then is, in lieu of unrealistic approaches to control publications, what is the best way to assure reliable results. Results, if reliable, DO NOT have be the outcome of an unflawed hypothesis. They can the outcome of a flawed notion, as long as the results are replicable. The hypothesis is the focus of testing and imagination, not the result of testing. Perhaps the best approach to all of this is a highly geographically integrated team with a non-project oversight group, similar to clinical trials. The downside to this, is a significant increase in cost per project. But in the end, if 20-60% (maybe more, depending on research specialty) of all research is not replicable, is it really less cost-effective? The elimination of poor research, fraud, and waste will never be 100%. But new approaches appear to be necessary to stem the tide of lousy research.

2. Data, Software, and Model Sharing: As my comments above suggest, the sharing data sets (ORIGINAL), tightening the data capture procedures, videoing some work in progress (although invasive), auditing source documents, having a CCRA-type person with authority as part of a department or team, etc., may necessary. All of these things are invasive and may be counterproductive, e.g., stifling innovation. But some measures need to be adopted in order to maintain scientific integrity (which has suffered profoundly over the last 6 years, at least).

3. Training and Mentorship: In my experience, mentorship and outcome are solely dependent on the particular mentor and how their operations are run---- very individual dependent. This is a very difficult matter to manage. It is also institutionally impacted. Many mentors seek to be very hands on while institutionally, they are warned, "Hands off". Any change in how this is handled needs to be a "do IT or lose the funding" approach. If you show institutional favor, all is lost. But this too, is a problem. You cannot force this onto institutions where the requirements impose exorbitant costs, where such an approach favors the wealthy institutions. This will be simply more of same IVY crap that is evident among some study sections.

4. Collaboration: I am all for it. But in order to accomplish this, you need to impose rules that institutionally reward the participants. During my tenure at a major academic institution, we positioned the entirety of the research effort to become a team/collaborative effort. This did NOT go well with department chairs. Systems of reward, both monetary and productivity based, were not adapted to service a team effort. Push-back was tremendous. But the solution was to revamp the rewards and

promotion system. In the end it was do it or else (or else your department won't get anything). The days of a single or two investigators working behind closed doors, publishing lone results, and waiting for the promotion, ended. Perhaps this is really telling with regard to current institutional structures and rewards systems that need a revamping.

5. Entrepreneurship and Translation: Franky, this is the only thing that matters. If this is not the focus, what are you really doing???? Bench-to-bedside MUST BE a major part of the equation. It took center stage at my institution, and it paid off (not without the pain of restructuring).

6. Foundational Scientific Exploration: See "Rigor and Responsibility"

7. Public Impact: If the NIH mission is to be a consideration for the expenditure of any tax dollars, then public impact must be outcome optic. But really, this is a public health and safety optic. And it flows from ALL FEDERAL STRUCTURES. When one fails, e.g., CDC or FDA or NIH, all are seen as a failure. Better integration of the Federal (and State) agencies better soon happen, or the public, fed up with catastrophes over the past 6-12 years, will demands cuts or major reorganization. Perhaps it is time for MAJOR reorganization.

8. Other Areas: 1. With all that is considered in these sections, I would also suggest a complete overhaul of the study section system. My experience is that it is institutionally biased. Heavy northeast and west coast. And please don't tell me this is where the great science is REALLY DONE. Find a better way to assess applications or to divey up funding.

2. Indirect costs: In the end if you wish to find a way to improve rewards, consider a significant reduction for indirect costs. This can give you the money you need without an increase in budget. After all, why should any institution get more than another. If their overhead is 20-40% higher, tough!! Fixed fee overhead is the only way to go (try 25%). Use the difference to structure a rewards program for collaborations, for clinical translation, for outcomes, for true impactful science breakthroughs. You really have to ask yourself, why the hell is Harvard, with the largest endowment of any university, getting 68.5% overhead when other universities and small business biotech get ~45% and 10%, respectively? Com'on, you can do better than that, and the optics....not a good picture.

112. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Other

Date Submitted: 7/24/2026

1. Rigor and Reproducibility: While the effort to maximize

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: The notion of improving impact and rigor is laudable. It has hardly been the first time that folks (researchers included) have thought about this and the issue it raises.

However, the reason it has remained an ongoing concern is that the solution is not easy/straightforward, and there is no answer that is without negative repercussions.

How are we to measure how impactful something is? Take a paper for example. Is it how many people read it? How many people cite it? How many marketable products it generates? Whether it leads to another grant? Whether it leads to private funding? Importantly, what is the timeline? There is plenty of research that doesn't make a high mark on any of these metrics in the immediate short term but then years or decades later ends up informing other discoveries and products. In many of those cases, the primary investigator is onto different types of projects or is retired altogether. So how can you reward 'impact' in that situation?

This effort broadly strikes me as a way to weaponize results against researchers and their future efforts. In other words, especially in the new "unified funding strategy" it gives administrators at NIH just one more reason they can use to pick winners and losers, potentially under influence by politicians, rather than letting the scientific review process play out.

113. Jessica Paredes

I am responding: On behalf of myself

Name: Jessica Paredes

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 7/24/2026

1. Rigor and Reproducibility: Rigor and reproducibility are essential to scientific progress but are difficult to measure and reward. Independent replication often requires specialized expertise, unique resources, and substantial time and funding. In many cases, expecting individual laboratories to routinely reproduce influential findings may not be feasible without dedicated resources or programs focused on replication.

One potential approach would be to create mechanisms for researchers to confidentially report unsuccessful replication attempts, including methods used, limitations encountered, and possible explanations for discrepancies. These reports should not automatically be interpreted as evidence that original findings are incorrect, as differences in expertise, materials, methods, or experimental conditions may contribute. However, aggregating such information could identify areas where formal replication efforts are most valuable.

It is also important to recognize that reproducibility often emerges through the cumulative process of science. Multiple laboratories investigating similar questions using different approaches contribute to confidence in findings over time. Reproducibility should therefore be viewed not only as direct replication of individual studies, but also as the broader process by which scientific knowledge is validated.

NIH should also recognize that creating reusable resources—including validated datasets, harmonized data, and well-maintained repositories—is an important scientific contribution. These activities can enable future discoveries but currently receive less recognition than publications. Grant evaluation should consider the long-term value of shared resources, data validation, and reuse in addition to traditional metrics such as papers and citations. Incentives and funding mechanisms that support data curation, integration, and validation would encourage researchers to invest in these activities while maintaining career competitiveness.

2. Data, Software, and Model Sharing: Making scientific data, software, and models accessible is essential for accelerating biomedical discovery. As an early-career scientist, access to existing datasets and computational resources creates opportunities to address important questions that may otherwise be limited by the cost and infrastructure required for new data generation. Shared resources enable validation, collaboration, and greater return on publicly funded research investments.

However, there are legitimate concerns from investigators who generate these resources. Large datasets require years of funding, specialized expertise, and substantial effort to collect, curate, and validate. Researchers may worry about losing scientific priority, having ideas pursued before they can complete analyses, receiving insufficient credit, or creating disincentives for future data generation.

A potential solution is to distinguish between data accessibility and release of the entire research process. Making raw or minimally processed data available with appropriate metadata and documentation can enable reuse while preserving opportunities for original investigators to develop additional analyses. Importantly, datasets must include sufficient context—such as experimental design, quality-control procedures, and limitations—to be scientifically interpretable. Data that are accessible but poorly documented may have limited value.

NIH should also recognize that generating knowledge from existing datasets requires significant expertise and effort. Developing models, integrating datasets, and performing secondary analyses are important scientific contributions that should be rewarded. At the same time, new data generation remains essential for rigor and reproducibility because independent data collection is often needed to validate findings, address limitations, and test new hypotheses.

NIH could promote this balance by:

Recognizing datasets, software, models, and secondary analyses as meaningful scientific outputs.
Providing credit for data generation, stewardship, validation, and reuse.
Supporting infrastructure for data documentation, attribution, and long-term maintenance.
Encouraging responsible sharing practices that protect contributor recognition while maximizing scientific value.

A successful sharing ecosystem should reward both researchers who generate high-quality resources and researchers who use those resources to make new discoveries.

3. Training and Mentorship: Mentorship is essential for sustaining the biomedical workforce, but it often competes with incentives that prioritize publications and grants. Effective mentorship requires substantial time and effort, yet these contributions are difficult to measure and frequently undervalued.

The stability of the training pipeline is also a major concern. Postdoctoral positions are limited, and compensation may not be sufficient to support researchers and their families given current costs of living. These pressures can discourage talented individuals from pursuing long-term research careers. In addition, many postdocs face uncertainty about academic career prospects and may not have sufficient opportunities to develop the grant writing, leadership, and independent research skills needed for long-term success. Creating additional pathways that allow postdoctoral researchers to pursue more independent research projects with appropriate mentorship and support could help prepare the next generation of scientific leaders.

Current incentives may also limit access to training opportunities by favoring trainees with existing publication records rather than recognizing potential, growth, and diverse paths into research. Similarly, many early research experiences, including volunteer opportunities, are more accessible to individuals

with financial flexibility. Expanding supported opportunities for research training could help reduce barriers for talented individuals who cannot afford unpaid or low-paid experiences. Personally, I believe all scientists should be rewarded for outreach (as connection to the public is critical to the long-term stability of the science).

Mentorship metrics should avoid becoming another publication-based measure that rewards only already successful trainees. Instead, they should consider trainee development, skill growth, career outcomes, and the creation of supportive research environments.

NIH should consider:

Recognizing mentorship as a meaningful scientific contribution.
Rewarding researchers who invest in training, including mentorship that does not immediately result in publications.
Supporting sustainable postdoctoral and trainee positions.
Expanding accessible, supported research training opportunities.
Measuring mentorship impact through trainee growth and long-term contributions to science.

Training future scientists is itself a critical research investment, and the biomedical enterprise should better recognize and support those who make that investment.

4. Collaboration: As biomedical research becomes increasingly interdisciplinary, collaboration and team-based science are essential for addressing complex problems. However, policies and incentives should be aligned with this goal. Restricting the number of times a principal investigator can serve as a multi-PI (MPI) on grants may unintentionally create barriers to collaboration, limit productive research partnerships, and reduce opportunities for trainees to participate in interdisciplinary science.

MPI structures can be particularly valuable for early-career researchers by providing access to complementary expertise, infrastructure, and mentorship that may not exist within a single laboratory. Restricting these collaborations may unintentionally reinforce advantages for large, established laboratories while limiting opportunities for trainees and emerging investigators to develop within collaborative research environments.

NIH should consider approaches that recognize individual contributions within team science, including leadership, coordination, mentorship, methodological expertise, and development of shared resources. Evaluation and funding policies should encourage effective collaboration while ensuring accountability, rather than creating barriers that discourage the interdisciplinary partnerships needed for modern biomedical research.

5. Entrepreneurship and Translation: Translating discoveries into products and clinical advances requires researchers to take risks, explore new directions, and engage with partners outside traditional academic settings. However, pursuing entrepreneurial activities often requires significant time away from generating publications and grants, which can create career risks for investigators.

Researchers may reasonably ask how an applicant who dedicates time to developing a venture,

technology, or translational effort—but consequently produces fewer publications during that period—will be evaluated in future funding decisions. Without mechanisms that recognize entrepreneurial efforts as legitimate scientific contributions, investigators may be discouraged from pursuing opportunities that could have significant public impact.

NIH should consider ways to reduce these barriers by recognizing innovation, technology development, commercialization efforts, and industry partnerships as meaningful research outcomes. Evaluation criteria should consider the broader impact of a researcher's contributions rather than relying primarily on publication metrics.

Supporting entrepreneurial pathways may also require mechanisms that provide financial and career stability during periods of transition, allowing researchers to pursue translation without jeopardizing long-term scientific careers. A successful innovation ecosystem should reward thoughtful risk-taking and recognize that impactful discoveries may require time away from traditional academic productivity measures.

6. Foundational Scientific Exploration: Basic research and high-risk, high-reward initiatives are essential because scientific progress often depends on pursuing questions whose importance is not immediately apparent. As a society, we often underestimate how much remains unknown and how difficult it is to predict which discoveries will ultimately transform medicine or technology.

Many major advances have emerged because researchers were able to pursue ideas that initially appeared speculative, impractical, or disconnected from immediate applications. In many cases, the significance of a discovery only becomes clear after years of additional research, when foundational knowledge enables new technologies, therapies, or unexpected applications.

NIH should continue to support scientific exploration that allows investigators to pursue ambitious questions without requiring immediate evidence of translational potential. Evaluation systems should recognize that failure is an inherent part of discovery and that long-term scientific progress depends on maintaining a portfolio that includes both incremental advances and bold, uncertain ideas.

Supporting foundational research requires incentives that value scientific insight, development of new knowledge, and creation of future opportunities—not only short-term outcomes. The ability to pursue unanswered questions is essential because innovation often occurs only after researchers have invested the time needed to understand a field deeply enough to reveal its broader potential.

To encourage bold scientific exploration, funding structures should consider the career and financial risks associated with pursuing uncertain ideas. Although investigators may technically complete projects earlier than the allotted grant period, doing so may create financial and career disadvantages because future funding often depends on sustained productivity and continued support. More flexible funding mechanisms, shorter grant cycles for exploratory work, milestone-based funding, or opportunities to transition based on scientific contribution rather than only successful outcomes could allow researchers to pursue high-risk ideas while reducing career risk.

Importantly, high-risk research should not be evaluated solely by whether the original hypothesis was

confirmed. Studies that generate valuable knowledge, reveal limitations, develop new methods, produce important datasets, or redirect a field can represent meaningful scientific contributions even when the expected outcome is not achieved.

7. Public Impact: Biomedical research can have broad impacts beyond publications, including improvements in health outcomes, clinical practice, economic development, accessibility, and public trust. However, researchers often lack the resources, training, and opportunities needed to develop these broader impacts.

NIH could support public impact by creating funding opportunities, workshops, and conferences that help researchers build partnerships beyond traditional academic settings, including with clinicians, patients, communities, industry, educators, and policymakers. These opportunities could help investigators translate discoveries, communicate findings effectively, and identify applications that may not be apparent within a single research environment.

Public impact should also be recognized as a meaningful research contribution. Evaluation metrics could consider outcomes such as adoption of research findings, development of useful resources, contributions to clinical practice or policy, community engagement, improved accessibility, and successful translation of discoveries into practice.

However, public impact is often difficult to predict at the time research is funded. Many foundational discoveries produce societal benefits only after years of additional development and understanding. NIH should avoid creating incentives that prioritize only projects with immediate or easily measurable applications, as this may undervalue basic research and high-risk discoveries.

Supporting public impact requires both recognizing outcomes and creating the infrastructure, partnerships, and opportunities that allow discoveries to reach society. A successful approach should value near-term engagement as well as long-term contributions to health and societal advancement.

8. Other Areas: N/A

114. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 7/26/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: The biomedical scientific community is in a dramatic funding crisis and many people are leaving academia or leaving the country. In the current funding climate, there are two very basic and quite straight-forward changes that could –and should–be made to better distribute NIH funding so that more of the people who can produce the results–at all levels–that we all desire will be funded.

First, eliminate funding based on an individual's connection to a particular disease; rather, fund researchers on the Howard Hughes model, which funds the scientific rigor, abilities and accomplishments of the individual or team. There are WAY too many clinicians who receive funding because they are experts in a particular disease area, not because they are good scientists, but because there's an urgent–sometimes desperate–need to find therapies/cures. Instead of pushing a field forward, much of the output from these groups is derivative and incremental and cannot meet NIH's goals. A genuine, deep desire to help one's patients, or how accomplished someone may be in their clinical specialty never made anyone a scientist, let alone a good one.
There are plenty of M.D.s who are great scientists (Harold Varmus, etc.), but the only reliable way to distribute VERY limited funds is to get those funds to the best scientists, period - not to people who work on a particular disease area. One of the best ways to do this is to use the Howards Hughes model and have a panel of outside experts (study sections have far too many conflicts) to evaluate a grant applicant's abilities as a scientist.

Second, award funding to people who do scientific research FULL TIME. There are, again, WAY too many clinicians who do "research" as an afterthought. You would not hire a lawyer who practiced law half a day or one day a week, but such a schedule is routine in virtually all major medical centers, again because there is disease-based grant money available. Success in scientific research requires a major

investment of time and effort, and it is a privilege that the taxpayer expects to have carried out on a full-time basis.

If you do just these two things, you will jump-start the research enterprise in this country and accelerate progress towards understanding and cures (not to mention the enormous positive affect it has on the economy). If you want to be even more effective, increase NIH funding.

7. Public Impact: N/A

8. Other Areas: N/A

115. N/A

I am responding: On behalf of an organization

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Other

Role - Other: Evaluation team

Date Submitted: 7/27/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: Various indicators/tools for Training and Mentoring impact evaluation such as:
- High-level success story – to share to public on the trainee journey and their research outcome
- Post-award progress reporting for academic output tied to funding (presentations, posters, papers), follow on grants for calculating ROI.
- Utilization of Flight Tracker for Scholars for post-award progress reporting. Flight Tracker is a REDCap module developed by Vanderbilt University, used to track the academic career development of CTSI education program scholars by scraping data from several databases and repositories, including NIH REPORTER and PubMed. FT will help research staff in reporting the outcome and impact more automatically, which helps staff's bandwidth.
- Metric to measure the percent of protected time (need to set a minimum threshold) across institutions.
- Standardized survey to measure satisfaction of training across institutions

4. Collaboration: Assessing:
Evaluating team science needs to define success across diverse disciplines and project phases, ensure fair credit distribution among team members, and navigate complex organizational structures. For implementation approach, team science needs supporting staff who can make ongoing observations of team dynamics and implement best practice interventions.
Key outcomes include achieving scientific goals, improving collaborative skills, and fostering a positive team environment. Metrics for these outcomes include productivity, publications, grant funding, team satisfaction, percent of potential academic and industry collaborations outside home institutions utilized and engagement.
Evaluating team science needs to focus on both team processes and outcomes including assessing communication, collaboration, and conflict resolution within the team, as well as the impact of the team's work on research outputs, scientific advancement, and practical applications.
For collaboration with community, CIROP (Community Impacts of Research-Oriented Partnerships) survey can be used to measure the impact of research partnership from the perspective of community

members. This survey will help to demonstrate impact of research partnership with community members in the areas of: 1) personal knowledge development, 2) personal research skill development, 3) organizational/group access to and use of information, and 4) community and organizational development.

Recognizing:
Tie the achievement of met goals to promotion opportunities.
Best practice articles/case studies/success stories.

5. Entrepreneurship and Translation: Assess:
Various tools for Entrepreneurship and Translation impact evaluation such as:
- Number and type of scientific and/or operational innovations developed and/or disseminated
- Number and type of patents or trademarks filed and/or received (e.g. IP data, implementation science etc.)
- Number and type of FDA approvals received by CTSA Hubs
Utilization of TSBM framework for impact evaluation will be suitable for translating/ demonstrating the impact on real world in four areas: four areas: Clinical and Medical Benefits, Community and Public Health Benefits, Economic Benefits, Policy and Legislative Benefits.

Incentivize:
Tie the achievement of met goals to promotion opportunities.
Best practice articles/case studies/success stories.

Integrate:
Develop NIH common metrics and approaches

6. Foundational Scientific Exploration: N/A

7. Public Impact: Assessing:
Utilization of Translational Science Benefit Model (TSBM) for impact evaluation. TSBM is a framework to help demonstrate impact of translational science in the real world. The model identifies benefits of clinical and translational research in four areas:
• Clinical and Medical Benefits (person-involved care, procedures & guidelines, tools & products)
• Community and Public Health Benefits (community engagement & wellness, health activities & products, health care characteristics, health promotion)
• Economic Benefits (commercial products, financial savings & benefits, healthcare workforce)
• Policy and Legislative Benefits (advisory activities, community-centered policy, policies & legislation)
TSBM has available tools to help guide data collection, building case study, and building impact profile (Washington University Tool Kit). TSBM helps to guide researchers to share their research impact in real world setting to public. Since research impact can take many years, TSBM framework allows both demonstrated (benefits that had been achieved) and potential (benefits that had not yet been achieved but could be achieved by the project) indicators of benefits.

Recognizing:
Tie the achievement of met goals to promotion opportunities.
Best practice articles/case studies/success stories.

8. Other Areas: N/A

116. Abdel Abdel-Rahman

I am responding: On behalf of myself

Name: Abdel Abdel-Rahman

Name of Organization: East Carolina University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/28/2026

1. Rigor and Reproducibility: Replicability of human pathophysiological or diseases in basic research models needs to be recognized and rewarded by the NIH.

2. Data, Software, and Model Sharing: Data reuse is a great idea. However, there have to be safeguards to protect the rights of the original data depositor both when the reused data is reused in NIH applications or submitted manuscripts.

3. Training and Mentorship: N/A

4. Collaboration: This effort has significantly grown and is currently well reconized by the NIH. Perhaps, less emphasis needs to be devoted to the names of the established researchers and institutions than current practices by reviewers. The efforts of the PI and collaborators need to be evaluated based on their capabilities to conduct the proposed studies rather than on their standing and reputation.

5. Entrepreneurship and Translation: Both factors are important. However, in recent years, reviewers have been paying too much attention to catchy techniques/methodologies when evaluating the innovation of the application. While the efforts of the reviewers are highly appreciated, generational changes led to downplaying the importance of animal studies using basic and translationally important techniques.

6. Foundational Scientific Exploration: This is a great idea that reinforces the comment just above this idea.

7. Public Impact: Yes, this endeavor deserves continued recognition.

8. Other Areas: N/A

117. Judith Grisel

I am responding: On behalf of myself

Name: Judith Grisel

Name of Organization: Bucknell University

Type of Organization: Academic Institution

Role: Other

Role - Other: Professor at an undergraduate institution

Date Submitted: 7/28/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: Science is a tedious process of knowledge accretion. Students depend on access to the latest research/thinking/findings to adapt to, and prepare for, future careers.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: I'm especially interested in slowing or reversing the rise in substance use disorders. Toward that end, I've begun a project in which students (high school, college, graduate) work in teams to translate one empirical paper at a time for social media. Thus scientific findings are amplified by young people, for young people. In this way, scientific understanding about this critical topic is able to reach the people who need it most so that they are able to make informed decisions.

8. Other Areas: N/A

118. Hoda Badr

I am responding: On behalf of myself

Name: Hoda Badr

Name of Organization: Sidney Kimmel Comprehensive Cancer Center at Thomas Jefferson University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/29/2026

1. Rigor and Reproducibility: Rigor should be evaluated using measures that reflect the full biomedical research enterprise, including basic, clinical, behavioral, implementation, and population sciences. Reproducibility should encompass transparent reporting of study methods, intervention protocols, analytic approaches, implementation procedures, and community engagement methods so that findings can be evaluated and applied in diverse settings.

Potential indicators: Public availability of study protocols, statistical analysis plans, intervention manuals, analytic code, common data elements, and successful independent replication across healthcare settings and populations.

Implementation: NIH could incorporate these measures into progress reports and final reports using standardized reporting expectations across funding mechanisms.

Potential unintended consequences: Expectations should be proportional to study design and resources so that requirements do not create unnecessary barriers for early-stage investigators or smaller institutions.

2. Data, Software, and Model Sharing: Scientific value extends beyond datasets and software. Many federally funded studies produce reusable resources such as intervention manuals, implementation toolkits, REDCap instruments, qualitative codebooks, common data elements, and longitudinal cohorts that enable future discovery.

Potential indicators: Number of external investigators using shared resources, incorporation into multicenter studies, citations of shared datasets or tools, and new grants or publications enabled by these resources.

Implementation: NIH could encourage standardized repositories and recognize resource development as a scholarly product within biosketches and progress reports.

Potential unintended consequences: Expectations for sharing should account for participant privacy, community agreements, and the additional resources required to curate high-quality datasets.

3. Training and Mentorship: Training should be evaluated by the success of the next generation of scientists rather than the number of trainees supported. Strong mentorship develops independent investigators, promotes interdisciplinary collaboration, and prepares scientists to address increasingly complex biomedical challenges.

Potential indicators: Transition to independent funding, scientific leadership, interdisciplinary publications, successful careers in biomedical research, and contributions to collaborative science.

Implementation: Standardized mentoring outcomes could be reported through existing NIH progress reports while recognizing differences across career stages and disciplines.

Potential unintended consequences: Evaluation should recognize that meaningful mentorship often extends beyond grant-funded trainees and that career trajectories differ across scientific fields.

4. Collaboration: Many of today's most important biomedical questions require multidisciplinary teams. Advances in prevention, treatment, survivorship, and chronic disease management increasingly depend on collaboration among laboratory scientists, clinicians, behavioral scientists, population scientists, implementation scientists, data scientists, and community partners.

Potential indicators: Sustained multidisciplinary collaborations, shared leadership, contribution-based authorship statements, development of shared research infrastructure, and collaborative grant funding.

Implementation: NIH could encourage standardized reporting of scientific contributions that complements traditional publication metrics.

Potential unintended consequences: These measures should complement rather than replace traditional indicators of scientific productivity, recognizing that collaboration opportunities vary across disciplines and career stages.

5. Entrepreneurship and Translation: Translation should be defined broadly to include both commercialization and implementation into routine healthcare and community settings. Many NIH-supported discoveries improve health through evidence-based care models, behavioral interventions, survivorship programs, caregiver interventions, patient navigation, and improvements in healthcare delivery rather than commercial products.

Potential indicators: Adoption by healthcare organizations, incorporation into professional guidelines, sustained implementation after funding, and documented improvements in patient care.

Implementation: Investigators could report implementation outcomes through existing NIH reporting mechanisms using standardized definitions.

Potential unintended consequences: Translation metrics should recognize that opportunities for commercialization differ substantially across scientific disciplines.

6. Foundational Scientific Exploration: Transformative discoveries arise across the full spectrum of biomedical science. In addition to laboratory discoveries, foundational advances may include innovative behavioral interventions, healthcare delivery models, implementation strategies, digital health approaches, and new conceptual frameworks for improving health.

Potential indicators: Development of novel scientific methods, demonstration of feasibility for innovative concepts, subsequent external funding, and evidence that new approaches influence future research directions.

Implementation: NIH should continue to support high-risk, high-reward research while recognizing innovation across multiple scientific disciplines.

Potential unintended consequences: Evaluation should recognize that transformative discoveries often require substantial time before their ultimate impact becomes evident.

7. Public Impact: Scientific impact should be measured not only by what is discovered, but also by what changes because of the discovery. Publications and citations remain important measures of scientific productivity, but they should be complemented by indicators that reflect improvements in health, healthcare delivery, and public health.

Research should be recognized for improving physical health, mental health, quality of life, functional status, symptom management, healthcare delivery, and long-term outcomes. Mental health and behavioral outcomes are fundamental components of health and should be valued alongside traditional biomedical outcomes.

Cancer survivorship illustrates this opportunity. Advances in prevention, detection, and treatment have created millions of cancer survivors living with the long-term physical, psychological, social, and financial effects of cancer and its treatment. Research that improves survivorship care, symptom management, care coordination, and quality of life translates scientific discovery into meaningful improvements in people's lives.

Potential indicators: Improvements in patient-reported outcomes, implementation into clinical practice, incorporation into professional guidelines, reductions in avoidable healthcare utilization, and measurable improvements in population health.

Implementation: These measures could be incorporated into NIH progress reports and final reports using standardized definitions appropriate to different research disciplines.

Potential unintended consequences: Because opportunities to influence policy or clinical practice vary across disciplines and career stages, these measures should complement rather than replace traditional scientific metrics.

8. Other Areas: ---

Other Areas

Biomedical research increasingly demonstrates that health is shaped not only by biology but also by behavior, families, healthcare systems, and the communities in which people live. NIH's framework for measuring scientific impact should explicitly recognize these dimensions.

Behavioral science, population science, implementation science, and health services research provide the essential bridge between discovery and improved health. These disciplines help determine whether scientific advances are adopted, sustained, and ultimately improve the lives of patients and families.

Family caregivers play a central role in prevention, treatment, survivorship, serious illness, and long-term management of chronic disease. They influence treatment adherence, symptom management, healthcare utilization, and patient well-being, while often experiencing substantial physical, emotional, and financial burdens themselves. Research that improves caregiver health and effectiveness has the potential to improve outcomes for both patients and families and should be recognized as an important contributor to NIH's mission.

Finally, NIH should recognize the value of community-engaged research that is conducted in partnership with the populations it is intended to serve and that disseminates findings back to those communities. Scientific impact is strengthened when discoveries are translated into sustainable improvements in health for individuals, families, healthcare systems, and communities.

119. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 7/31/2026

1. Rigor and Reproducibility: Would like further clarification if qualitative research would follow under the requirements for rigor and reproducibility. Commonly, qualitative research is appropriately used in biomedical and public health research. However, the nature of qualitative research does not easily follow into the same classification as quantitative. Reproducibility may not always be the end goal with qualitative, simply being that qualitative research is extremely context dependent. Qualitative research still relies heavily on rigor, such as the use of triangulation, thick description, and transparent detail on how the researchers own experience can influence the interpretation of the data. With all the being said, would want to stress, that high-quality qualitative research, may not be reproducible, and these standards, in no way, limit high-quality, high-rigorous qualitative research.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: In addition to focus on quantifying mentorship, would also high-stress the importance of gathering qualitative data - focus on the quality of mentoring.

4. Collaboration: Collaboration is important. How will NIH assess international collaboration, especially where there is increasing standards and policies (including from the NIH) to limit international collaboration?

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

120. hamid khosrowshahi

I am responding: On behalf of myself

Name: hamid khosrowshahi

Name of Organization: FloSure Technologies

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 7/31/2026

1. Rigor and Reproducibility: PUBLIC SUBMISSION TO THE NATIONAL INSTITUTES OF HEALTH (NIH)

Title: Restoring Gold Standard Science: Addressing Implementation Failures, Protocol Integrity, and Unvetted Assumptions in Federally Funded Clinical Trials
Target Mechanism: NIH Request for Information (RFI): Measuring Scientific Impact, Research Rigor, and Reproducibility (Notice NOT-OD-26-087)
Respondent: Hamid Khosrowshahi, President, FloSure Technologies LLC | Independent Forensic Medical Researcher & Patient Safety Advocate
Date of Submission: July 31, 2026

1. Executive Summary & Statement of Intent
The National Institutes of Health's initiative to refine how scientific impact, rigor, and reproducibility are measured is essential for safeguarding public trust and ensuring taxpayer funds deliver meaningful clinical outcomes. Traditional trial evaluation metrics often fail to distinguish between intervention futility and implementation failure. When a multi-million-dollar, federally funded trial suffers from execution flaws, unverified physical assumptions, and protocol non-compliance, publishing its conclusions as proof that a technology "does not work" misleads the medical community, distorts clinical practice guidelines, and harms patient care.
This submission presents a forensic case study of the NIH-funded PreVent 2 trial to illustrate how protocol degradation and unmeasured mechanical variables compromise research integrity. It outlines concrete governance reforms to ensure future NIH-funded interdisciplinary and device/interface research meets the highest standards of scientific rigor.
(Case Study Reference: PreVent 2 trial supported by the National Institutes of Health / National Heart, Lung, and Blood Institute under Grant Numbers R61HL138650 and R33HL138650).

2. Case Study in Research Degradation: PreVent 2 vs. Historical Verification Standards
A primary example of federal grant allocation yielding misleading outcomes is the NIH-funded PreVent 2 trial. Designed to evaluate Subglottic Secretion Drainage (SSD) and endotracheal tube (ETT) cuff materials, the trial suffered from fatal execution gaps and ignored established biophysical literature, invalidating its conclusions.

A. Redundant Human Exposure to Known Physical Hazards (The Berra 2004 Precedent)

The trial's protocol relied heavily on Continuous Aspiration Subglottic Suctioning (CASS) without incorporating mandatory verification or pressure-relief controls. It has been well-established in critical care literature for over two decades, specifically demonstrated in the animal model study by Berra et al. (2004), that unmonitored continuous suction creates a "vacuum-lock" against the tracheal mucosa, causing severe tissue trauma, port occlusion, and drainage failure.
Critically, the study's own investigators explicitly acknowledged in earlier 2010 and 2016 publications, including phase-one grant work (PreVent 1 under R61HL138650), that continuous subglottic suctioning CASS carries well-documented risks of severe mucosal injury and that intermittent methods offer safety advantages.
Despite this established evidence, the PreVent 2 trial utilized unmonitored CASS in human subjects under a $10M federal grant, predictably replicating the exact outcome of the 2004 animal model by generating an 83–84% tracheal mucosal injury rate. Using substantial public funds to subject human patients to a known mechanical hazard, only to confirm 20-year-old animal trial data and then misattribute the resulting mucosal trauma and port blockages to "technology futility", represents a profound failure in trial design, grant oversight, and research ethics.

B. The 30-Year Research Regression (1995 vs. 2025)
A forensic comparison between foundational verification studies (e.g., Vallés, Rello, et al., 1995) and PreVent 2 reveals a complete degradation of trial rigor (See Attached Figure: The 30-Year Research Regression):
1. The Verification Gap: Foundational 1995 protocols mandated 4-hour saline flushes to confirm subglottic port permeability. PreVent 2 abandoned active verification, leaving port functional status completely unknown throughout the trial ("zombie ports").
2. The Fluid Volume Blindspot: 1995 protocols recorded daily aspirated fluid volume to track physical fluid dynamics (addressing the 800–1500 mL/day of upper airway secretions documented in human physiology – Guyton & Hall 2020). PreVent 2 failed to measure or document evacuated fluid volume; claiming an evacuation device failed without measuring evacuated volume is a fundamental scientific contradiction.
3. Expanded Monitoring Windows: 1995 protocols required 4-hour cuff pressure checks to maintain seal integrity. PreVent 2 expanded monitoring gaps to 12-hour intervals, creating deliberate structural windows where microaspiration inevitably occurred regardless of cuff material.

C. Editorial Amplification and Guideline Distortions:
The systemic impact of funding methodologically flawed research is amplified when companion editorial commentaries immediately use implementation failures to advocate for clinical abandonment. Key leadership involved in national guideline panels (such as the SHEA/IDSA updates) authored accompanying editorials in The Lancet calling to "pull the plug" on subglottic secretion drainage, using PreVent 2's flawed implementation outcomes as justification.
This creates a closed feedback loop where trials designed with catastrophic execution gaps, such as unverified lines and unmonitored CASS protocol hazards, are immediately cited by editorialists to dismantle mechanical source control guidelines. Prematurely stripping away mechanical prevention measures directly compromises national patient safety initiatives targeting Ventilator-Associated Pneumonia (VAP), secondary Sepsis, Antimicrobial Resistance (AMR), and Post-Intensive Care Syndrome (PICS).

This forensic assessment reflects the established consensus of ventilator-associated pneumonia (VAP) research pioneers: SSD is highly effective when implemented well with active verification, but fails when basic operational controls are stripped away. Measuring poor implementation and misrepresenting it as "intervention futility" misleads the medical community and damages national clinical practice guidelines.

3. Actionable Policy Recommendations for NIH Grant Metrics & Research Oversight
To prevent taxpayer resources from funding trials that generate misleading clinical paradigms, the NIH should adopt four structural criteria for evaluating, funding, and auditing clinical research:

I. Mandatory Physical & Engineering Verification in Trial Design
Clinical trials evaluating physical, mechanical, or anatomical interventions (such as fluid evacuation, airway management, or mechanical ventilation) must undergo independent engineering and biophysical review before grant approval. Trial protocols must account for basic physical dynamics (e.g., fluid accumulation rates, surface tension, and gravity vectors) rather than relying solely on observational or behavioral checklists.

II. Structural Separation of "Implementation Failure" from "Intervention Futility"
The NIH must mandate that multi-center clinical trials institute verifiable, continuous protocol-adherence tracking. If a trial fails to enforce its own intervention protocol (e.g., unverified drainage ports, unmonitored 12-hour pressure gaps, or unmeasured volume outputs), the study outcome must be formally classified and indexed as an implementation failure, prohibiting authors from claiming intervention futility.

III. Formal Audit Mechanisms for Guideline-Impacting Federal Research
When forensic analysis reveals that an NIH-funded trial contained major protocol execution flaws that threaten national standard-of-care guidelines, the NIH Office of Research Integrity (ORI) or Office of Science Policy (OSP) should maintain a rapid-review mechanism. This panel would audit raw compliance data and issue formal scientific clarifications to prevent premature guideline downgrades.

IV. Institutional & Grantee Accountability Controls
To enforce reproducibility, NIH grant agreements should incorporate strict compliance milestones. Principal investigators and institutions that receive multi-million-dollar grants but abandon fundamental verification controls, misreport implementation failures as clinical futility, or refuse to address verified methodological discrepancies should face formal funding audits and restrictions on future federal grant eligibility.

4. Conclusion
Measuring true scientific impact requires looking beyond publication metrics to inspect the underlying mechanical and methodological integrity of funded research. By holding clinical trials accountable to basic physical reality and rigorous implementation standards, the NIH can eliminate the waste of public funds, prevent the premature abandonment of lifesaving mechanical interventions, and ensure that federal research drives true, reproducible progress in patient safety.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

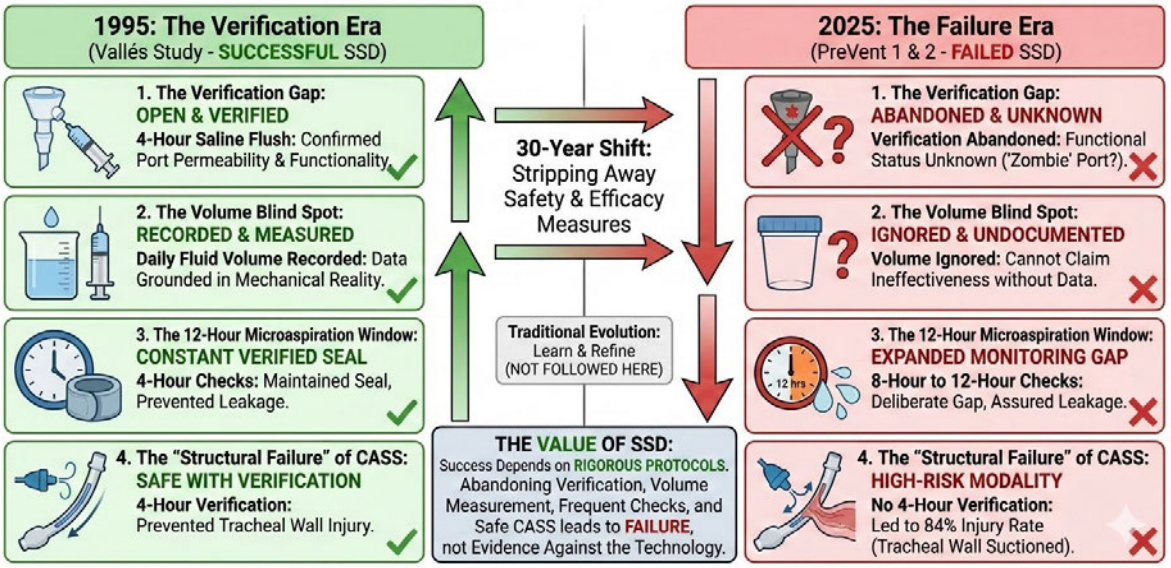
8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/PreVent-2-The-30-Year-Regression-Image.pdf

Description: The 30-Year Research Regression - From Verification to Failure

The 30-Year Research Regression: From Verification to Failure

(Illustrating the Value of Rigorous SSD Protocols)



121. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 8/3/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Enabling researchers to investigate the impact of research on policy decisions is critical and should not be misunderstood as lobbying. Research funded by the NIH/tax payers should be meaningfully shared with decision-makers including policy actors. This ensures that research funded by tax payers is one of many inputs for them to consider. Terminating or deciding to not fund research because it seeks to measure the impact of research on policy decisions would prevent real public impact from being achieved and be a bad return on investment for the public tax payers. An actionable recommendation is to encourage researchers to share their results with different audiences including the public and policy actors, and to not penalize researchers who aim to study the impact of research on policy decisions.

8. Other Areas: N/A

122. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/3/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: The ending of millions of dollars of funding for the research enterprise of the USA does not demonstrate a support of training or any interest in "promoting the growth and stability of the biomedical research workforce." There has been a clear impact of the NIH's current modus operandi on the research workforce. It is overwhelmingly negative with thousands of jobs lost, stark decreases in the number of trainees pursuing research careers, and a mass exodus of scientific talent from to US to countries that are much more transparent in their support of science.

4. Collaboration: N/A

5. Entrepreneurship and Translation: The NIH is not a business or a place to grow profits. The NIH exists to further the health and knowledge of the American people, not to line the pockets of millionaire and billionaire investors. This purpose is stated clearly in its mission statement which reads as follows: "NIH's mission is to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability." Nowhere in that statement is there a mention of entrepreneurship or products. Innovation to further knowledge and to expand the scientific toolset should be rewarded and incentivized, whether or not it results in a new patent or product.

6. Foundational Scientific Exploration: N/A

7. Public Impact: See my comment above on "Entrepreneurship and Translation". Economic growth is not the end goal of scientific research. Additionally, what the Trump administration and Dr. Bhattacharya view as "societal benefit" should not be used as a gold standard. Famously, they both are willing to perpetuate bigotry, disinformation, and hate backed up by research that has been proven to be fabricated, misrepresented, or without basis in scientific fact.

8. Other Areas: N/A

123. Aimee Payne, MD, PhD

I am responding: On behalf of myself

Name: Aimee Payne, MD, PhD

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/3/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: T/F/K awards are essential to support our biomedical research workforce. One issue that is most affecting the success of these efforts currently is the uncertainty of future funding and university cutbacks in many areas of the research mission (eg, fewer graduate students trained, less dollars going to research faculty recruitment in favor of investing in the clinical enterprise, etc). I am hesitant to recommend putting more administrative burden on universities receiving T/F/K awards to prove how well they are doing (beyond what is already required) during a period when institutions are cutting back on hiring due to uncertainty of federal support for research.

4. Collaboration: To a certain extent, recognition of team-based science is already captured through the h-index, which reflects highly cited papers agnostic to where the researcher is in the author list (this can disproportionately reward individuals who participate in reviews and clinical guidelines and not recognize papers where the PI was corresponding or first author - to my knowledge no index measures that important measure of impact). I agree that team-based science is important (which often occurs through P/U mechanisms), but an equally important issue going forward is that we need to ensure that our foundation of investigator-initiated and single PI-based studies remain funded as they are an essential component of developing new knowledge. The US research enterprise became powerful because we trained students and junior faculty through individual and institutional T/F/K awards and then supported the most competitive applicants to pursue new knowledge in the research fields they know best through R/P/U mechanisms, with additional awards such as ARPA-H and DOD funding to support targeted initiatives of national interest. When individuals are strong, team science becomes powerful. If there is not a diverse and deep foundation of expertise across individual faculty, then team science can't achieve its potential.

5. Entrepreneurship and Translation: ARPA-H and SBIR/STTR address this to a certain extent. One challenge is when the call for applications becomes overly proscriptive, as this can run the risk of duplicating what the private sector is already prioritizing and missing out on innovative ideas that need help to advance before they become commercially viable.

6. Foundational Scientific Exploration: 100% agree. There needs to be funding support for faculty who create new knowledge and devote their time and careers to this pursuit.

7. Public Impact: N/A

8. Other Areas: N/A

124. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/3/2026

1. Rigor and Reproducibility: First of all, goals (short-term or long-term) of majority of research is identify targets or develop interventions. If the researcher able to identify targets, develop target-specific interventions and advance from preclinical to commercialization. That is a clear indication of rigor and reproducibility and trying to full fill the promised goals. If the research does not go anywhere except several grants and publications, it is not helping taxpayers or society and simply helping the researchers from friendly peer review both in terms of a greater number of grants and publications. This also short chains another researcher who earnestly trying to contribute to the society. This may be one of the reasons we do have cures for all most all major diseases despite decades and decades public funding for research.

2. Data, Software, and Model Sharing: Data, software and model sharing is useful but not sure how much this would advance. Best way to encourage and reward the research is to reward to those who transitioned basic preclinical bench research to clinical trial and commercialization. This will encourage others follow the same path.

3. Training and Mentorship: Useful for developing next generation scientists. However, majority of scientists main goal is to get more grants and publish more to advance their career and promotion beyond that this type of research goes nowhere.

4. Collaboration: Interdisciplinary collaboration is very helpful to advance in short time.

5. Entrepreneurship and Translation: This really needs to be encouraged, supported and rewarded. This represents rigor, reproducibility and achievement of goals. Discourage researcher obtained grants after grants for 20-30 years with a promise of interventions but never advanced anything except publications and grants with help of friendly review and career advancement in research administration.

6. Foundational Scientific Exploration: This is not helpful in my opinion.

7. Public Impact: Public impact such as health outcomes, economic growth and societal benefit should be the measure and not publications, grants, impact scores or career advancement of the researcher. In fact, researchers above Departmental chair has hardly time to devote for research and funds better utilized on those whose devote whole time on research and not career advancement in research administration.

8. Other Areas: Encourage those try to patent and commercialize their research societal use and reward those who advanced their research through clinical trials. They may develop and advance many more bench to bedside transition through their research if encouraged properly and others will follow.

125. Peter Karp

I am responding: On behalf of myself

Name: Peter Karp

Name of Organization: SRI International

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/3/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: In the bioinformatics field, our primary outputs include online databases and software tools. Both can be characterized and measured to a degree.

For example, to assess the contributions of a researcher who has developed a database, we would ask questions such as what is the
size of the database? What is its scope in terms of the types of data it captures? Does the database occupy a novel niche, and did
the researcher invent that niche, or did they simply maintain a database that was conceptualized by someone earlier? What type
of search mechanisms, data quality improvements, visualizations, and analysis tools does the database provide?

To assess the contributions of a researcher who developed a software tool, we should also consider the size of that
software tool, since software can vary in size by several orders of magnitude. Someone who has authorized a
software tool consisting of a million lines of code has made a contribution that is 100x more significant (along that dimension) than someone
who developed a tool containing ten thousand lines of code. As with the case of a database, the software contribution
should also be assessed based on its originality and the scope of problems that it solves.

3. Training and Mentorship: N/A

4. Collaboration: Perhaps have each of the principal contributors in a team assess the magnitude of the contributions
made by the other contributors, and average the results.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: I have been disappointed in recent years to see that NIGMS has measured investigator productivity based simply

on the number of publications they published. Although publication counting could be an element of a larger
approach to measuring productivity, by itself this approach is much too simplistic because publications vary
tremendously in significance.

7. Public Impact: N/A

8. Other Areas: N/A

126. Jun Yang

I am responding: On behalf of myself

Name: Jun Yang

Name of Organization: St Jude Children's Research Hospital

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/3/2026

1. Rigor and Reproducibility: Rigor and reproducibility are critical to science. However, relying on validation by independent third-party labs is often inefficient and resource-intensive. A far more effective strategy is integrating cross-validation into the active research phase through multi-lab collaborations. For example, NIH initiatives such as MPI grants can require participating groups to independently test and validate each other's key findings prior to publication. This proactive co-validation model offers two major advantages:
(1) Resource Efficiency: It leverages existing grant infrastructure, reagents, and expertise across collaborating sites rather than requiring a third party to rebuild a project from scratch. (2) Proactive Rigor: It identifies potential discrepancies, batch effects, or protocol sensitivities before data reaches the public domain, directly curbing the reproducibility crisis. By embedding internal replication into the research design itself, funding agencies can maximize the impact of research dollars while significantly raising the standard of scientific reliability.

2. Data, Software, and Model Sharing: To accelerate scientific discovery and enhance data integrity, NIH could introduce a payline bonus for open science. Under this model, PIs and laboratories that consistently practice active, transparent sharing including raw datasets, custom software, analytical code, and research models would qualify for a favorable adjustment to their grant paylines (e.g., a percent-line boost or a dedicated score discount during review).

3. Training and Mentorship: Evaluating mentorship and training productivity presents a challenge. Institutions vary wildly in size: some have a low turnover of graduate students, while others struggle to attract top postdoc talent due to location or resources. To create a fair, standardized metric for grant evaluation, NIH could track the publication productivity and fellowship grant awards of the specific trainees funded by the grant. Specifically, NIH can assess how many peer-reviewed papers are published and/or grants are obtained by the postdocs and graduate students explicitly listed in the grant's budget justification.

4. Collaboration: Scientific collaboration should be defined by intellectual and operational contributions, not resource gatekeeping. When tools, mouse models, software, or reagents are developed using NIH grant funding, they are public assets that should be shared openly with the research community. Simply providing access to an existing grant-funded resource does not inherently constitute co-authorship or formal collaboration.

5. Entrepreneurship and Translation: NA

6. Foundational Scientific Exploration: N/A

7. Public Impact: Such impact may take a long term to manifest. NIH could use AI to predict the impact.

8. Other Areas: N/A

127. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/3/2026

1. Rigor and Reproducibility: No comment

2. Data, Software, and Model Sharing: No comment

3. Training and Mentorship: No comment

4. Collaboration: No comment

5. Entrepreneurship and Translation: No comment

6. Foundational Scientific Exploration: I think it is time to seriously asking the question: why the fundamental breakthroughs like: 1) PCR; 2) AlphaFold; 3) Large Language Model/AI are not made by American academia.

7. Public Impact: No comments

8. Other Areas: No comments

128. Mikail Rubinov

I am responding: On behalf of myself

Name: Mikail Rubinov

Name of Organization: Vanderbilt University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/4/2026

1. Rigor and Reproducibility: • Many scientific results are exciting, replicable, and accepted in the literature, but are also redundant with existing knowledge, in a way that is often hard to discern. The accumulation of these results in the literature creates a false impression of progress even as it actually indicates stagnation (and leads to wasted time, money, and effort).
• This problem is especially important in basic biomedical research where rapid advances in methods, data, and analyses can coexist with vast, scattered, and implicit bodies of existing knowledge. See this article for an illustration of the scope and importance of this problem in systems and network neuroscience: https://elifesciences.org/articles/79559
• The NIH needs to incentivize the development of new rigorous statistical tests of proposed novelty against established existing knowledge. The ideal deliverables of such incentives would be new hypothesis tests for basic research fields that match the highest standards of evidence of other fields (most notably randomized controlled trials).

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

129. Richard Kahn

I am responding: On behalf of myself

Name: Richard Kahn

Name of Organization: Emory University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/4/2026

1. Rigor and Reproducibility: I'm very concerned about the dramatic increase in predatory journals or those with weak review/acceptance standards. With the increase in AI tools to scrape the literature we are headed for a huge amount of misleading or incorrect data and interpretations that will slow and misdirect efforts and funding. Tools used for assessing productivity and impact must establish standards for the journals used in all analyses

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

130. Professor

I am responding: On behalf of myself

Name: Professor

Name of Organization: Major university

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/4/2026

1. Rigor and Reproducibility: Unfortunately, this is being approached by the NIH/NHLBI from a small game perspective, e.g. high enough Ns, good stats etc. This does not address the problem of publication of highly ambitious papers that no one even tries to reproduce because of their specialized technical nature. Much of this is being driven by for-profit journals such as Nature and Cell press. It is much less of a problem in lower impact but highly respected journals. The NIH can most improve their analysis by mandating service on NIH study sections. The lack of expertise is very noticeable and influences peer review in a very negative way.

2. Data, Software, and Model Sharing: This seems like less of a problem in my mind. The present arrangements mostly work. Again, the major issues are with the journals.

3. Training and Mentorship: There is irony here. The NIH seems to be favoring the funding of younger investigators to an extent that their natural mentors will need to retire early and thus they create the vacuum they are most concerned about.

4. Collaboration: N/A

5. Entrepreneurship and Translation: This is over-emphasized by the NIH and NHLBI. The SBIR/STTR is quite effective, but the main role of the institute should be to advance basic science knowledge. The free market should and does ensure timely translation.

6. Foundational Scientific Exploration: The concept of "high risk high reward" usually means high speculation little data in the real world. I think that most exciting ideas are funded in a reasonable time frame. Throwing money at promises that mostly cannot be kept is not the answer here. Instead, the institute should not lose sight of the importance of funding basic science research.

7. Public Impact: When the head of HHS and NIH are actively pursuing strategies known to negatively impact the health of American citizens, it is pointless to address this question.

8. Other Areas: N/A

131. Lawrence P. Reynolds, PhD

I am responding: On behalf of myself

Name: Lawrence P. Reynolds, PhD

Name of Organization: Center for Nutrition and Pregnancy, North Dakota State University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/4/2026

1. Rigor and Reproducibility: As far as I am concerned, this focus on reproducibility is a bit of a boondoggle (also know as a "pipe dream")!!! Two observations during my 50 years as a "basic" scientist: (1) Results are often not strictly reproducible. This in fact 'drives' science and is in fact very important; i.e., if something is not reproducible, that drives investigators to investigate why! (2) Observation #1 no doubt is highly dependent on the number of observations! This means that funding agencies need to recognize the importance of non-reproducible results, as well as negative results, and provide the funding sources to replicate the results and determine why they are providing what on the surface seem to be conflicting results!!!

2. Data, Software, and Model Sharing: This effort, again, is in my opinion a bit of a boondoggle. This already is a component of every grant I submit to review, and I have served, since 1986, on 75 federal grant-review panels for NIH (61, chair of 7), USDA (10, chair of 2), NSF (1, ad hoc reviewer), and the Fulbright Scholar Program (3). I have also reviewed grants for federal programs in Australia, Brazil, Canada, France, Germany, Holland, Hong Kong, Israel, Italy, the U.K., Norway, Poland, Spain, and Switzerland.

3. Training and Mentorship: As with any assessment of research quality and impact, this can and should be reflected by publications, grants, and positions attained by mentees, as well as the participation of mentees in their profession; e.g., grants received, collaborations, invited talks, etc. This already is done and evaluated for training grants, so, again, a bit of a boondoggle!!!

4. Collaboration: Other than joint publications and lists of collaborators, and perhaps letters of support from current and past collaborators, I can see no other "approaches" for "assessing and recognizing contributions to team-based science." Again, this "seeking" is a boondoggle and makes me think the "seekers" have had little experience in actually doing science!!!

5. Entrepreneurship and Translation: The only thing I can see in terms of "More [that] can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission" is to provide more support for grant programs designed to promote "translation of discoveries into products for people!!!"

6. Foundational Scientific Exploration: In my view, the only thing that will "Further [embed] these approaches within the academic enterprise" is to provide additional funding and programs to do so. Again, another boondoggle.

7. Public Impact: This is the only topic that is not a boondoggle. Additional support, including grant programs and funding, for "assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit" is certainly warranted. In fact, based on this NIH effort on which I am commenting, even the NIH leadership has little understanding of how science is done, how it progresses, and the needs for advancing science, which, to me, from a federal standpoint is very well in-hand for many decades now, but simply needs more monetary and public (meaning from those connected to the public, like politicians for example) support!!! As I often tell my students and mentees, "This is not magic - the bottom line is the need for support, especially monetary - see places like the EU for example!!!

8. Other Areas: Glad to discuss this with NIH officials at any time!!!

132. Danielle Tatum

I am responding: On behalf of myself

Name: Danielle Tatum

Name of Organization: Tulane University School of Medicine

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/4/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: Rubrics commonly utilized for promotion metrics on the research track in academic medicine remain focused on metrics tailored to individual investigators and do little to acknowledge or reward significant contributions to team science, particularly if the contributor was not named first or senior author on any resultant publication. This essentially penalizes collaboration by not viewing it as an important and valuable contribution towards science and towards the growth of the investigator's research program and institution. A byproduct of this is that investigators know that having their name anywhere other than first or last is not seen as significant intellectual contribution, which can be misleading and paint an inaccurate picture of an individual's important contributions, and as such, may be hesitant to engage with groups where they are not the lead. This is harmful for individual career trajectories, talent retention, and addressing the more challenging questions that our research is trying to address.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

133. Erin Mead-Morse

I am responding: On behalf of myself

Name: Erin Mead-Morse

Name of Organization: UConn Health

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/5/2026

1. Rigor and Reproducibility: Although the number of times an article is cited by another publication is important, understanding the impact of our research beyond this traditional measure is equally important to provide a total view of impact as well as rich context for that citation. However, to accomplish this, the NIH needs to provide resources for personnel to implement, monitor, and address these alternative metrics. There needs to be an additional budget item, beyond the budget max for the research, for this effort. Research dollars are already maxed out on many projects; we are often asked to do more with less. We will not be able to effectively have a broader impact and track that broader impact without additional funds, even small amounts like $50k per year.

Suggested indicators:
- # times researchers post their scientific findings on a public facing site or social media platform, and then the number of times that post is shared, mentioned, commented on, and forwarded.
- # times article is posted by someone else, and then the number of times that post is shared, mentioned, commented on, and forwarded.
- # times study, article, or researcher is cited in a news media piece or interviewed for the news media

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: '- # papers a researcher publishes with a student
- where student collaborators are employed after graduation (government, corporate, NGO, etc.)
- student evaluations of researcher mentorship: quantitative Likert scale ratings of satisfaction, amount learned, preparation for the workforce

4. Collaboration: '- # partners from different sectors who collaborate on a project, including those from the community, NGO, etc.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: '- # testimonials researcher gives to legislatures when considering a piece of legislation
- # times article is cited in policy document or piece of legislation
- research submissions to public dockets for regulatory rule-making

8. Other Areas: Although the number of times an article is cited by another publication is important, understanding the impact of our research beyond this traditional measure is equally important to provide a total view of impact as well as rich context for that citation. However, to accomplish this, the NIH needs to provide resources for personnel to implement, monitor, and address these alternative metrics. There needs to be an additional budget item, beyond the budget max for the research, for this effort. Research dollars are already maxed out on many projects; we are often asked to do more with less. We will not be able to effectively have a broader impact and track that broader impact without additional funds, even small amounts like $50k per year.

134. Jean Liew

I am responding: On behalf of myself

Name: Jean Liew

Name of Organization: Boston University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/5/2026

1. Rigor and Reproducibility: Measurable indicators should include, but are not limited to:
Independent replication of findings in external datasets or cohorts. Investigators should be encouraged to pool resources. Collaboration is key to avoid findings from single centers that are not reproducible.
Public sharing of analytic code and statistical protocols to allow transparency of what was done, particularly for clinical research.
Publication of well-conducted negative or confirmatory studies. Less emphasis on publication of positive studies, which further leads to publicationi bias.
Adherence to reporting guidelines (e.g., CONSORT, STROBE, PRISMA).
Correction of errors when identified, including retractions of publications if necessary.

Rewarding reproducibility would encourage researchers to prioritize robust science over novel but irreproducible findings.

However, NIH should avoid creating burdensome documentation requirements that divert investigators' time from conducting research. The change from the NIH biosketch template to requiring ScienCV formatted biosketches was one example of such a requirement.

2. Data, Software, and Model Sharing: Investigators should be able to provide persistent links (DOIs, GitHub repositories, data repositories) to allow for public sharing of analytic code and statistical protocols to allow transparency of what was done, particularly for clinical research. Shared resources often generate far greater scientific impact than a single publication because they enable future discoveries by many investigators.

NIH should support collaborative data access across institutions. Siloed research is not the most rigorous or reproducible research.

3. Training and Mentorship: Training and mentorship should be recognized as core measures of scientific impact because one of NIH's most enduring contributions is the development of the next generation of biomedical researchers. The impact of a scientist extends beyond their own publications and discoveries to the people they train, mentor, and sponsor, who in turn advance science throughout their careers.

As an NIH-funded early-career physician-scientist, my perspective on mentorship is informed by my own lived experience. I was the child of immigrants and a first-generation college graduate, and I have

experienced firsthand how access to opportunity is often shaped by circumstances beyond one's control. I also know the transformative impact that mentors and sponsors can have in opening doors that otherwise might have remained closed. These experiences have shaped my approach to academic medicine and reinforced my belief that current faculty/research PIs have a responsibility not only to train future scientists, but also to recognize potential, invest in others, and create opportunities for them to succeed.

My own mentorship has centered on clinical research, where I have worked with medical students, residents, fellows, and junior faculty. Through projects supported by foundation awards and my NIH K23 award, trainees have presented research at regional and national scientific meetings, published peer-reviewed manuscripts, and continued into successful academic careers. Equally important has been serving as a sponsor by introducing trainees to collaborators, encouraging them to pursue ambitious opportunities, and helping them build the professional networks that sustain long-term scientific careers. It has been especially meaningful to mentor trainees who, like me, are first-generation college graduates, international medical graduates, or individuals from backgrounds that have historically faced barriers to careers in medicine and biomedical research.

NIH should therefore evaluate mentorship using meaningful outcomes rather than simple counts of trainees. Examples include evidence that trainees have developed independent research skills, presented their work at scientific meetings, published peer-reviewed research, obtained research funding, or advanced into careers that contribute to biomedical science and health care. Evaluations should also recognize sponsorship, interdisciplinary mentorship, and sustained commitment to trainee development.

At the same time, expectations should be appropriate for career stage. Early-career investigators have had fewer years to establish mentoring relationships, and therefore should not be expected to demonstrate decades of trainee outcomes. Instead, evaluation should focus on evidence of thoughtful mentorship practices and meaningful contributions to trainee development relative to opportunity.

Recognizing mentorship as a component of scientific impact would reinforce NIH's long-standing investment in building a strong, innovative, and sustainable biomedical research workforce. The discoveries that improve health tomorrow will depend not only on today's research, but also on today's trainees becoming tomorrow's scientific leaders.

4. Collaboration: NIH should ask investigators to describe their intellectual contributions to team science. Modern biomedical research increasingly depends on collaboration across institutions, disciplines, and countries. These collaborations improve scientific rigor by testing whether findings are reproducible in different populations. Important to the idea of multidisciplinary collaboration is INTERNATIONAL collaboration. For example, I recently completed a study with well regarded investigators in my field who are based in Europe. Each group analyzed its own patient cohort using harmonized methods, and we compared results across populations. Neither institution transferred patient-level data, as per the legal data use agreements at each site, but together we produced substantially stronger evidence than either group could have generated independently. Evaluations should recognize this type of collaboration as a marker of scientific impact. We should not be discouraged from international collaboration, if the expertise of tour collaborators is warranted. The NIH

should avoid paperwork that puts up barriers to international collaboration. Such scientific collaboration is fundamental to rigorous biomedical research and should be rewarded rather than discouraged.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Foundational science often has delayed impact and may not immediately generate high citation counts or clinical applications. Supporting foundational research preserves the pipeline of discoveries that ultimately lead to new diagnostics and therapies. Careers and downstream research are also being supported. Short-term productivity metrics should not disadvantage high-risk, high-reward science whose greatest impact may emerge years later. It is also important to recognize that sudden terminations or delays in awards causes downstream effects that last over years and potentially decades.

7. Public Impact: The ultimate measure of NIH-supported research is its ability to improve the health and well-being of the public. Accordingly, public health impact should be evaluated based on measurable improvements in health, evidence-based communication, and the translation of scientific discoveries into policies, clinical practice, and interventions that benefit patients and communities.

Improving the health of the public means improving the health of ALL people. Achieving that goal requires recognizing that disease burden, access to care, and health outcomes are not distributed equally across the population. Many communities experience disproportionately high rates of disease because of longstanding differences in access to healthcare, preventive services, environmental exposures, socioeconomic conditions, and other social determinants of health. Research that identifies these disparities, seeks to understand their underlying causes, and develops interventions to reduce them is ESSENTIAL to NIH's mission of improving health for the entire nation.

Scientific impact should therefore include research that advances health equity by ensuring that discoveries are relevant to the populations most affected by disease. This includes conducting studies in populations that have historically been underrepresented in biomedical research, developing interventions that are effective across diverse communities, and generating evidence that helps clinicians and policymakers reduce avoidable differences in health outcomes. Focusing on populations that face the greatest barriers to health is not at odds with improving health for everyone; rather, it is often the most effective way to improve the health of the population as a whole.

Public impact should also recognize investigators who communicate science responsibly to clinicians, patients, policymakers, and the public. Accurate, accessible communication is essential for translating research into practice, maintaining public trust in science, and ensuring that scientific discoveries ultimately improve health. Measures of public impact should emphasize evidence-based educational activities, contributions to clinical guidelines and public health recommendations, development of patient and clinician educational resources, and demonstrated improvements in healthcare delivery or health outcomes. Popularity on social media or media attention alone should not be considered evidence of scientific impact.

Finally, NIH should continue to value a biomedical research workforce that brings diverse perspectives and experiences to scientific discovery. Diversity in backgrounds, disciplines, life experiences, and scientific approaches strengthens research by fostering innovation, challenging assumptions, broadening

the questions that are asked, and improving the relevance of research to the populations NIH serves. Evaluation of public health impact should recognize investigators who build inclusive research teams, engage meaningfully with the communities affected by their research, and produce science that is applicable across the diversity of the U.S. population.

8. Other Areas: As an NIH-funded early-career physician-scientist, I encourage NIH to recognize that scientific impact is created not only by individual discoveries, but also by the activities that make those discoveries rigorous, reproducible, and useful. Collaboration, dissemination through peer-reviewed publications and scientific meetings, sharing of data and analytic tools, and mentorship should be viewed as central components of scientific impact rather than secondary activities. The United States has become a global leader in biomedical research because it has fostered openness, collaboration, and scientific exchange. This necessarily includes international collaborations, as well as supporting researchers' ability to attend conferences to further collaborate and disseminate results, and support researchers in publishing their findings in peer reviewed journals. Outside of research funding, we do not have alternative means to support our attendance at conferences or publications in journals that charge publication fees. Removing our ability to include such items in grant budgets does a huge disservice not just to us but to scientific research in this country overall. NIH's evaluation framework should reinforce these strengths while avoiding unnecessary administrative burdens that reduce the time investigators can devote to conducting high-quality research.

135. Erick Kiprotich Yegon, PhD

I am responding: On behalf of myself

Name: Erick Kiprotich Yegon, PhD

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 8/5/2026

1. Rigor and Reproducibility: Current and proposed indicators of rigor share a common weakness: they measure whether a practice occurred, not whether it constrained anything. Pre-registration is the clearest example. It is easy to count and easy to satisfy without a single analytic decision being bound by it. If NIH adopts registration as an indicator, the predictable unintended consequence is decorative registration — plans written loosely enough to accommodate any result, filed to satisfy a checkbox.

I would suggest three indicators that are equally countable but substantially harder to game.

1. Presence and content of a numbered deviation log. A registered plan that produced no deviations is either trivially simple or not being read honestly. What distinguishes bound work is a public, itemised record of where the analysis departed from the plan and why. This is machine-checkable: does a deviation log exist, is it numbered, does each entry state the registered specification, the change, and the rationale?

2. Whether the registered primary outcome is reported regardless of direction. The indicator is the ratio of registered analyses to reported analyses, and whether the registered primary endpoint appears in the abstract when it returns a null. A registration that produced a null primary result and a paper that leads with a secondary finding is a detectable pattern.

3. Advance declaration that all outcome branches are publishable. In a recent pre-registered analysis of county-level Medicare fee-for-service preventable hospitalizations across 3,061 counties (OSF registration DOI 10.17605/OSF.IO/9U4AE), I declared in the registration that all four possible outcome branches — material, mechanical, robust/null, and ambiguous — were publishable before any model was fit. This costs nothing to state, is verifiable against the frozen registration, and removes the author's own exit route from a null.

On unintended consequences, I want to name one from direct experience. That study's registered primary contrast returned a null. Its most substantive finding was unregistered and unfavourable to my own work: three of the four highest-ranked predictors partially re-measured the outcome, because inpatient-setting encounters were counted inside the outcome definition — a fact documented only in an appendix of the source methodology, not in any variable label or data dictionary. Reporting it reduced spatially blocked model performance from R^2 0.321 to 0.236 and reordered the attribution ranking more

than the registered question had. Disclosing this made the paper harder to place, not easier. Under any indicator that counts registrations, deposits, or publications, that study is indistinguishable from one where the registration changed nothing. Self-correction is currently paid for entirely by the researcher who performs it, and no existing metric can see it. If NIH wants a single indicator worth building, it is this one.

2. Data, Software, and Model Sharing: The current standard is availability: data are deposited, code is posted, a statement appears in the manuscript. Availability is necessary but it is a weak indicator, because it is satisfied by deposits that no one can execute and that no one has checked against the paper's claims. I would propose shifting the measured quantity from availability to verifiability, and offer three concrete, implementable indicators.

1. Proportion of numeric claims in the manuscript machine-checked against the deposited data. In the study referenced above I deposited two executable verification programs alongside the dataset: one that recomputes every shared figure in the analysis reports directly from the data, and one that parses the report text and checks every numeric claim in the prose against the underlying dataset, flagging any mismatch. Both passed with no mismatches. This is not exotic tooling — it is a few hundred lines of Python — but it converts "the data are available" into "the claims in this paper have been mechanically checked against the data." The indicator is the fraction of reported numbers under automated verification, which a repository can compute.

2. Cryptographic binding of the dataset and feature set at registration. My registration records two SHA-256 hashes, one fixing the analytic dataset and one fixing the feature list, both checkable against the deposited files. This makes silent substitution of the data between registration and publication detectable by anyone, at essentially zero cost. Repositories could compute and display this automatically.

3. Machine-readable derivation provenance for every variable, not just labels. The contamination described above was undetectable from variable names and would have remained undetectable from any ordinary data dictionary. It was visible only in an appendix describing how the outcome measure was constructed. The measurable requirement is that shared datasets carry derivation formulas and source-code composition at a level sufficient to detect overlap between a feature and the outcome. As secondary analysis of administrative and claims data becomes more common, and as machine learning makes feature counts larger, this failure mode will become more frequent, not less — and post-hoc explanation methods will faithfully rank the contaminated variables at the top.

On unintended consequences: any indicator based on counting deposits or download volumes will reward the deposit of many low-value datasets and disadvantage fields where a single well-documented dataset represents years of work. Weighting by verification and reuse, rather than by count, avoids this.

3. Training and Mentorship: Existing indicators count trainees. A more informative measure is whether trainees can independently reproduce work they did not perform. I taught graduate research methods as an adjunct for twelve years and currently supervise doctoral research. In my experience the single most transferable skill is not statistical technique but the discipline of pre-specifying an analysis and reporting deviations from it — and it is almost never assessed. A measurable indicator: the proportion of trainees who have registered an analysis plan and published a deviation log before completing their degree.

4. Collaboration: Team science is increasingly multinational, and NIH-funded work frequently depends on investigators and institutions in the countries where data are collected. Credit for those contributions is systematically under-recorded, and it is measurable.

I have directed research teams across six countries and worked with more than fifty partner organisations. In that setting the recurring pattern is that in-country investigators design instruments, negotiate government access, secure ethics approval, and collect and clean the data, then appear in acknowledgments rather than authorship, or in middle positions that no metric registers.

Three implementable indicators:

Authorship distribution relative to data origin. For studies using data collected outside the US, the proportion of first and senior authorships held by investigators based in the country of data collection. This is computable from existing metadata and is already tracked in some bibliometric work.
Completeness and machine-readability of CRediT statements. CRediT exists but is applied inconsistently and rarely audited. The indicator is whether every listed author has at least one assigned role, and whether the statement is deposited in structured form rather than as prose.
Whether data-collection partners appear as authors at all. A binary, checkable field.

On unintended consequences: any indicator that counts collaborators, institutions, or co-authors will reward gift authorship and large consortia over genuine contribution. Distribution and role-completeness are harder to inflate than headcount.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Publication and citation counts are poor proxies for whether findings changed anything, and the most consequential outcomes of applied health research are currently uncounted.

Two examples from my own work. Findings from a task-shifting evaluation of community health workers in Burkina Faso were adopted by the national government and incorporated into programme policy. Evidence from an adolescent sexual and reproductive health programme in Kenya contributed to revision of the national adolescent reproductive health policy. Neither appears in any citation-based metric, and neither would register under any indicator NIH currently uses.

Suggested indicators:

Citations in policy documents, clinical guidelines, and government technical reports, tracked with the same infrastructure now used for scholarly citations. The technical capability already exists commercially; what is missing is that it does not count toward anything.
Time from publication to first documented guideline or policy incorporation. This measures whether the research reached the people who act on it, and it rewards accessibility and clear reporting rather than novelty.
Explicit inclusion of non-US and low-resource settings. NIH-funded evidence is frequently applied in health systems that produce no indexed literature of their own, and any tracking built only on US and

European policy corpora will systematically under-count it.

On unintended consequences: altmetric and media-mention indicators are already available and should be resisted as impact measures. They reward controversy and volume, and they are trivially manipulable in ways that policy citation is not.

8. Other Areas: The RFI asks for considerations of feasibility across career stages, disciplines, and institution types. I would ask NIH to consider a group the phrase "institution types" does not quite reach: researchers with no institution at all.

I am an independent researcher. The pre-registered study described in my earlier responses was designed, built, analysed, and written without institutional affiliation, library access, departmental support, or funding. Every open-science practice NIH wishes to reward — registration, data deposit, verification tooling, transparent reporting of a null — was available to me at no cost, and I used all of them. That is worth stating plainly: open-science indicators are more feasible for unaffiliated researchers than almost any other marker of scientific standing.

The obstacles sit elsewhere, and they will distort whatever indicators NIH builds:

Article processing charges are borne personally. An indicator rewarding open-access publication penalises exactly the researchers for whom open science is otherwise cheapest.
Verification and replication work earns nothing. There is no mechanism by which someone who checks another group's deposited data and finds an error receives credit. An ORCID-linked, citable record of verification contributions would be a specific and implementable start.
Registration infrastructure is open to anyone; institutional review is not. Independent researchers working with public secondary data can register and deposit but often cannot access an IRB. Any indicator conditioning on institutional review excludes this population by construction.

This is not only a fairness argument. Replication and verification are precisely the activities that require no institutional resources, and they are going undone in a system where they earn nothing. A population capable of performing them exists and is currently invisible to every metric NIH uses.

136. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/5/2026

1. Rigor and Reproducibility: I agree that scientific research needs to be rigorous. However, the definition of what constitutes rigor is not clear. Clarifying that definition will be critical to utilizing it as any sort of standard to compare research practice against. Reproducibility of research findings has always been a fundamental element of the scientific method. The goal of any Methods section of published work is provide enough detail that another scientist could replicate the approach and check their findings against others using the same approach. The current model of 'publish or perish' in the tenure-track system within universities does put a damper on publishing replication studies, as these are often not reviewed well when submitted for publication. However, how many replication studies are necessary for any given finding? Is 3 enough? Is 1,000 enough? How much research funding needs to go toward reproducibility before it is considered a waste of taxpayer money? And how closely does a replication study's finding need to be to the original finding to be considered 'reproduced'? Further, in the event where many groups publish findings on a similar topic, but use different methods, how does one determine which study is the original against which the others are compared? And who determines what the 'gold standard' is in any field of study? These are the practical challenges to implementing the ideals which this initiative seems to promote.

2. Data, Software, and Model Sharing: Data, software and model sharing are great ideals. Many areas of scientific research have been pushing these ideals for the past 20 years or more. However, there is currently not incentive to do so, besides recognition by colleagues that some ideal has been upheld. Instead, there is a competing ideal which our society seems to favor; that is the ideal that keeping things proprietary or protecting them with IP makes them more valuable, which can then financially incentivize researchers in the event that they may be able to license their data, software or models to industry, or start their own company to utilize the approach for financial gain. It seems that our society values the proprietary for-profit model more than it does the free sharing of knowledge. For ideal of sharing to be incentivized, one would need some certainty of financial benefit to offset the incentive of protecting ideas for financial benefit.

3. Training and Mentorship: Though there are many great NIH policies and programs that support training and protect trainee time for workforce development, the barrier to entry is quite large. Successful acquisition of fellowship funding through the existing F30/31/32 models is seemingly still based on reputation of PI or Institute rather than inherent quality and potential for growth of the trainee. On a slightly different topic, the changes made over the past 18 months to immigration and visa

applications has greatly reduced the number of high quality trainees who are able to join the scientific training that our country has to offer, thus reducing the future research workforce.

4. Collaboration: Recognition of team-based research needs to be implemented at the institution level so that researchers are rewarded for such collaboration in the form of high confidence of securing tenure and promotion. Additionally, the NIH needs to provide well-funded opportunities that are specifically targeted to researchers who are engaged in multi-institutional research, or who would like to. To often, the will and the interest is strong between researchers to collaborate, but the mechanisms to support the work does not exist or is not sufficient to carry out the work. This disincentives researchers from this ideal, causing them to pursue smaller, single-PI funding mechanisms.

5. Entrepreneurship and Translation: This ideal would seem to be contrary to the one listed previously regarding "Data, Software, and Model Sharing". The SBIR/STTR program is already a great program. Investing more resources into that program would be great! Additionally, I think it is necessary to provide even earlier pilot funding to researchers who are developing their ideas for translation to market but haven't yet formed a business.

6. Foundational Scientific Exploration: Agreed. There should be a well-supported mechanism for high-risk/high-reward proposal. Over the years, the R21 mechanism has become less likely to be funded, as it seems the reviewers tended to review these applications as if they were R01s, especially in the expectation that there would be strong preliminary findings. I think this reviewer bias was natural, because these have typically been reviewed in the same panels as R03 and R01 submissions. In an entire review panel's agenda, a minority of proposals would be R21s, and they seemed to not get sufficient focus. I would recommend that the review of R21 proposals (or similar high-risk mechanism) should occur in it's own focused panel, with very explicit instructions given to reviewers to not treat the proposals as if they were R01s.

7. Public Impact: I agree that there should be more benefit given to research that can demonstrate translation into real-world public impact. NIH has not been known as the federal agency which funds applied translational work, but rather has been generally known as the home for basic science discovery research. My own work has been easier to fund with industry-sponsored research agreements or with donor support than with NIH support, because my work is in the realm of human health benefiting from sport.

8. Other Areas: If the NIH wants to continue to be seen as the world-leader in biomedical scientific research, it is critical that it's reputation as a purely scientific entity be maintained. Because the NIH exists as a government-funded entity, there has always been risk that it's decisions will be overly influenced by the non-scientific perspectives of government officials or elected politicians. Over the last several months, announcements regarding executive orders or NIH policy statements have seemingly indicated that political appointees would be in positions to approve or deny funding for grant proposals which the non-government scientific experts reviewed favorably. These announcements have also seemed to suggest that previously awarded scientific projects could have their funding rejected by political appointees because they no longer match the current administration's priorities. This would utterly destroy the NIH's reputation as a scientific entity. There are historical examples in other countries of similar politicization of science completely ruining the otherwise strong research community. Perhaps the most poignant example of this was during the 1930's in the Soviet Union when political appointees destroyed the use of science in that country. One exceptional case was Trofim Lysenko, who rejected the

widely accepted principles of Mendelian genetics and drove agricultural decisions based on political ideals, rather than science. The loss of scientific principles in the Soviet Union led to the country falling several decades behind the rest of the world. The NIH needs to carefully manage its position not only within the government, but within the apolitical scientific community.

137. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/5/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: Data are very siloed. Creating an easily searchable database of all large datasets generated through NIH funding would be helpful. And there should be an incentive for particpating.

3. Training and Mentorship: Existing programs are onerous (T32 applications should be streamlined to encourage applications) and underfunded. There should be more support for program directors and less of a salary gap for trainees. Tracking trainees beyond their training program would be important.

4. Collaboration: Institutional promotional committees need to be educated to recognize team science. Increase funding fro program projects and create new mechanisms for collaborative projects.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Basic research has suffered the last few years because of the uncertainty of the funding environment. We are losing promising young researchers so we need to create a stable pathway. Having competitions for specific high-risk high yield programs could help.

7. Public Impact: N/A

8. Other Areas: N/A

138. Doug Brugge

I am responding: On behalf of myself

Name: Doug Brugge

Name of Organization: University of Connecticut School of Medicine

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/6/2026

1. Rigor and Reproducibility: Regarding reproducibility, the best way to incentivize more replication of prior research is to provide funding mechanisms that support such efforts. Currently, the incentives are against replication because NIH favors, quite strongly, innovation and originality. A program announcement that only funds replication would be one possible way to encourage it.

As to rigor, I think the review process at NIH, and I have served on over 45 study sections, including chair one and as a standing member on two, could be improved by incentivizing the most qualified reviewers to serve. As it currently stands, reviewing is largely a thankless job that consumes a lot of time. How about some advantage in the review process for reviewers, or higher compensation or a reward of an R03 after service on 20 panels?

2. Data, Software, and Model Sharing: NA

3. Training and Mentorship: Expand the K award programs to offer more awards?

4. Collaboration: In my field, public health, we collaborate all the time across disciplines already

5. Entrepreneurship and Translation: How about follow up funding to R01 funded research that explicitely focuses on translation into real world impacts?

6. Foundational Scientific Exploration: NA

7. Public Impact: I am confident that in public health, much of our research has real world impact, the challenge is documenting that. I think two things are needed.

One is to develop reliable and standardized metrics by which these impacts can be measured. Scientific impact is easily assessed by publications, citations and reads, but nothing like that exists, as far as I know, for real world impact. So, fund development of metrics.

Secondly, at the end of a research study, there is no funding to document this sort of impact. Either add on funding to do so or requiring that budgets include funding for this purpose, perhaps in the final year of the grant, a bit like requirements for dissemination, might help.

8. Other Areas: N/A

139. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Other

Role - Other: Over my long career I can check off participation in 7 of the proposed roles.

Date Submitted: 8/6/2026

1. Rigor and Reproducibility: This process will not work unless anonymity is provided for those challenging well known researchers or long accepted principles. Or - the replication process is reserved for established researchers who don't need grant support for research. There goes the citation index. On occasions where I examined such issues as bias of perceptual attention, restriction of range of response, bias of focus of response, repeated construct bias, I was published and then characterized as having no original thoughts of my own but only attacking others' research, faking data, using questionable stimulus manipulations, undermining a publishing journal by showing that some alterations in my own methods would produce slightly but not totally different outcomes, showing that a reliable outcome I was predicting and replicating, only required half of participants to show the "reliable direction of outcome" to obtain significant results. And - an extensive review of methods article was accepted for publication by the editor of a well-known methods journal with the codicil of the editor that it could end my career and that of my co-author in the general area of the research. The world of research is much more vicious now than it was when I was starting my career. Now a pretext and the misuse of findings to support that pretext is a common part of science.

2. Data, Software, and Model Sharing: Based on my previous comments what might the incentive be to be abused or misused politically for not supporting a current version of authorized reality.

More importantly perhaps is that the current context of science does not support the time and financial dollars needed to support replication and its nuances.

3. Training and Mentorship: N/A

4. Collaboration: AI can run simulations, e.g., what is the financial impact of limiting highly successful investigators/labs to N grants or N dollars, but AI is easily biased by existing biased information and cannot be trusted.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

140. Sara Kiesler

I am responding: On behalf of myself

Name: Sara Kiesler

Name of Organization: Carnegie Mellon Univ

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/7/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: I recommend using the so-called tire tracks approach, which identifies major discoveries and advances in the use and application of science, and looks backwards at how research enabled these contributions to society, medicine, or business. A major advantage of this method is that it recognizes that research is cumulative and can take many years, and may involve missteps and corrections as well. This method is probably most useful when the research is fast moving, as in computational research or some areas of bioscience. Other sciences are slower moving because they receive little funding, or require human RCTs, or involve processes that enfold over long periods. https://cccblog.org/2016/07/27/the-tire-tracks-diagram-corrected-and-humanized-by-national-academy-workshop-report/

7. Public Impact: See above

8. Other Areas: N/A

141. Imon Banerjee

I am responding: On behalf of myself

Name: Imon Banerjee

Name of Organization: Mayo Clinic

Type of Organization: Healthcare Institution

Role: Investigator/Researcher

Date Submitted: 8/7/2026

1. Rigor and Reproducibility: **Response on Rigor and Reproducibility**

NIH should strengthen policies and infrastructure that make biomedical research reproducible by design rather than as an optional post-publication activity. We recommend the following measurable and feasible approaches:

1. **Open-source code and model release.** Investigators should be encouraged, and when appropriate required, to release source code, model architectures, training configurations, and inference pipelines through publicly accessible repositories. Measurable indicators include the proportion of funded projects with publicly available code, complete environment specifications (e.g., containerized environments), version-controlled repositories, and executable workflows.

2. **Standardized data-sharing mechanisms.** When participant privacy and consent permit, de-identified datasets, benchmark cohorts, or standardized synthetic datasets should accompany publications. For restricted clinical data, NIH should encourage secure data enclaves or federated evaluation platforms. Success metrics could include the percentage of studies providing reusable datasets or approved mechanisms for independent external validation.

3. **Standardized APIs for AI models.** For projects developing machine learning or foundation models, NIH should require or strongly encourage deployment of standardized application programming interfaces (APIs) that enable independent researchers to evaluate models on external datasets without requiring access to proprietary code or sensitive training data. This would facilitate reproducibility while protecting patient privacy and intellectual property. Metrics could include API availability, documentation quality, version control, uptime, and the number of independent validation studies completed.

4. **Reproducibility and sustainability as required closeout criteria.** Every federally funded research project should undergo a reproducibility and sustainability assessment before grant closeout. Similar to financial and scientific reporting requirements, investigators should demonstrate that the project's primary outputs are reproducible through documented code, accessible software, appropriate data-sharing mechanisms, comprehensive documentation, and, where applicable, externally accessible APIs. In addition, projects should include a **maintenance and sustainability plan** describing how software,

APIs, and associated resources will be maintained after the award period, including bug fixes, security updates, dependency management, versioning, and long-term repository hosting. NIH could establish a reproducibility checklist with measurable indicators, such as successful execution of the computational workflow, availability of required artifacts, API functionality, repository accessibility, and evidence that an independent evaluator can reproduce the reported results.

Embedding reproducibility and long-term maintenance into grant closeout would create strong incentives for sustainable research practices, improve transparency, accelerate independent validation, and maximize the long-term value of federally funded research investments. To ensure feasibility across institution types and career stages, NIH should provide dedicated funding mechanisms for software engineering, data curation, cloud infrastructure, API hosting, and long-term maintenance of open-source repositories and research software beyond the initial project period.

2. Data, Software, and Model Sharing: NIH should expand its data-sharing policies to recognize that modern biomedical research depends not only on data, but also on reusable software, computational workflows, and AI models. Effective sharing requires that these resources remain accessible, interoperable, and maintainable beyond the initial funding period.

NIH should establish centralized infrastructure (or regional centers) dedicated to the long-term stewardship of federally funded research software, datasets, models, and APIs. Similar to national data repositories or coordinating centers, this Research Software and AI Sustainability Center would provide repository hosting, API management, cloud infrastructure, security monitoring, dependency updates, version control, user support, and software preservation. Such a center would reduce duplication across projects, ensure continuity after grants end, and maximize the long-term return on federal investments.

By treating data, software, AI models, and their supporting infrastructure as long-term scientific assets, NIH can maximize the return on federal research investments, improve reproducibility, foster collaboration, and accelerate biomedical discovery. A centralized logistics and sustainability center would ensure that valuable federally funded resources remain accessible, functional, and secure for the broader research community long after the original project has concluded.

3. Training and Mentorship: NIH should broaden its evaluation of training and mentorship beyond traditional metrics such as publications and grant funding to recognize activities that build a sustainable, interdisciplinary, and translational biomedical workforce. Training should equip researchers with the skills needed for modern biomedical science, including data science, artificial intelligence, software engineering, reproducible research, team science, entrepreneurship, and regulatory science.

Establish measurable mentorship outcomes. NIH should evaluate training programs using standardized indicators such as trainee retention, career progression, interdisciplinary collaborations, leadership development, transition to independent funding, and contributions to team science. Evaluations should also include trainee satisfaction and long-term career tracking.
Implement long-term assessment of NIH-supported trainees. NIH should establish a centralized system to longitudinally track the career trajectories of trainees supported through NIH awards (e.g., 5, 10, and 15 years after training). Success metrics should extend beyond academic appointments to include

leadership positions, scientific productivity, entrepreneurship, technology commercialization, clinical translation, public service, mentorship of future investigators, and contributions to open science. These longitudinal outcomes would provide a more meaningful assessment of the long-term return on NIH investments in workforce development.
Recognize open science and research software contributions. Training programs should incentivize the development and maintenance of open-source software, datasets, AI models, and reproducible workflows. Contributions to community resources, software documentation, benchmarking, and data curation should be recognized as valuable scholarly outputs alongside publications.
Support experiential and cross-disciplinary training. NIH should encourage programs that integrate computational sciences, engineering, biomedical research, and clinical practice through cross-disciplinary mentoring, industry partnerships, internships, and collaborative projects. Exposure to regulatory science, implementation science, and commercialization will better prepare trainees for translating research into practice.
Provide sustained mentorship incentives. Effective mentorship requires significant time and effort but is often undervalued. NIH should recognize mentorship excellence in funding decisions and provide dedicated support for faculty who consistently demonstrate successful trainee development, interdisciplinary mentoring, and workforce diversity.
Develop centralized training infrastructure. NIH should establish national training and mentorship hubs that provide shared educational resources, standardized curricula, online learning platforms, mentoring networks, career development resources, and access to computational infrastructure. These centers would reduce duplication across institutions, expand opportunities for investigators at resource-limited institutions, and ensure equitable access to high-quality training.
Require workforce development plans and outcomes. Every federally funded training grant—and research grant supporting trainees—should include a workforce development plan with measurable milestones and report outcomes at grant closeout. Metrics could include trainee career advancement, mentorship quality, reproducible research practices, software and data contributions, interdisciplinary collaborations, and community engagement. These grant-level outcomes should also feed into NIH's centralized longitudinal trainee tracking system to continuously evaluate and improve the effectiveness of federal training investments.

By systematically measuring both short-term training outcomes and long-term career impact, NIH can better understand which training models produce the greatest scientific and societal return, enabling evidence-based improvements to workforce development programs while strengthening the future biomedical research enterprise

4. Collaboration: Evaluation should extend beyond first- and senior-author publications to acknowledge leadership in data generation, software development, AI model development, project coordination, clinical implementation, regulatory activities, and shared infrastructure. Standardized contributor taxonomies should be incorporated into grant reporting and career evaluations.

Rather than simply counting collaborators or institutions, NIH should assess measurable outcomes such as jointly funded grants, shared intellectual property, co-developed software and datasets, successful multicenter validation studies, technology adoption, and sustained partnerships beyond a single funding cycle.

Many impactful collaborations require years to mature. NIH should provide mechanisms to sustain successful collaborative networks, shared research infrastructure, and coordinating centers beyond the duration of individual research awards. Long-term support for data platforms, software ecosystems, and collaborative governance will maximize the return on federal investments and reduce duplication across projects.

5. Entrepreneurship and Translation: Track long-term translational outcomes. NIH should establish a centralized system to longitudinally assess the downstream impact of federally funded innovations, including follow-on investment, regulatory approvals, commercialization milestones, healthcare adoption, patient outcomes, startup sustainability, job creation, and economic impact. These data would enable NIH to identify funding mechanisms and translational models that produce the greatest public benefit.

Support translational infrastructure. NIH should expand access to innovation hubs, incubators, regulatory science expertise, manufacturing resources, clinical validation networks, and implementation science programs that help investigators bridge the gap between laboratory discoveries and real-world deployment.

6. Foundational Scientific Exploration: N/A

7. Public Impact: Recognize contributions to public resources. Federally funded projects that produce openly available datasets, software, AI models, educational materials, clinical guidelines, benchmarking resources, and public repositories should receive recognition for creating infrastructure that enables future scientific discovery and broad community benefit.

Evaluate economic and workforce impact. NIH should measure downstream outcomes such as startup creation, technology licensing, job creation, workforce development, commercialization, industry partnerships, and regional economic growth resulting from federally funded research.

8. Other Areas: N/A

142. Sindy KY Tang

I am responding: On behalf of myself

Name: Sindy KY Tang

Name of Organization: Stanford University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/7/2026

1. Rigor and Reproducibility: NIH should expand the concept of rigor and reproducibility to encompass the complete measurement chain from biospecimen collection to reported result. For many cellular and functional assays, particularly those involving living or time-sensitive specimens (e.g., ex vivo cell activation tests), reproducibility is determined not only by experimental design but also by the interval and conditions between collection, processing, stabilization, transport, data acquisition, and analysis. Differences in these pre-analytical steps can substantially alter the measured biological signal and limit reproducibility across laboratories and clinical sites.

A key opportunity is to support technologies and validation frameworks that make time-sensitive biological measurements robust to realistic multicenter research workflows. These could include methods for decentralized specimen processing and stabilization, standardized transport conditions, automated or simplified sample handling, and approaches that separate the time and location of specimen collection from specialized downstream analysis.

Importantly, reproducibility should be evaluated at the individual-sample level, not only through aggregate measures such as cohort means or correlations. Strong population-level correlation can coexist with substantial deviations in individual measurements, including deviations large enough to alter biological or clinical classification. Useful quantitative indicators could therefore include sample-level bias and limits of agreement relative to a reference condition, preservation of classification or decision thresholds, robustness across defined time and temperature conditions, and reproducibility across sites, operators, instruments, and reagent lots.

Analytical reproducibility should be considered part of the same measurement chain. For complex datasets, NIH should encourage methods that preserve a traceable path from raw data through quality control, analysis, and the reported endpoint. This would make it possible to determine whether disagreement across studies arises from biology, specimen handling, instrumentation, batch effects, or analytical decisions.

A useful future NIH priority would therefore be the development and cross-site validation of end-to-end reproducibility platforms for complex biological measurements, with success defined by preservation of sample-level measurement fidelity from specimen collection through final analysis.

2. Data, Software, and Model Sharing: NIH should distinguish between data that are merely available and data that are genuinely re-analysis ready. Large amounts of valuable biomedical data have been generated through NIH-supported studies, but their future value can be limited when differences in instruments, assay configurations, reagent panels, batches, quality-control criteria, analytical methods, or reporting conventions make datasets difficult to compare or re-analyze together.

This challenge is particularly apparent in experimental data such as flow cytometry. Raw data from multiple studies may remain available, yet different laboratories may have used different marker panels, cytometers, population-identification strategies, gating conventions, quality-control rules, and thresholds. As a result, historical datasets that could otherwise support replication, validation, pooled analyses, or new biological questions may be difficult to synthesize objectively.

NIH should incentivize preservation of the raw data and metadata required for future re-analysis, including information about specimen collection and processing, assay configuration, instrumentation, reagents, batch, quality-control rules, and analytical procedures. A useful measure of scientific impact would be whether an independent investigator can reconstruct a reported endpoint from archived primary data and subsequently apply an updated or standardized analytical workflow.

NIH should also support development of computational and AI-enabled frameworks for retrospective cross-study re-analysis and auditing. Such tools could apply explicit analytical protocols across heterogeneous current and archived datasets, accommodate differences in measurement configurations, identify batch- or analysis-associated discrepancies, compare standardized analyses with historical manual analyses, flag results requiring expert review, and retain complete provenance of every analytical decision. The goal should not simply be automated prediction, but transparent and reproducible analysis that allows investigators to understand how each result was generated.

Measures of impact could include successful reproduction of published endpoints from raw data, number and diversity of independent datasets that can be analyzed using a common framework, external cross-laboratory validation, identification of previously unrecognized sources of analytical variability, and demonstrated reuse of archived datasets for cross-study validation or synthesis.

Such investments would increase the return on existing NIH-funded datasets while creating infrastructure for future studies to be designed for reproducibility and interoperability from the outset.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: NIH should recognize technologies that transform fragile, expert-dependent research measurements into robust and deployable measurement systems as an important form of translational impact.

Many promising biomarkers and functional assays perform well in specialized academic laboratories but are difficult to implement in multicenter studies or routine clinical settings because they depend on fresh specimens, narrow processing windows, highly trained operators, specialized laboratory infrastructure, or immediate local analysis. These logistical constraints can prevent otherwise informative

measurements from progressing from discovery studies into large clinical trials, longitudinal cohorts, community-based research, or eventual clinical use.

NIH could incentivize technologies that address these barriers across the full workflow. Examples include methods that preserve biological measurements during delayed analysis and transportation; simplified or automated sample-processing systems that reduce operator dependence; approaches that allow specimen processing to occur near the participant while specialized measurement or analysis occurs centrally; and computational systems that maintain consistent quality control and interpretation across sites.

Translation should be assessed not only by patents, licensing, company formation, or regulatory milestones, but also by demonstrated improvement in deployability and measurement reliability. Relevant indicators could include extension of allowable specimen-processing or analysis windows, preservation of sample-level measurements during shipping, reduction in specialized infrastructure or operator requirements, successful implementation at multiple external sites, agreement across instruments and laboratories, and preservation of performance under realistic rather than idealized laboratory conditions.

NIH could specifically support cross-site studies that evaluate whether an assay developed in an expert laboratory can retain quantitative measurement fidelity when deployed in decentralized or multicenter settings. This type of translational validation bridges an important gap between proof-of-concept assay development and broad clinical implementation.

A future funding priority focused on technologies that enable robust translation of complex biological measurements from specialized laboratories into multisite and decentralized settings could accelerate the clinical utility of promising functional biomarkers while simultaneously improving rigor and reproducibility.

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: Measurement Science and Assay Harmonization

NIH should consider measurement science and assay harmonization as a distinct dimension of scientific impact. Biomedical research depends on the ability to generate measurements that are reliable not only within one laboratory or experiment, but also across laboratories, studies, instruments, operators, batches, and time.

Differences between studies are often attributed to biological heterogeneity when part of the disagreement may instead arise from specimen handling, assay implementation, instrumentation, reagent variation, batch effects, or analytical interpretation. Conversely, imposing a single assay protocol on all laboratories may be impractical and can make valuable historical datasets effectively unusable. There is therefore a need for technologies and reference frameworks that can both reduce avoidable measurement variability prospectively and account for unavoidable heterogeneity retrospectively.

NIH could support an integrated measurement-science framework with two complementary goals. First, prospective measurement stabilization and standardization should make sensitive biological assays reproducible under realistic multicenter conditions, including delays, transportation, different operators, and decentralized specimen collection. Second, retrospective analytical harmonization should make current and archived datasets generated with different assay configurations more re-analyzable, auditable, and comparable.

This could include development of reference materials and benchmark datasets; quantitative sample-level stability criteria; cross-site and cross-instrument validation studies; interoperable metadata describing specimen handling and assay conditions; standardized quality-control frameworks; and transparent computational tools capable of applying common analytical principles across heterogeneous datasets.

NIH could assess the impact of such work through measures including reduction in inter-site measurement variability, successful reproduction of results across independent laboratories, preservation of individual-sample measurements under realistic handling conditions, successful re-analysis of independently generated or historical datasets, and demonstration that data from multiple studies can be objectively compared or synthesized.

Dedicated investment in measurement science and assay harmonization would address an important gap between experimental rigor, data science, and translation. It would improve confidence in new studies while also increasing the scientific value of prior NIH investments by making existing data more useful for validation, replication, and new discovery.

143. mPower, Inc.

I am responding: On behalf of an organization

Name: Julian McElveen

Name of Organization: mPower, Inc.

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Organizational Official

Date Submitted: 8/7/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: Building and Measuring Collaborative Scientific Ecosystems: A Pilot Framework for Advancing Team Science and Scientific Impact

Executive Introduction and Background
The National Institutes of Health (NIH) is to be applauded for seeking input on future approaches to measuring collaborative science and scientific ecosystems. This Request for Information reflects NIH's continued leadership in advancing evidence-based approaches to understanding, evaluating, and strengthening modern biomedical research.

Traditional measures, including publications, citations, grant funding, patents, and clinical outcomes, remain fundamental indicators of scientific productivity and impact. However, these measures primarily assess the results of scientific activity rather than the organizational capabilities, collaborative behaviors, scientific networks, and research environments that enable discovery.

Over the past decade, NIH, the National Academies of Sciences, Engineering, and Medicine (NASEM), the National Cancer Institute (NCI), the Clinical and Translational Science Awards (CTSA) Program, and other Team Science initiatives have emphasized the importance of strengthening interdisciplinary collaboration and developing more effective approaches for evaluating collaborative research. These efforts have advanced the science of collaboration and provide a strong foundation for further understanding the organizational conditions that consistently produce high-performing scientific teams and sustain scientific impact.

We recommend that NIH complement, not replace, traditional outcome measures with an evidence-based framework that also evaluates the collaborative scientific ecosystem supporting research. The proposed Framework for Measuring Collaborative Scientific Ecosystems integrates leading indicators of collaborative capacity with lagging indicators of scientific outcomes to provide a more comprehensive understanding of scientific performance.

The framework is founded on three principles:
1. Traditional scientific outcomes remain essential and should continue to serve as the foundation for evaluating research productivity and impact.
2. Collaborative capacity can be measured through organizational capabilities, scientific networks, leadership, governance, collaborative behaviors, and workforce competencies that enable and incentivize effective Team Science.
3. Evidence should be drawn from multiple complementary sources to provide a comprehensive understanding of collaborative biomedical research.

We propose that NIH evaluate the Framework for Measuring Collaborative Scientific Ecosystems through a phased pilot conducted across multiple Institutes and Centers, scientific disciplines, career stages, and organizational settings. The pilot would provide practical experience, strengthen the evidence base, and assess the feasibility, reliability, administrative resiliency, predictive value, and any potential unintended consequences associated with implementation.

Building upon NIH's longstanding leadership in Team Science, this approach provides a practical pathway for strengthening Scientific Portfolio Stewardship, advancing organizational learning, and informing future guidance for evaluating collaborative scientific ecosystems through evidence-based implementation and continuous refinement.

1. Framework for Measuring Collaborative Scientific Ecosystems
Biomedical research has become increasingly collaborative, interdisciplinary, data-intensive, and translational. Traditional measures of scientific productivity, including publications, citations, grant funding, patents, and clinical outcomes, remain essential indicators of scientific success. However, these measures primarily assess the results of scientific activity rather than the organizational capabilities that enable discovery. As such, they function primarily as lagging indicators, reflecting achievements after they occur.

To better understand, strengthen, and sustain modern biomedical research, NIH should also consider leading indicators that assess the collaborative capacity supporting successful Team Science. These indicators evaluate the capabilities, behaviors, relationships, leadership, and organizational conditions that create an environment where scientific innovation can flourish before traditional outcomes become evident.

The proposed Framework for Measuring Collaborative Scientific Ecosystems (Figure 1 in attached PDF version of this Response) integrates leading and lagging indicators into a complementary approach for evaluating both scientific outcomes and the collaborative environment that produces them. The framework recognizes that modern biomedical discovery is increasingly generated through interconnected scientific ecosystems composed of investigators, multidisciplinary teams, collaborative behaviors, shared infrastructure, leadership, governance, institutional policies, incentive structures, recognition of collaborative contributions, technology platforms, and collaborative scientific networks.

Leading Indicators: Building Collaborative Scientific Capacity

Leading indicators evaluate whether an organization is developing the collaborative capabilities necessary to support sustained scientific excellence. The framework proposes four complementary dimensions.

> Individual Competencies assesses the scientific, technical, leadership, and collaborative capabilities that enable investigators to participate effectively in interdisciplinary research. These competencies may be evaluated using existing workforce development measures, competency assessments, education, and professional development activities.

> Collaborative Behaviors examines observable practices that advance Team Science, including data sharing, mentoring, knowledge exchange, shared decision-making, resource sharing, and cross-disciplinary problem solving. These behaviors provide evidence of how collaboration functions within research environments and may be assessed through structured interview analysis, direct observations, and complementary team analytic methods.

> Social Network Analysis (SNA) is an established quantitative, evidence-based methodology for measuring and analyzing collaborative relationships, patterns of connectivity, interdisciplinary interactions, and knowledge flows that support scientific collaboration. SNA provides objective measures of how investigators, teams, and organizations collaborate and helps assess the health of collaborative scientific networks over time.

> Scientific Ecosystems assesses the organizational environment that enables collaboration. This includes governance, leadership support, policies, incentives, shared infrastructure, technology platforms, workforce development, and organizational culture. Evaluating these conditions provides insight into an organization's collaborative capacity, readiness for Team Science, and ability to sustain scientific innovation over time.

Lagging Indicators: Measuring Scientific Results and Impact

Lagging indicators continue to provide the primary evidence of scientific achievement and long-term mission success.

> Scientific Portfolio Outcomes include measures of scientific innovation, translational progress, research productivity, workforce development, stewardship of scientific resources, and effective use of shared infrastructure.

> Patient and Public Impact reflects the ultimate purpose of biomedical research through improvements in health outcomes, public benefit, scientific advancement, equitable access to innovation, and return on public investment.

These outcome measures remain essential but become more informative when interpreted alongside the leading indicators that help explain how collaborative scientific success is achieved.

Recognition of Contributions to Team Science

Modern biomedical discovery increasingly depends on contributions that extend beyond traditional measures of authorship and principal investigator roles. Collaborative leadership, mentoring, data stewardship, development of shared research infrastructure, cross-disciplinary coordination, scientific facilitation, and contributions to collaborative networks all play important roles in advancing Team Science. Expanding recognition of these diverse contributions would better reflect how modern collaborative science is conducted while reinforcing behaviors that strengthen scientific ecosystems.

A Complementary Measurement Approach

The proposed framework is intended to complement, not replace traditional approaches for evaluating scientific productivity. By integrating leading indicators of collaborative capacity with lagging indicators of scientific outcomes, NIH can develop a more complete understanding of how collaborative scientific ecosystems contribute to discovery, translation, and public health impact.

This multidimensional approach provides opportunities to strengthen organizational learning, improve Scientific Portfolio Stewardship, identify opportunities to strengthen collaboration, and evaluate investments in Team Science. Because collaborative scientific environments differ across Institutes and Centers, the framework is intended for validation through a phased pilot that can assess feasibility, reliability, and predictive value before broader implementation.

2. Multiple Sources of Evidence

A fundamental principle of the Framework for Measuring Collaborative Scientific Ecosystems is that a comprehensive understanding of modern biomedical research is achieved through the integration of multiple complementary sources of evidence. Scientific collaboration reflects the interaction of individual capabilities, team dynamics, organizational culture, leadership, governance, infrastructure, and scientific outcomes. Together, these factors provide a balanced understanding of collaborative scientific performance and the conditions that enable successful Team Science.

The framework integrates qualitative and quantitative evidence to provide a comprehensive view of how collaborative scientific ecosystems function, support scientific discovery, and contribute to long-term scientific success. Each source of evidence offers unique insight into scientific performance, and together they provide a richer understanding than any single measure alone.

Proposed Evidence Sources

Quantitative Measures

> Quantitative measures provide objective evidence of scientific activity, collaborative capacity, and research outcomes. Examples may include bibliometric indicators, grant activity, collaboration metrics, shared publications, data-sharing metrics, utilization of shared scientific resources, translational outputs, organizational performance indicators, and other validated measures appropriate to the scientific environment.

Qualitative Measures

> Qualitative measures provide context for quantitative findings and help explain the organizational conditions that influence collaborative scientific performance. Examples may include interviews, focus groups, structured observations, case studies, after-action reviews, expert assessments, and other validated qualitative methods appropriate to the scientific environment. Together, these sources provide insight into collaborative experiences, leadership effectiveness, organizational culture, emerging practices, and the factors that contribute to successful Team Science and collaborative scientific ecosystems.

Social Network Analysis

> Social Network Analysis (SNA) is an established quantitative, evidence-based methodology for

measuring and analyzing collaborative relationships, patterns of connectivity, interdisciplinary interactions, and knowledge flows that support scientific collaboration. SNA provides objective measures of how investigators, teams, and organizations collaborate by identifying patterns of connectivity, boundary-spanning interactions, knowledge exchange, and the overall health of collaborative scientific ecosystems. These measures complement traditional scientific outcomes by providing evidence of the collaborative capacity that enables discovery, innovation, and long-term scientific impact.

Leadership Assessment
> Leadership plays a critical role in shaping collaborative scientific environments. Leadership assessments evaluate how vision, governance, decision-making, incentive structures, and organizational support foster Team Science and strengthen Scientific Portfolio Stewardship. These assessments provide insight into the leadership practices that encourage collaboration, recognize diverse scientific contributions, promote rigorous and transparent scientific practices, and create the organizational conditions necessary to sustain effective Team Science and collaborative scientific ecosystems.

Organizational Data
> Organizational data provide evidence regarding the environment in which scientific research is conducted and the organizational conditions that enable successful Team Science. Examples may include governance structures, institutional policies, resource allocation, shared infrastructure, workforce characteristics, training investments, technology platforms, incentive structures, recognition practices, and other organizational indicators appropriate to the scientific environment. Together, these data provide insight into organizational readiness, collaborative capacity, leadership support, and the conditions that sustain collaborative scientific ecosystems and long-term scientific excellence.

Each source of evidence contributes a unique perspective on collaborative scientific performance. No single measure can fully capture the complexity of modern scientific ecosystems. Integrating complementary sources of evidence provides a more comprehensive understanding of scientific performance, collaborative capacity, organizational effectiveness, and the conditions and incentives that support scientific excellence while reducing reliance on any individual measure. For example:

> Quantitative measures provide objective evidence of scientific activities, collaborative capacity, performance, and outcomes.
> Qualitative measures provide context for those findings by capturing organizational experiences, culture, leadership practices, and collaborative behaviors.
> Social Network Analysis illustrates how investigators, expertise, and knowledge are connected across collaborative scientific networks.
> Leadership assessments evaluate how vision, governance, decision-making, and organizational support foster Team Science and collaborative scientific ecosystems.
> Organizational data demonstrate how governance, institutional policies, infrastructure, resources, incentive structures, and recognition practices contribute to collaborative scientific environments.

Together, these complementary sources of evidence provide a more complete understanding of scientific performance, organizational capability, collaborative capacity, and the factors that enable successful Team Science. Integrating multiple sources of evidence also strengthens interpretation, supports

informed decision-making, and helps identify potential unintended consequences across diverse scientific environments.

Value of Multiple Sources of Evidence
Integrating multiple sources of evidence strengthens organizational learning, supports informed decision-making, and provides a more comprehensive understanding of collaborative scientific ecosystems. This multidimensional approach enables leadership to recognize organizational strengths, strengthen collaboration, evaluate investments in Team Science, encourage rigorous and transparent scientific practices, and continuously enhance the scientific ecosystem.

This approach aligns with the principles of Scientific Portfolio Stewardship by providing leadership with evidence that supports strategic decision-making across the research enterprise while complementing established measures of scientific productivity and impact. By recognizing diverse scientific contributions and reinforcing collaborative behaviors, the framework encourages the capabilities required for modern Team Science while preserving flexibility across scientific disciplines, career stages, institution types, and organizational missions.

3. Proposed Phased Pilot Implementation Framework
The proposed framework provides NIH with an opportunity to advance the evaluation of collaborative biomedical research through a structured pilot implementation. Conducting a pilot across multiple Institutes and Centers, scientific disciplines, career stages, institution types, and organizational settings would generate practical experience, strengthen the evidence base, and support the continued evolution of measures and incentives for collaborative scientific ecosystems.

The primary objective is to evaluate leading indicators that complement traditional measures of scientific productivity while providing additional insight into the organizational capabilities, collaborative behaviors, scientific networks, leadership practices, and ecosystem conditions that contribute to long-term scientific success and Scientific Portfolio Stewardship.

The pilot is intended to support organizational learning, evaluate the practical application of the framework, and inform future NIH guidance for measuring collaborative scientific ecosystems. Pilot implementation would assess the feasibility, usefulness, consistency, and applicability of the proposed measures across diverse scientific environments. The pilot would also examine relationships among leading and lagging indicators, evaluate implementation considerations, identify potential unintended consequences, and determine where flexibility may be appropriate across scientific disciplines, career stages, institution types, and organizational missions. Findings from the pilot would help refine the framework, identify effective implementation practices, and strengthen evidence-informed approaches that complement existing methods for evaluating scientific productivity and impact.

Proposed Pilot Phases
The proposed pilot consists of five sequential phases designed to support continuous learning, evaluation, and refinement.

Phase 1 – Framework Preparation

Conduct a review of existing evidence and leading practices; engage scientific and organizational stakeholders; identify candidate indicators; and develop the initial evaluation framework, implementation guidance, and data collection strategy.

Phase 2 – Pilot Design
Select participating Institutes, Centers, or research programs that represent diverse scientific disciplines, organizational structures, career stages, institution types, and research missions; establish baseline measures; define evaluation objectives; prepare implementation guidance; and orient participating organizations to the pilot framework and evaluation approach.

Phase 3 – Pilot Implementation
Implement the framework by collecting evidence across the proposed leading and lagging indicators, including individual competencies, collaborative behaviors, scientific networks, scientific ecosystem characteristics, scientific portfolio outcomes, and patient and public impact. Multiple complementary sources of evidence, including quantitative measures, qualitative assessments, Social Network Analysis, leadership assessments, and organizational data – would be examined together to provide a comprehensive understanding of collaborative scientific performance, organizational capability, collaborative capacity, and the organizational conditions that support scientific excellence.

Phase 4 – Evaluation and Learning
Evaluate pilot findings to assess the usefulness, consistency, feasibility, and practical application of the proposed measures. Examine relationships among leading and lagging indicators, identify implementation challenges and potential unintended consequences, assess flexibility across diverse scientific environments, capture lessons learned, and refine the framework and implementation guidance based on evidence gathered throughout the pilot.

Phase 5 – Guidance and Broader Application
Use the knowledge gained through the pilot to refine the framework, develop NIH guidance, establish recommended practices, and identify implementation strategies that support broader adoption of collaborative scientific ecosystem measures, where appropriate, across NIH.

Benefits and Implementation Considerations
The proposed framework is intended to complement existing evaluation approaches while remaining sufficiently flexible for use across diverse scientific environments. Because collaborative research differs by career stage, scientific discipline, organizational mission, institution type, and organizational context, pilot implementation across multiple Institutes and Centers will help determine which measures provide meaningful insight, which are broadly applicable, and where flexibility may be appropriate. This phased approach allows NIH to evaluate feasibility, administrative burden, consistency, predictive value, and potential unintended consequences before broader implementation.

The pilot is intended to inform future NIH guidance by identifying effective implementation practices, strengthening evidence-informed approaches to Scientific Portfolio Stewardship, and demonstrating how complementary sources of evidence can enhance understanding of collaborative scientific ecosystems while complementing established measures of scientific productivity and impact.

4. Recommendations to NIH

As biomedical research continues to evolve toward increasingly collaborative, interdisciplinary, and data-intensive models, NIH has an opportunity to complement existing measures of scientific productivity with approaches that also assess the collaborative capabilities and organizational conditions that enable scientific discovery. Based on the proposed Framework for Measuring Collaborative Scientific Ecosystems, we respectfully recommend that NIH consider the following actions:

1. Continue using traditional scientific outcome measures as the foundation for evaluating scientific productivity, scientific impact, and public benefit.
2. Expand scientific evaluation to include leading indicators that assess collaborative capacity, including individual competencies, collaborative behaviors, scientific networks, leadership practices, and the broader scientific ecosystem that supports Team Science and scientific excellence.
3. Adopt multiple sources of evidence approach that combines quantitative measures (including Social Network Analysis), qualitative assessments, leadership assessments, and organizational data to provide a more comprehensive understanding of collaborative scientific performance, organizational capability, and the conditions that support collaborative scientific ecosystems.
4. Conduct a phased pilot implementation across multiple Institutes and Centers, scientific disciplines, career stages, institution types, and organizational settings to evaluate the feasibility, usefulness, flexibility, and practical application of the proposed framework; strengthen the evidence base; identify effective implementation practices; and assess potential unintended consequences before broader adoption.
5. Use findings from the pilot to refine the framework, inform future NIH guidance, strengthen Scientific Portfolio Stewardship, and support evidence-informed approaches that complement established methods for evaluating scientific productivity and impact.

The proposed framework is intended to build upon NIH's longstanding leadership in Team Science by providing an evidence-informed approach for understanding not only the outcomes of scientific research but also the collaborative capabilities and organizational conditions that contribute to sustained scientific excellence. By integrating leading and lagging indicators through multiple sources of evidence, NIH can advance a more comprehensive approach to evaluating collaborative scientific ecosystems while strengthening Scientific Portfolio Stewardship, organizational learning, and patient and public benefit.

References

Team Science

> National Academies of Sciences, Engineering, and Medicine. 2025. The Science and Practice of Team Science. Washington, DC: The National Academies Press.
> Hall KL, Vogel AL, Croyle RT. Introduction. In Hall KL, Vogel AL, Croyle RT, eds. Strategies for Team Science Success: Handbook of Evidence-Based Principles for Cross-Disciplinary Science and Practical Lessons Learned from Health Researchers. New York, NY: Springer, Cham;2019:3-17. doi:10.1007/978-3-030-20992-6_1.
> National Research Council. Enhancing the Effectiveness of Team Science. Washington (DC): National Academies Press (US); 2015. Available from: https://www.nationalacademies.org/read/19007.

Social Network Analysis
> Stephenson K. Beyond Hierarchy: Heterarchy in the 21st Century [unpublished manuscript]. 2026.
> Borgatti SP, Everett MG, Johnson JC. Analyzing Social Networks. 2nd ed. Thousand Oaks (CA): SAGE Publications; 2018.
> Wasserman S, Faust K. Social Network Analysis: Methods and Applications. New York (NY): Cambridge University Press; 1994.
> Stephenson K, Zelen M. Rethinking centrality: methods and examples. Soc Networks. 1989;11(1):1-37. doi:10.1016/0378-8733(89)90016-6.

Organizational Learning & Performance
> Edmondson AC. Teaming: How Organizations Learn, Innovate, and Compete in the Knowledge Economy. San Francisco (CA): Jossey-Bass; 2012.
> Kaplan RS, Norton DP. The balanced scorecard—measures that drive performance. Harv Bus Rev. 1992;70(1):71-79.

Scientific Ecosystems
> Adner R. Ecosystem as structure: an actionable construct for strategy. J Manage. 2017;43(1):39-58. doi:10.1177/0149206316678451.

NIH Resources
> National Institutes of Health. Office of the Director. Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087) [Internet]. Bethesda (MD): National Institutes of Health; 2026 Jun 18 [cited 2026 Jul 09]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html.
> National Cancer Institute. Division of Cancer Control and Population Sciences. Team Science Toolkit [Internet]. Bethesda (MD): National Cancer Institute; [cited 2026 Jul 09]. Available from: https://cancercontrol.cancer.gov/brp/research/team-science-toolkit.
> National Institutes of Health. Office of Extramural Research. Data Management & Sharing Policy Overview [Internet]. Bethesda (MD): National Institutes of Health; [cited 2026 Jul 09]. Available from: https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/dms/policy-overview.

Foundational Work Supporting the Proposed Framework
The following prepublication drafts informed development of the proposed framework and are included to provide additional context regarding its conceptual foundations:
> Stephenson, K., & Levosky, L. (2026). Measuring 21st Century Science: Social Network Analysis—An Established Methodology for Evaluating Team Science. Prepublication draft, unpublished manuscript.
> Stephenson, K., & Levosky, L. (2026). Building and Measuring Collaborative Scientific Ecosystems: A Conceptual Framework for Advancing Team Science and Scientific Impact. Prepublication draft, unpublished manuscript.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/mPower_Response_to_NIH_OSP_RFI_NOT-OD-26-087_April_2026.pdf

Description: mPower's Response to NIH OSP RFI NOT-OD-26-087 (April 2026)

mPower, Inc.
1445 Research Blvd, Suite 301, Rockville, MD 20850
mPower-Inc.com

Building and Measuring Collaborative Scientific Ecosystems: A Pilot Framework for Advancing Team Science and Scientific Impact

Executive Introduction and Background

The National Institutes of Health (NIH) is to be applauded for seeking input on future approaches to measuring collaborative science and scientific ecosystems. This Request for Information reflects NIH's continued leadership in advancing evidence-based approaches to understanding, evaluating, and strengthening modern biomedical research.

Traditional measures, including publications, citations, grant funding, patents, and clinical outcomes, remain fundamental indicators of scientific productivity and impact. However, these measures primarily assess the results of scientific activity rather than the organizational capabilities, collaborative behaviors, scientific networks, and research environments that enable discovery.

Over the past decade, NIH, the National Academies of Sciences, Engineering, and Medicine (NASEM), the National Cancer Institute (NCI), the Clinical and Translational Science Awards (CTSA) Program, and other Team Science initiatives have emphasized the importance of strengthening interdisciplinary collaboration and developing more effective approaches for evaluating collaborative research. These efforts have advanced the science of collaboration and provide a strong foundation for further understanding the organizational conditions that consistently produce high-performing scientific teams and sustain scientific impact.

We recommend that NIH complement, not replace, traditional outcome measures with an evidence-based framework that also evaluates the collaborative scientific ecosystem supporting research. The proposed **Framework for Measuring Collaborative Scientific Ecosystems** integrates leading indicators of collaborative capacity with lagging indicators of scientific outcomes to provide a more comprehensive understanding of scientific performance.

The framework is founded on three principles:

1. **Traditional scientific outcomes remain essential** and should continue to serve as the foundation for evaluating research productivity and impact.
2. **Collaborative capacity can be measured** through organizational capabilities, scientific networks, leadership, governance, collaborative behaviors, and workforce competencies that enable and incentivize effective Team Science.
3. **Evidence should be drawn from multiple complementary sources** to provide a comprehensive understanding of collaborative biomedical research.

We propose that NIH evaluate the Framework for Measuring Collaborative Scientific Ecosystems through a phased pilot conducted across multiple Institutes and Centers, scientific disciplines, career stages, and organizational settings. The pilot would provide practical experience, strengthen the evidence base, and assess the feasibility, reliability, administrative resiliency, predictive value, and any potential unintended consequences associated with implementation.

Building upon NIH's longstanding leadership in Team Science, this approach provides a practical pathway for strengthening Scientific Portfolio Stewardship, advancing organizational learning, and informing

future guidance for evaluating collaborative scientific ecosystems through evidence-based implementation and continuous refinement.

1. Framework for Measuring Collaborative Scientific Ecosystems

Biomedical research has become increasingly collaborative, interdisciplinary, data-intensive, and translational. Traditional measures of scientific productivity, including publications, citations, grant funding, patents, and clinical outcomes, remain essential indicators of scientific success. However, these measures primarily assess the results of scientific activity rather than the organizational capabilities that enable discovery. As such, they function primarily as **lagging indicators**, reflecting achievements after they occur.

To better understand, strengthen, and sustain modern biomedical research, NIH should also consider **leading indicators** that assess the collaborative capacity supporting successful Team Science. These indicators evaluate the capabilities, behaviors, relationships, leadership, and organizational conditions that create an environment where scientific innovation can flourish before traditional outcomes become evident.

The proposed **Framework for Measuring Collaborative Scientific Ecosystems** (*Figure 1.)* integrates leading and lagging indicators into a complementary approach for evaluating both scientific outcomes and the collaborative environment that produces them. The framework recognizes that modern biomedical discovery is increasingly generated through interconnected scientific ecosystems composed of investigators, multidisciplinary teams, collaborative behaviors, shared infrastructure, leadership, governance, institutional policies, incentive structures, recognition of collaborative contributions, technology platforms, and collaborative scientific networks.

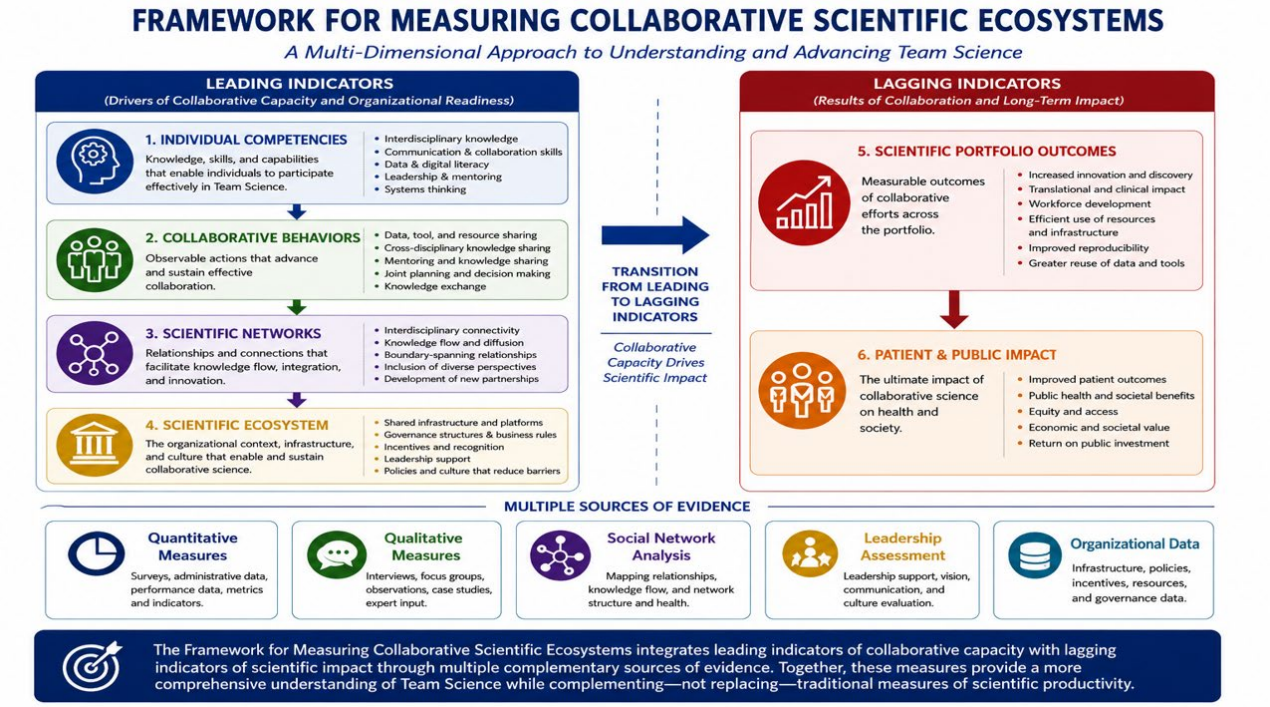


***Figure 1**. Framework for Measuring Collaborative Scientific Ecosystems*

Leading Indicators: Building Collaborative Scientific Capacity

Leading indicators evaluate whether an organization is developing the collaborative capabilities necessary to support sustained scientific excellence. The framework proposes four complementary dimensions.

- **Individual Competencies** assesses the scientific, technical, leadership, and collaborative capabilities that enable investigators to participate effectively in interdisciplinary research. These competencies may be evaluated using existing workforce development measures, competency assessments, education, and professional development activities.
- **Collaborative Behaviors** examines observable practices that advance Team Science, including data sharing, mentoring, knowledge exchange, shared decision-making, resource sharing, and cross-disciplinary problem solving. These behaviors provide evidence of how collaboration functions within research environments and may be assessed through structured interview analysis, direct observations, and complementary team analytic methods.
- **Social Network Analysis (SNA)** is an established quantitative, evidence-based methodology for measuring and analyzing collaborative relationships, patterns of connectivity, interdisciplinary interactions, and knowledge flows that support scientific collaboration. SNA provides objective measures of how investigators, teams, and organizations collaborate and helps assess the health of collaborative scientific networks over time**.**
- **Scientific Ecosystems** assesses the organizational environment that enables collaboration. This includes governance, leadership support, policies, incentives, shared infrastructure, technology platforms, workforce development, and organizational culture. Evaluating these conditions provides insight into an organization's collaborative capacity, readiness for Team Science, and ability to sustain scientific innovation over time.

Lagging Indicators: Measuring Scientific Results and Impact

Lagging indicators continue to provide the primary evidence of scientific achievement and long-term mission success.

- **Scientific Portfolio Outcomes** include measures of scientific innovation, translational progress, research productivity, workforce development, stewardship of scientific resources, and effective use of shared infrastructure.
- **Patient and Public Impact** reflects the ultimate purpose of biomedical research through improvements in health outcomes, public benefit, scientific advancement, equitable access to innovation, and return on public investment.

These outcome measures remain essential but become more informative when interpreted alongside the leading indicators that help explain how collaborative scientific success is achieved.

Recognition of Contributions to Team Science

Modern biomedical discovery increasingly depends on contributions that extend beyond traditional measures of authorship and principal investigator roles. Collaborative leadership, mentoring, data stewardship, development of shared research infrastructure, cross-disciplinary coordination, scientific

facilitation, and contributions to collaborative networks all play important roles in advancing Team Science. Expanding recognition of these diverse contributions would better reflect how modern collaborative science is conducted while reinforcing behaviors that strengthen scientific ecosystems.

A Complementary Measurement Approach

The proposed framework is intended to **complement, not replace traditional approaches for evaluating scientific productivity.** By integrating leading indicators of collaborative capacity with lagging indicators of scientific outcomes, NIH can develop a more complete understanding of how collaborative scientific ecosystems contribute to discovery, translation, and public health impact.

This multidimensional approach provides opportunities to strengthen organizational learning, improve Scientific Portfolio Stewardship, identify opportunities to strengthen collaboration, and evaluate investments in Team Science. Because collaborative scientific environments differ across Institutes and Centers, the framework is intended for validation through a phased pilot that can assess feasibility, reliability, and predictive value before broader implementation.

2. Multiple Sources of Evidence

A fundamental principle of the **Framework for Measuring Collaborative Scientific Ecosystems** is that a comprehensive understanding of modern biomedical research is achieved through the integration of multiple complementary sources of evidence. Scientific collaboration reflects the interaction of individual capabilities, team dynamics, organizational culture, leadership, governance, infrastructure, and scientific outcomes. Together, these factors provide a balanced understanding of collaborative scientific performance and the conditions that enable successful Team Science.

The framework integrates qualitative and quantitative evidence to provide a comprehensive view of how collaborative scientific ecosystems function, support scientific discovery, and contribute to long-term scientific success. Each source of evidence offers unique insight into scientific performance, and together they provide a richer understanding than any single measure alone.

Proposed Evidence Sources

Quantitative Measures

- Quantitative measures provide objective evidence of scientific activity, collaborative capacity, and research outcomes. Examples may include bibliometric indicators, grant activity, collaboration metrics, shared publications, data-sharing metrics, utilization of shared scientific resources, translational outputs, organizational performance indicators, and other validated measures appropriate to the scientific environment.

Qualitative Measures

- Qualitative measures provide context for quantitative findings and help explain the organizational conditions that influence collaborative scientific performance. Examples may include interviews, focus groups, structured observations, case studies, after-action reviews, expert assessments, and other validated qualitative methods appropriate to the scientific environment. Together, these sources provide insight into collaborative experiences, leadership effectiveness, organizational culture, emerging practices, and the factors that contribute to successful Team Science and collaborative scientific ecosystems.

Social Network Analysis

- **Social Network Analysis (SNA)** is an established quantitative, evidence-based methodology for measuring and analyzing collaborative relationships, patterns of connectivity, interdisciplinary interactions, and knowledge flows that support scientific collaboration. SNA provides objective measures of how investigators, teams, and organizations collaborate by identifying patterns of connectivity, boundary-spanning interactions, knowledge exchange, and the overall health of collaborative scientific ecosystems. **These measures complement traditional scientific outcomes by providing evidence of the collaborative capacity that enables discovery, innovation, and long-term scientific impact.**

Leadership Assessment

- Leadership plays a critical role in shaping collaborative scientific environments. Leadership assessments evaluate how vision, governance, decision-making, incentive structures, and organizational support foster Team Science and strengthen Scientific Portfolio Stewardship. These assessments provide insight into the leadership practices that encourage collaboration, recognize diverse scientific contributions, promote rigorous and transparent scientific practices, and create the organizational conditions necessary to sustain effective Team Science and collaborative scientific ecosystems.

Organizational Data

- Organizational data provide evidence regarding the environment in which scientific research is conducted and the organizational conditions that enable successful Team Science. Examples may include governance structures, institutional policies, resource allocation, shared infrastructure, workforce characteristics, training investments, technology platforms, incentive structures, recognition practices, and other organizational indicators appropriate to the scientific environment. Together, these data provide insight into organizational readiness, collaborative capacity, leadership support, and the conditions that sustain collaborative scientific ecosystems and long-term scientific excellence.

Each source of evidence contributes a unique perspective on collaborative scientific performance. No single measure can fully capture the complexity of modern scientific ecosystems. Integrating complementary sources of evidence provides a more comprehensive understanding of scientific performance, collaborative capacity, organizational effectiveness, and the conditions and incentives that support scientific excellence while reducing reliance on any individual measure. For example:

- **Quantitative measures** provide objective evidence of scientific activities, collaborative capacity, performance, and outcomes.
- **Qualitative measures** provide context for those findings by capturing organizational experiences, culture, leadership practices, and collaborative behaviors.
- **Social Network Analysis** illustrates how investigators, expertise, and knowledge are connected across collaborative scientific networks.
- **Leadership assessments** evaluate how vision, governance, decision-making, and organizational support foster Team Science and collaborative scientific ecosystems.

- **Organizational data** demonstrate how governance, institutional policies, infrastructure, resources, incentive structures, and recognition practices contribute to collaborative scientific environments.

Together, these complementary sources of evidence provide a more complete understanding of scientific performance, organizational capability, collaborative capacity, and the factors that enable successful Team Science. Integrating multiple sources of evidence also strengthens interpretation, supports informed decision-making, and helps identify potential unintended consequences across diverse scientific environments.

Value of Multiple Sources of Evidence

Integrating multiple sources of evidence strengthens organizational learning, supports informed decision-making, and provides a more comprehensive understanding of collaborative scientific ecosystems. This multidimensional approach enables leadership to recognize organizational strengths, strengthen collaboration, evaluate investments in Team Science, encourage rigorous and transparent scientific practices, and continuously enhance the scientific ecosystem.

This approach aligns with the principles of Scientific Portfolio Stewardship by providing leadership with evidence that supports strategic decision-making across the research enterprise while complementing established measures of scientific productivity and impact. By recognizing diverse scientific contributions and reinforcing collaborative behaviors, the framework encourages the capabilities required for modern Team Science while preserving flexibility across scientific disciplines, career stages, institution types, and organizational missions.

3. Proposed Phased Pilot Implementation Framework

The proposed framework provides NIH with an opportunity to advance the evaluation of collaborative biomedical research through a structured pilot implementation. Conducting a pilot across multiple Institutes and Centers, scientific disciplines, career stages, institution types, and organizational settings would generate practical experience, strengthen the evidence base, and support the continued evolution of measures and incentives for collaborative scientific ecosystems.

The primary objective is to evaluate leading indicators that complement traditional measures of scientific productivity while providing additional insight into the organizational capabilities, collaborative behaviors, scientific networks, leadership practices, and ecosystem conditions that contribute to long-term scientific success and Scientific Portfolio Stewardship.

The pilot is intended to support organizational learning, evaluate the practical application of the framework, and inform future NIH guidance for measuring collaborative scientific ecosystems. Pilot implementation would assess the feasibility, usefulness, consistency, and applicability of the proposed measures across diverse scientific environments. The pilot would also examine relationships among leading and lagging indicators, evaluate implementation considerations, identify potential unintended consequences, and determine where flexibility may be appropriate across scientific disciplines, career stages, institution types, and organizational missions. Findings from the pilot would help refine the framework, identify effective implementation practices, and strengthen evidence-informed approaches that complement existing methods for evaluating scientific productivity and impact.

Proposed Pilot Phases

The proposed pilot consists of five sequential phases designed to support continuous learning, evaluation, and refinement.

Phase 1 – Framework Preparation

Conduct a review of existing evidence and leading practices; engage scientific and organizational stakeholders; identify candidate indicators; and develop the initial evaluation framework, implementation guidance, and data collection strategy.

Phase 2 – Pilot Design

Select participating Institutes, Centers, or research programs that represent diverse scientific disciplines, organizational structures, career stages, institution types, and research missions; establish baseline measures; define evaluation objectives; prepare implementation guidance; and orient participating organizations to the pilot framework and evaluation approach.

Phase 3 – Pilot Implementation

Implement the framework by collecting evidence across the proposed leading and lagging indicators, including individual competencies, collaborative behaviors, scientific networks, scientific ecosystem characteristics, scientific portfolio outcomes, and patient and public impact. Multiple complementary sources of evidence, including quantitative measures, qualitative assessments, Social Network Analysis, leadership assessments, and organizational data – would be examined together to provide a comprehensive understanding of collaborative scientific performance, organizational capability, collaborative capacity, and the organizational conditions that support scientific excellence.

Phase 4 – Evaluation and Learning

Evaluate pilot findings to assess the usefulness, consistency, feasibility, and practical application of the proposed measures. Examine relationships among leading and lagging indicators, identify implementation challenges and potential unintended consequences, assess flexibility across diverse scientific environments, capture lessons learned, and refine the framework and implementation guidance based on evidence gathered throughout the pilot.

Phase 5 – Guidance and Broader Application

Use the knowledge gained through the pilot to refine the framework, develop NIH guidance, establish recommended practices, and identify implementation strategies that support broader adoption of collaborative scientific ecosystem measures, where appropriate, across NIH.

Benefits and Implementation Considerations

The proposed framework is intended to complement existing evaluation approaches while remaining sufficiently flexible for use across diverse scientific environments. Because collaborative research differs by career stage, scientific discipline, organizational mission, institution type, and organizational context, pilot implementation across multiple Institutes and Centers will help determine which measures provide meaningful insight, which are broadly applicable, and where flexibility may be appropriate. This phased approach allows NIH to evaluate feasibility, administrative burden, consistency, predictive value, and potential unintended consequences before broader implementation.

The pilot is intended to inform future NIH guidance by identifying effective implementation practices, strengthening evidence-informed approaches to Scientific Portfolio Stewardship, and demonstrating how complementary sources of evidence can enhance understanding of collaborative scientific ecosystems while complementing established measures of scientific productivity and impact.

4. Recommendations to NIH

As biomedical research continues to evolve toward increasingly collaborative, interdisciplinary, and data-intensive models, NIH has an opportunity to complement existing measures of scientific productivity with approaches that also assess the collaborative capabilities and organizational conditions that enable scientific discovery. Based on the proposed **Framework for Measuring Collaborative Scientific Ecosystems**, we respectfully recommend that NIH consider the following actions:

1. **Continue using traditional scientific outcome measures** as the foundation for evaluating scientific productivity, scientific impact, and public benefit.
2. **Expand scientific evaluation** to include leading indicators that assess collaborative capacity, including individual competencies, collaborative behaviors, scientific networks, leadership practices, and the broader scientific ecosystem that supports Team Science and scientific excellence.
3. **Adopt multiple sources of evidence approach** that combines quantitative measures (including Social Network Analysis), qualitative assessments, leadership assessments, and organizational data to provide a more comprehensive understanding of collaborative scientific performance, organizational capability, and the conditions that support collaborative scientific ecosystems.
4. **Conduct a phased pilot implementation** across multiple Institutes and Centers, scientific disciplines, career stages, institution types, and organizational settings to evaluate the feasibility, usefulness, flexibility, and practical application of the proposed framework; strengthen the evidence base; identify effective implementation practices; and assess potential unintended consequences before broader adoption.
5. **Use findings from the pilot** to refine the framework, inform future NIH guidance, strengthen Scientific Portfolio Stewardship, and support evidence-informed approaches that complement established methods for evaluating scientific productivity and impact.

The proposed framework is intended to build upon NIH's longstanding leadership in Team Science by providing an evidence-informed approach for understanding not only the outcomes of scientific research but also the collaborative capabilities and organizational conditions that contribute to sustained scientific excellence. By integrating leading and lagging indicators through multiple sources of evidence, NIH can advance a more comprehensive approach to evaluating collaborative scientific ecosystems while strengthening Scientific Portfolio Stewardship, organizational learning, and patient and public benefit.

References

Team Science

- National Academies of Sciences, Engineering, and Medicine. 2025. *The Science and Practice of Team Science*. Washington, DC: The National Academies Press.
- Hall KL, Vogel AL, Croyle RT. Introduction. In Hall KL, Vogel AL, Croyle RT, eds. *Strategies for Team Science Success: Handbook of Evidence-Based Principles for Cross-Disciplinary Science and Practical Lessons Learned from Health Researchers.* New York, NY: Springer, Cham;2019:3-17. doi:10.1007/978-3-030-20992-6_1.
- National Research Council. *Enhancing the Effectiveness of Team Science*. Washington (DC): National Academies Press (US); 2015. Available from: https://www.nationalacademies.org/read/19007.

Social Network Analysis

- Stephenson K. *Beyond Hierarchy: Heterarchy in the 21st Century* [unpublished manuscript]. 2026.
- Borgatti SP, Everett MG, Johnson JC. *Analyzing Social Networks*. 2nd ed. Thousand Oaks (CA): SAGE Publications; 2018.
- Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. New York (NY): Cambridge University Press; 1994.
- Stephenson K, Zelen M. Rethinking centrality: methods and examples. *Soc Networks*. 1989;11(1):1-37. doi:10.1016/0378-8733(89)90016-6.

Organizational Learning & Performance

- Edmondson AC. *Teaming: How Organizations Learn, Innovate, and Compete in the Knowledge Economy*. San Francisco (CA): Jossey-Bass; 2012.
- Kaplan RS, Norton DP. The balanced scorecard—measures that drive performance. Harv Bus Rev. 1992;70(1):71-79.

Scientific Ecosystems

- Adner R. Ecosystem as structure: an actionable construct for strategy. J Manage. 2017;43(1):39-58. doi:10.1177/0149206316678451.

NIH Resources

- National Institutes of Health. Office of the Director. Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087) [Internet]. Bethesda (MD): National Institutes of Health; 2026 Jun 18 [cited 2026 Jul 09]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html.
- National Cancer Institute. Division of Cancer Control and Population Sciences. Team Science Toolkit [Internet]. Bethesda (MD): National Cancer Institute; [cited 2026 Jul 09]. Available from: https://cancercontrol.cancer.gov/brp/research/team-science-toolkit.

- National Institutes of Health. Office of Extramural Research. Data Management & Sharing Policy Overview [Internet]. Bethesda (MD): National Institutes of Health; [cited 2026 Jul 09]. Available from: https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/dms/policy-overview.

Foundational Work Supporting the Proposed Framework

The following prepublication drafts informed development of the proposed framework and are included to provide additional context regarding its conceptual foundations:

- Stephenson, K., & Levosky, L. (2026). *Measuring 21st Century Science: Social Network Analysis—An Established Methodology for Evaluating Team Science.* Prepublication draft, unpublished manuscript.
- Stephenson, K., & Levosky, L. (2026). *Building and Measuring Collaborative Scientific Ecosystems: A Conceptual Framework for Advancing Team Science and Scientific Impact.* Prepublication draft, unpublished manuscript.

144. Lois Sayrs, PhD

I am responding: On behalf of myself

Name: Lois Sayrs, PhD

Name of Organization: Rady Children's Health and UCI School of Medicine

Type of Organization: Healthcare Institution

Role: Other

Role - Other: DIrector, Biostatistics

Date Submitted: 8/7/2026

1. Rigor and Reproducibility: As a biostatistician and methodologist, I have been particular concerned about the lack of rigor within "data science." This field has no theoretical basis, and no standards. there may be as many 600 "guidelines" but none are comprehensive and none effectively promote rigor, reproducibility and transparency. The field has evolved to serve themselves with improved computational techniques to save time and space. Similar issues attach to software engineering. Letting data drive any endeavor is a terrible idea. to move forward, we have to return to the fundamentals-theory, theory, theory. The problem with prediction is that there is no theory; as a result, we have returned to some 19th century understanding of philosophy of science. I have commented elsewhere on the misuse of the EHR in clinical research; I have also belied the lack of transparency in a recent paper and proposed a comprehensive framework for reproducibility and transparency as well as an audit procedure using curated AI queries--all of which can be automated as a digital tool to review research prior to publication. For funding, a tool like this could be utilized as a condition of funding. Overall, empirical research has gone too far in the direction of external validation with little or no though given to internal validity. I am also a strong proponent of simulation to assess the robustness of assumptions and theoretical propositions. Replication, in my opinion, has killed progress in clinical research. Now we are using meta-analysis to indemnify the baisses inherent in the replications. No thanks! The first action to take is to require 2 semesters of philosophy of science, not targeting bioethics which are important, but not the key problem. Trained researchers have misguided faith in the RCT and its replicants. Coming from social science to medicine, I was horrified at the lack of knowledge regarding the scientific endeavor. Decades of overvaluing the RCT has led to a dearth of skill in actual research. Now with data driving research and no theory, as well as unreliable data sets designed to track billing and not patient experience, are we really surprised we can't replicate results? If decades of pediatric cancer research with 80% of all NIH money cannot undertake research to reduce chemotherapy dosing and prevent horrific side effects, something is wrong with our research and funding mechanisms.

2. Data, Software, and Model Sharing: I can't agree with a replication approach to validation. Again, it misses the importance of theory and internal validation. You can never get to causality from replication. In fact, you can't really get anywhere because all it really takes is one counterexample to upend your research. As I described above, I am working on a comprehensive framework with AI auditing to support confirmation of results. Yes, all coding should be posted on GitHub or equivalent but more often than not, you find numerous problems with the code after the paper is published. i suggest demonstrating

validation via in silico (simulation testing) of your model and data. Yes, use alternate data sets but not for replication-rather for exploration via simulation to understand failure points and edges. We just completed this type of effort in a recent paper (abstract attached).

3. Training and Mentorship: Training and mentorship is incredibly time consuming and requires a mutli-year commitment. I have to applaud the new ASU Medical School for diving into the deep end and deciding every MD will also be a biomedical engineer. I am sure they will find, however, without deep mentorship in extant start-ups and with experienced developers/founders-the technical training only goes so far. Trainees must be embedded into MedTech ecospheres that extend beyond one community or university. founders (engineers) are notoriously ignorant of basically everything else they need to know about commercializing their tech. In addition, many are looking for lifelong careers for a single science project not business opportunities. In the Irvine Tech Hub, we are embedding MBAs, regulatory specialists, trial specialists etc. to educate founders. Beyond the founders, traditional training in biostats, data science, or regulatory science will not serve biomedical engineering. Even the bioscience side has huge gaps, for example, purchasing an enzyme for animal studies that works but can't be translated clinically because the company doesn't want the liability associated with their enzyme. WHAT? More biomedical engineers will not make the workforce you need for translation. So counting them won't help. I recommend supporting partnerships where start-ups are funded a trainee and assigned multiple projects or components to expand their applied skillset. The number of translated projects and investor funded projects from those partnerships would also be a metric to consider.

4. Collaboration: One of the best collaborations I have observed is the Pediatric Cardiac Research and Learning Collaborative. The group sets guidelines for clinical practice, engages patients and families, test new treatments and is well-poisiotned to evaluate new cardiac devices, as well as hold their own the with FDA (thus reducing costly and time-consuming replication). the group are the experts, they understand the theory, they can explain the science supporting the device or intervention, they engage patients and their families to ensure their work is patient and family first. 90% of the time, research is clinician-focused--what they want or need (and heaven help us -what they need for tenure and promotion) not what the patient wants or needs; not what the health system or the payor could benefit from. Personally, I target the latter in all my consults. what good does it do to have a slightly more accurate treatment but it requires in-patient surgery where the existing tx is outpatient? Let's refocus research on the big issues--reducing the cost of Medicaid not providing more and more elaborate treatments for already very expensive patients. Let's move 90% of hospital activity to the home--let's target devices that do that. Let's stop rampant replication of single institution findings where the noise is always larger than the signal. Let's engender research and learning collaboratives that meet all the requirements of effective research and learning consortia. Let's offer them funding to "solve" big health system issues and not only micro improvements. Let's incentivize these collabs to solve problems and NOT COLLECT DATA! All that data led only to career making papers but no real knowledge. Does cooling actually help a TBI? 20,000 patients later and we don't know. There are so many problems out there to address, let's incentivize collabs to address them.

5. Entrepreneurship and Translation: Oh goodness! the NIH mission may need to be revamped away from cool science projects with tons of promise or promoting academic careers, towards patient centered problems or reducing health system costs. 20 years of genomics and we are only now starting to integrated genome sequencing to target infections. And it's expensive and the lab work can only be done by one university with no plans to democratize it or bring down the cost. It's a public good

monopolized by a private university. How do we explain that? I would encourage the NIH to skip the science projects completely because from my observation the funding of these has been diverted to sustain a 20-year academic career with R01 grants and zero, zero, zero translation.

6. Foundational Scientific Exploration: Let a thousand flowers bloom! Open the doors to research that hasn't spent a decade being primed in an NIH grant machine. Revamp the grant application to work backward from the vision of the translation to impact (and commercialization, if possible). we need to incentivize foundational scientific exploration to align with the biggest problems we have today not the esoteric whims of highly specialized researchers who are not targeting major problems. the one-step-at-a-time approach only incentivized academic careers not impact. My favorite example is related to a young investigator how really identified an impactful treatment but simply refused to translate it before she received additional grant money to test it in related populations. first, this example shows how hesitant investigators are to translate their work; second, it shows her misguided reliance on external validation; third it shows her commitment to herself and her academic career over the patients she could have helped. I cant forget this example because it was so egregious in my view. going forward, this researcher will likely not get r01 money because her research is not foundational and the novelty of is quickly eroding. this doesn't preclude her being clinically impactful and seeking out private partnerships to enhance her impact. However, her academic career comes first. I also saw another investigator who had lived on r01s for 20 years derisking his medtech. by the time he was ready to commercialize, investors said "no thanks."

We have to define: foundational to what? To 300 year old specialties? I encourage you to rethink your entire approach to research and replace the solipsism with problem solving. we are too used to thinking science moves ahead by some quirk or step by step process. it can move ahead all sorts of ways but never based on someone putting their career ahead of the problem. Cool science projects for science's sake in my view must target problems, otherwise, we are just supporting somebody's interest regardless of its overall value to and impact on the public (who pay for this research).

7. Public Impact: Obviously, I believe this should come first, second, and third. I work extensively with clinicians who look at changing clinical practice as a modest one-step-at -a-time endeavor that more often than not serves them and not the patient. Does a patient need remote monitoring? What does it actually do for them, specifically. I guarantee you that the clinician hasn't given it a second of thought. For the clinician it only provides a lot more information that they absolutely don't need, won't use, and can only be integrated into the EHR in summary data or notes. The whole endeavor is to ping a patient to make an appointment-they won't be automating recommended adjustments to their treatment plan because hospitals require a "human in the loop" which at this stage means "make an appointment." We are also too incremental in our envisioning of patient-centered care. We think about it in steps-push them onto a digital platform, integrate telehealth, navigate towards some automation, take 5 years to complete. NIH should not fund anything where patients are not a key stakeholder in the research, not just an afterthought-full stop. Never. designing treatments and interventions and improvements in delivery or care must come from the patient's needs first. However, don't get me wrong-I work in peds so I have the luxury of early, early intervention planning and prevention. More funding and faster funding (and FDA approval) for game changing diagnostics and treatments should be the top pediatric target. Surgeons want to cut not prevent. I recently consulted on a grant application for epilepsy which was attempting to use qEEG to better predict time to surgery for Sturge-Weber Syndrome. I suggested

very strongly that they align with the major research that is looking to predict as early as possible in the pathway of the disease not fix it at the margins. We are simply not training our clinicians to solve big problems and when they do get exposed through journal club or collaboration, their immediate approach is replication. It's safe, its easy, its publishable. NIH will never break that horrific approach without a completely different incentive structure. Medicine is now overtly creating two classes of clinicians-researchers and non-researchers. This is akin to universities who have researchers who don't teach and a teaching corps of underpaid service workers. My suggestion is that no NIH money should be granted to any elite research group unless they are also training multiple young investigators outside of their own institution. This not only prevents the elite research group from hoarding but also limits their ability to exploit their fellows and residents (which they do). By taking steps like these, we are asking researchers to include people from excluded communities. in turn, those trained can better hone their research to solve their local community problems.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Bias_Aware_ABSTRACTv3.docx

Description: Abstract of a current paper under review on transparency.

A Bias-Aware Analytic Framework for Real-World Data with AI-Audited Workflows: Application to Opioid Ordering in Pediatric Oncology

T. Morphew, H. Pease, M. Fotier, Z. Kain and L. Sayrs (2026)

Background The growing use of computational modeling of real-world data (RWD) in clinical research introduces significant risks for data scientists already managing inconsistency, unexamined bias, and analytical opacity. Without the formal theory supporting biostatistical models and despite more than 600 guideline-based checklists for data quality and transparency reporting in observational RWD, reliance on *ad hoc* model specification and verification is still the norm. No structured guidance exists to help clinicians and researchers systematically assess and manage the risks generated by computational results using RWD. In this paper, we propose a framework that reflects a rigorous, bias-aware methodology for complex analyses of RWD. We also propose the curated use of a large language model (LLM) supporting AI-generated code-auditing to ensure data quality. Our structured hierarchical approach coupled with AI-auditing is the first phase in the development of full-stack automation of RWD analytic validation.

Methods Our analytic framework spans 10 domains across three analytic phases. Each domain includes prespecified criteria governing verification with potential biases, mechanisms, and mitigation strategies systematically mapped throughout. Phase 1 entails *Specification* (Domains 1–4: objectives and estimands, data inputs and outputs, cohort and filters, and variable coding); Phase 2:*Modeling* (Domains 5–7: model testing, diagnostics, and reasonableness checks); and Phase 3: *Robustness and Interpretation* (Domains 8–10: missing data handling, temporal and stratified checks, and scope boundaries). We utilize specific prompt-directed large language model (LLM) auditing of model elements to ensure data quality. The methodology is illustrated with examples from an empirical characterization of inpatient opioid data from Cerner *Real World* (2014–2021).

Results The Cerner RWD data base contained 16,734 pediatric cancer patients (76,395 encounters) across 55 U.S. health systems. During the specification phase, total daily dose could not be estimated with sufficient validity from Cerner RWD, and "as-needed" orders, warranted empirically supported exclusion. The modeling phase revealed residual variability in opioid ordering attributable primarily to patient-level differences within health systems (ICC=19.9%) rather than between health systems (ICC=2.0%). Temporal analysis revealed a downward trend in morphine and fentanyl ordering trajectories from 2017-onward. We then utilized curated prompting to enforce an AI-generated coding requirement to produce valid likelihood ratio tests across the prespecified model sequence accounting for temporal trends.

Conclusions The 10-domain framework provides a structured, reproducible mechanism for governing RWD analyses, supporting transparent exposure definition, systematic bias assessment, prespecified analytic decision-making, and auditing via curated AI prompting. To our knowledge, this is the first structured framework to 1) identify analytic metrics at risk for unreliability; 2) incorporate AI-audited workflows in RWD research; and 3) demonstrate utility in a large multi-institutional real-world data base. The approach reflects the initial development in automating of RWE data analytics.

145. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Research Trainee/Fellow

Date Submitted: 8/7/2026

1. Rigor and Reproducibility: Maybe there could be an intramural branch at NIH dedicated to replicating studies that have the potential to be very impactful, and therefore, must be verifiably accurate. There could be a signifier (think social media "verified check") on PubMed that is displayed on papers that have been independently verified at NIH or other institutions.

2. Data, Software, and Model Sharing: Continuing to require that data/repositories/etc. that are produced using NIH grants be made publicly accessible is important. As we move into the age of AI writing most code that biologists use, it may be appropriate to require NIH grant awardees to also make their code publicly available upon publication of a project it is used in.

3. Training and Mentorship: This needs to be taken seriously at the level of university hiring practices and grant application review. If the primary investigator of a grant (excluding training grants) does not specify how they are going to involve trainees in conducting the project, it should be considered less favorably because it was not well thought out regarding the inclusion of trainee development.

4. Collaboration: Do not implement the grant cap proposed earlier this summer if you truly value collaboration. Investigators will be much less likely to work with other labs in any kind of formal capacity if writing a grant with other investigators directly takes away from their ability to fund their own lab. Since there is a constant cycle of trying to get money to fund science and doing science to get money, it is not realistic to assume that investigators will prioritize collaborating with other labs if there is no money involved because that informal collaboration work takes away the resources and personnel required to work on projects that will allow the lab to keep bringing in money.

5. Entrepreneurship and Translation: Generally, most biosciences researchers are either entrepreneurially-minded or they're not. Having resources available to support investigators who are motivated to study directly translatable questions or work in the drug development pipeline is great, but for those of us who do science for the advancement of knowledge and enjoyment of learning about the world, any attempts to force these people to abandon their studies and engage instead in an entrepreneurial agenda will be met with opposition.

6. Foundational Scientific Exploration: There needs to be more money set aside for grants that can be given to investigators with a proven track record of producing innovative and quality science using the traditional, "safer" funding mechanisms and who could therefore be reasonably trusted to explore meaningful questions with longer, more unstructured funding. Constant progress reports are a hindrance

to being able to work on anything in a timely manner and prioritize "successful results" over following the data naturally and trying bold ideas even if they don't work. If researchers are punished for not having positive results, they don't dare to study anything risky.

7. Public Impact: In the grant awarding process, include a section for applicants to discuss their involvement in science communication, policy, and other public engagement efforts. This requires reviewers to deliberately change their criteria for assessing a productive investigator.

8. Other Areas: Many of these areas can be addressed by reviewers and upper level members within each institute deciding to "put their money where their mouth is". If these topics are truly valued by NIH, making these changes only requires people to make a conscious effort to reward investigators who already do these things but are limited in how much time and resources they can expend on these other pursuits, as they are currently not rewarded by NIH.

146. Annette Greer

I am responding: On behalf of myself

Name: Annette Greer

Name of Organization: Scholarly Integrity Review, LLC

Type of Organization: Other

Type of Organization - Other: Professional consulting and advocacy

Role: Bioethicist

Date Submitted: 8/10/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/SIR-Briefing-Measuring-and-Rewarding-Scientific-Impact-NIH-RFI-NOT-OD-26-087-3.docx

SIR BRIEFINGS

Scholarly Integrity Review

Measuring and Rewarding Scientific Impact:

An Ethical, Interprofessional Response to NIH RFI NOT-OD-26-087

Author: **Annette Greer, PhD, MSN, RN, FNAP** — Founder, Scholarly Integrity Review LLC

Personal capacity; affiliation used only for identification. Not an ECU or BSOM institutional response.

Response window: June 18, 2026 – August 19, 2026.[1]

Version: Publication-ready draft after SIR internal audit x 3 (source verification, bias review, audience-impact assessment).

[1]National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

Executive summary

NIH's Request for Information on Measuring and Rewarding Scientific Impact rightly recognizes that biomedical research is now a collaborative, interdisciplinary enterprise, and that publication counts and citation metrics are a poor proxy for the mission of enhancing health, lengthening life, and reducing illness.[2] The reform opportunity is real. So is the risk: any new indicator can become a new gaming target, and metric change without structural incentive change will not produce more rigorous or more trustworthy science.

This briefing offers a source-anchored, interprofessional response grounded in three convergent literatures — responsible metrics (**Leiden Manifesto**, **DORA**, **CRediT**),[345] team science (**National Academies 2015, 2025**),[67] and rigor and reproducibility (**Ioannidis 2005**, **Begley & Ellis 2012**, **NASEM 2017**, **NIH R&R Initiative**).[891011] It draws on the **IPEC 2023 Core Competencies** and Cochrane evidence on interprofessional collaboration to name what team-science measurement actually requires when the "team" includes nurses, social workers, pharmacists, community researchers, and patients — not only principal investigators.[1213]

Core recommendation. NIH should adopt a two-track indicator system: (1) an individual-level **Contribution Portfolio** anchored in CRediT roles, verifiable data/software/model deposits, mentorship

---

[2]National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

[3]Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015;520:429-431. https://www.nature.com/articles/520429a

[4]DORA. San Francisco Declaration on Research Assessment. 2013 (ongoing). https://sfdora.org/read/

[5]Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publishing. 2015;28(2):151-155. CRediT — Contributor Roles Taxonomy (ANSI/NISO Z39.104-2022). https://credit.niso.org/

[6]National Research Council. Enhancing the Effectiveness of Team Science. Washington, DC: National Academies Press; 2015. doi:10.17226/19007. https://www.nationalacademies.org/read/19007

[7]National Academies of Sciences, Engineering, and Medicine. The Science and Practice of Team Science. Washington, DC: National Academies Press; 2025. https://www.nationalacademies.org/read/29043/chapter/4

[8]Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

[9]Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature. 2012;483:531-533. doi:10.1038/483531a. https://www.nature.com/articles/483531a

[10]National Academies of Sciences, Engineering, and Medicine. Fostering Integrity in Research. Washington, DC: National Academies Press; 2017. doi:10.17226/21896. https://nap.nationalacademies.org/catalog/21896/fostering-integrity-in-research

[11]Bhattacharya J. NIH Strategic Vision and Plan for Strengthening Replication and Reproducibility. NIH Office of the Director briefing; 2026. https://dpcpsi.nih.gov/sites/g/files/mnhszr346/files/2026-01/1215PM-NIH-Update-Bhattacharya.pdf

[12]Interprofessional Education Collaborative. IPEC Core Competencies for Interprofessional Collaborative Practice, Version 3. Washington, DC: IPEC; 2023. https://www.ipecollaborative.org/ipec-core-competencies

[13]Reeves S, Pelone F, Harrison R, Goldman J, Zwarenstein M. Interprofessional collaboration to improve professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2017;6(6):CD000072. doi:10.1002/14651858.CD000072.pub3. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD000072.pub3/full

outcomes, and reproducibility events; and (2) a team-level **Public-Impact Trajectory** anchored in health-outcomes linkage, evidence uptake, community/practitioner engagement, and independent replication. Both tracks must be paired with narrative context, career-stage calibration, and an explicit anti-gaming and equity review — otherwise the reform risks recreating h-index harms in new dress.

1. Why this RFI matters now

The RFI is issued alongside NIH's Replication and Reproducibility Initiative and its Plan to Drive Gold Standard Science.[14][15] It names seven target areas: rigor and reproducibility; data, software, and model sharing; training and mentorship; collaboration; entrepreneurship and translation; foundational scientific exploration; and public impact.[16] Two decades of published evidence support the diagnosis:

- **Preclinical irreproducibility is well documented.** Begley and Ellis reported that only 11% of 53 landmark preclinical oncology findings could be reproduced by an experienced industry team; Ioannidis showed on statistical grounds that most published findings can be expected to be false under common design conditions.[17][18]
- **Individual metrics reward the wrong things.** The Leiden Manifesto and DORA both conclude that journal impact factors and citation counts are poor proxies for scientific quality and misalign incentives with research missions.[19][20]
- **Team science is the norm, not the exception.** The National Academies concluded in 2015, and reaffirmed in 2025, that biomedical research increasingly requires interdependent expertise across disciplines, sectors, and career stages — with attribution challenges that individual metrics cannot solve.[21][22]

[14]National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

[15]Bhattacharya J. NIH Strategic Vision and Plan for Strengthening Replication and Reproducibility. NIH Office of the Director briefing; 2026. https://dpcpsi.nih.gov/sites/g/files/mnhszr346/files/2026-01/1215PM-NIH-Update-Bhattacharya.pdf

[16]National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

[17]Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature. 2012;483:531-533. doi:10.1038/483531a. https://www.nature.com/articles/483531a

[18]Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

[19]Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015;520:429-431. https://www.nature.com/articles/520429a

[20]DORA. San Francisco Declaration on Research Assessment. 2013 (ongoing). https://sfdora.org/read/

[21]National Research Council. Enhancing the Effectiveness of Team Science. Washington, DC: National Academies Press; 2015. doi:10.17226/19007. https://www.nationalacademies.org/read/19007

[22]National Academies of Sciences, Engineering, and Medicine. The Science and Practice of Team Science. Washington, DC: National Academies Press; 2025. https://www.nationalacademies.org/read/29043/chapter/4

- **Public trust is fragile but recoverable.** Pew Research found that 76% of U.S. adults have a great deal or fair amount of confidence in scientists to act in the public's best interests in 2024 — up from pandemic lows, but with sustained concerns about communication, humility, and policy overreach.[23] Metrics that ignore public impact and transparency will not repair that trust.

2. Ethical grounding: measurement is a moral act

Choosing what to count is a value statement about what science is for. Three ethical commitments should govern any new NIH indicator set.

2.1 Human accountability before algorithmic ranking

Leiden Manifesto Principle 1 is unambiguous: quantitative evaluation should support — not supplant — qualitative, expert assessment.[24] Every proposed indicator should have a named human decision-maker responsible for its use, with authority to override the number when context demands. This mirrors interprofessional ethics: teams are accountable, algorithms are not.[25]

2.2 Transparency, reproducibility, and honest failure

NASEM's **Fostering Integrity in Research** frames research integrity as a shared responsibility across investigators, institutions, journals, and sponsors, and treats transparency as constitutive of rigor rather than an add-on.[26] Indicators must reward negative results, pre-registration, failed replications, and corrections — not only positive discoveries.

2.3 Equity across career stages, disciplines, and institution types

The RFI explicitly asks for indicators that are feasible across career stages, disciplines, and institution types.[27] Early-career investigators, minority-serving and rural institutions, community-engaged and clinical-education researchers, and interprofessional non-physician scientists (nurses, social workers, pharmacists, allied health) will be disproportionately harmed by any metric set that presumes MD/PhD principal-investigator norms.

---

[23]Pew Research Center. Public Trust in Scientists and Views on Their Role in Policymaking. November 14, 2024. https://www.pewresearch.org/science/2024/11/14/public-trust-in-scientists-and-views-on-their-role-in-policymaking/

[24]Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015;520:429-431. https://www.nature.com/articles/520429a

[25]Interprofessional Education Collaborative. IPEC Core Competencies for Interprofessional Collaborative Practice, Version 3. Washington, DC: IPEC; 2023. https://www.ipecollaborative.org/ipec-core-competencies

[26]National Academies of Sciences, Engineering, and Medicine. Fostering Integrity in Research. Washington, DC: National Academies Press; 2017. doi:10.17226/21896. https://nap.nationalacademies.org/catalog/21896/fostering-integrity-in-research

[27]National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

3. What "collaboration" must mean in an interprofessional biomedical enterprise

The RFI names collaboration as one of seven indicator areas but leaves the term open.[28] In modern biomedical research, collaboration is not simply co-authorship across departments; it is interprofessional and inter-sectoral, and it operates under a competency framework that the workforce already uses to teach and assess team behavior.

IPEC 2023 Core Competencies (Version 3) define four domains — values and ethics, roles and responsibilities, communication, and teams and teamwork — that shape how modern health-professions teams learn to work together and how team performance can be observed.[29] Cochrane evidence on practice-based interprofessional collaboration interventions shows small-to-modest improvements in adherence to practice guidelines, healthcare processes, and use of resources, with the strongest signal coming from structured team routines rather than one-off training.[30]

Implication for NIH. Team-science indicators should draw on this existing, validated competency architecture rather than invent parallel scales. Contribution attribution should be structured (CRediT), reproducibility should be team-attributable (not only PI-attributable), and collaboration quality should be observable through team routines, decision logs, and shared data/software deposits — not through survey self-report alone.[31][32]

4. Recommended measurable indicators, by RFI area

For each area, SIR proposes candidate indicators that satisfy the citation and evidence rigor standard: they are observable, verifiable, resistant to trivial gaming, and feasible for early-career and non-R1 investigators. Each is paired with a named benefit and a named unintended consequence to be actively monitored.

---

[28] National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

[29] Interprofessional Education Collaborative. IPEC Core Competencies for Interprofessional Collaborative Practice, Version 3. Washington, DC: IPEC; 2023. https://www.ipecollaborative.org/ipec-core-competencies

[30] Reeves S, Pelone F, Harrison R, Goldman J, Zwarenstein M. Interprofessional collaboration to improve professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2017;6(6):CD000072. doi:10.1002/14651858.CD000072.pub3. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD000072.pub3/full

[31] Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publishing. 2015;28(2):151-155. CRediT — Contributor Roles Taxonomy (ANSI/NISO Z39.104-2022). https://credit.niso.org/

[32] National Research Council. Enhancing the Effectiveness of Team Science. Washington, DC: National Academies Press; 2015. doi:10.17226/19007. https://www.nationalacademies.org/read/19007

RFI Area	Candidate indicator	Watch-out
Rigor & Reproducibility	Independent replication attempts documented per lab per 5-year period; pre-registration rate; open protocol deposit rate; corrections and retractions handled transparently.	Do not create incentives for shallow "confirmatory" replications; require methodological independence.
Data, Software & Model Sharing	FAIR-compliant deposit rate at time of publication; downstream reuse counts (data citations, package installs); computational reproducibility receipts.	Reuse counts favor mature fields — pair with narrative context; protect participant privacy in data-sharing metrics.
Training & Mentorship	Trainee independence outcomes at 5 and 10 years (grants, first-author work, retention in research); mentee-reported climate and integrity indicators; documented mentoring plans.	Do not penalize mentors whose trainees leave academia for equally valuable public-impact roles.
Collaboration	CRediT-based contribution portfolios; team composition diversity (disciplines, professions, institution types); documented team routines and decision logs.	Avoid rewarding "author inflation"; use CRediT roles rather than raw co-author counts.
Entrepreneurship & Translation	Verified translation events (licensed technology, guideline adoption, policy citation, deployed clinical decision-support); post-translation safety and equity monitoring.	Do not conflate patents filed with public health benefit delivered; require outcome linkage.
Foundational Exploration	High-risk portfolio share at institution level; time-to-first-independent-award for exploratory investigators; documented negative-result publication.	Protect small-lab and non-R1 exploratory work; do not concentrate high-risk funding in a few elite sites.
Public Impact	Guideline citation; measurable population health signal linked to funded lines of work; community/patient-partner co-authorship and governance; independent media and educator reuse.	Guard against political capture of "impact"; adopt bias review, not popularity, as a proxy.

5. A two-track measurement model

Track A — Contribution Portfolio (individual). An investigator-level, career-stage-calibrated portfolio combining verified CRediT roles across published outputs, verifiable data/software/model deposits, mentorship outcomes, independent replication events (both attempted and successful), and a narrative

that names contribution to team science and public impact.[3][3][4][3][5] Institutions review the portfolio holistically; NIH samples it for validation.

Track B — Public-Impact Trajectory (team/line of work). A team- or program-level trajectory that follows a research line into practice: guideline uptake, changes in clinical or policy standards, community-partner co-governance, patient-reported benefit where measurable, and equity of that benefit across populations.[3][6][3][7] The unit of analysis is the research question, not the individual.

Why two tracks. A single composite score would recreate the h-index harm the RFI is trying to escape. Two independent tracks preserve individual accountability and team accountability without collapsing them into each other. NIH review committees can weight the tracks differently by mechanism (fellowships weight Track A; centers and consortia weight Track B).

6. Named risks and unintended consequences

- **Goodhart's Law.** Every metric becomes a target. Any new indicator should be paired with a stated tolerance for gaming behavior and a scheduled review at 3 and 5 years, with authority to sunset indicators that no longer measure what they claim.
- **Equity harms.** Reproducibility investment favors well-resourced labs. Data-sharing infrastructure favors large centers. Translation metrics favor commercially attractive lines of work. Each area needs an equity offset that protects early-career, minority-serving, rural, and non-physician-led science.[3][8]
- **Political capture of "public impact."** "Public impact" can be defined narrowly to reward politically convenient work. NIH should bind public-impact indicators to independent scientific and

[3][3]Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publishing. 2015;28(2):151-155. CRediT — Contributor Roles Taxonomy (ANSI/NISO Z39.104-2022). https://credit.niso.org/

[3][4]Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi:10.1038/sdata.2016.18. https://www.nature.com/articles/sdata201618

[3][5]National Institutes of Health. Final NIH Policy for Data Management and Sharing. Effective January 25, 2023. https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/dms/policy-overview

[3][6]National Academies of Sciences, Engineering, and Medicine. The Science and Practice of Team Science. Washington, DC: National Academies Press; 2025. https://www.nationalacademies.org/read/29043/chapter/4

[3][7]Pew Research Center. Public Trust in Scientists and Views on Their Role in Policymaking. November 14, 2024. https://www.pewresearch.org/science/2024/11/14/public-trust-in-scientists-and-views-on-their-role-in-policymaking/

[3][8]National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

community review, not to the sitting administration's policy agenda. This is a durable-institution question, not a partisan one.[3940]

- **Erosion of foundational exploration.** Basic and high-risk work has slow, uncertain public impact by design. Any indicator system that penalizes slow trajectories will hollow out the pipeline that produced the discoveries the RFI now wants to reward.[4142]
- **AI-enabled metric inflation.** Generative tools can now produce plausible-looking outputs, citations, and contribution claims. NIH should require verifiable deposit and identifier-resolvable attribution (ORCID, ROR, DOI) before any indicator is credited, and should treat unverifiable AI-authored outputs as unattributed.[4344]

7. Implementation approach

- **Pilot before mandate.** Pilot the two-track model in 2–3 ICs across mechanism types (R01, K, U54, T32) for a full cycle before scaling.[45]
- **Publish an anti-gaming register.** For each indicator, publish the known gaming behaviors, the monitoring plan, and the sunset trigger. This inverts the h-index history in which harms were named only after damage.
- **Anchor in existing standards.** Use CRediT (ANSI/NISO Z39.104-2022), FAIR principles, and the existing Data Management and Sharing Policy as the technical backbone.[464748]

---

[39]Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015;520:429-431. https://www.nature.com/articles/520429a

[40]National Academies of Sciences, Engineering, and Medicine. Fostering Integrity in Research. Washington, DC: National Academies Press; 2017. doi:10.17226/21896. https://nap.nationalacademies.org/catalog/21896/fostering-integrity-in-research

[41]National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

[42]National Academies of Sciences, Engineering, and Medicine. The Science and Practice of Team Science. Washington, DC: National Academies Press; 2025. https://www.nationalacademies.org/read/29043/chapter/4

[43]Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publishing. 2015;28(2):151-155. CRediT — Contributor Roles Taxonomy (ANSI/NISO Z39.104-2022). https://credit.niso.org/

[44]Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi:10.1038/sdata.2016.18. https://www.nature.com/articles/sdata201618

[45]National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

[46]Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publishing. 2015;28(2):151-155. CRediT — Contributor Roles Taxonomy (ANSI/NISO Z39.104-2022). https://credit.niso.org/

[47]Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi:10.1038/sdata.2016.18. https://www.nature.com/articles/sdata201618

[48]National Institutes of Health. Final NIH Policy for Data Management and Sharing. Effective January 25, 2023. https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/dms/policy-overview

- **Interprofessional review panels.** Include nursing, social work, pharmacy, allied-health, patient/community, and non-physician scientific reviewers on panels that assess team-science and public-impact indicators. IPEC competencies can structure that review.[4950]
- **Sunset the composite.** Do not publish a single institutional or investigator ranking derived from the new indicators. The Leiden Manifesto's warning against composite league tables applies with special force here.[51]

8. Direct response to the RFI's core questions

Specific measurable indicators: See Section 4 table and the two-track model in Section 5.

Implementation approaches: Section 7. Pilot, anti-gaming register, anchor in CRediT/FAIR/DMS policy, interprofessional panels, no composite ranking.

Potential benefits: Better alignment between reward structure and NIH mission; explicit recognition of team-based and interprofessional contributions; verifiable, sharable evidence of rigor; recovered trust through transparent public-impact linkage.[52]

Unintended consequences: Section 6. Goodhart's Law, equity harms, political capture, erosion of foundational exploration, AI-enabled metric inflation.

Feasibility across career stages, disciplines, and institution types: Two-track separation lets NIH weight tracks by mechanism, protecting early-career, non-R1, community-engaged, and interprofessional-non-physician investigators from being scored on the same rubric as senior R01 PIs.[535455]

9. Closing frame

[49]Interprofessional Education Collaborative. IPEC Core Competencies for Interprofessional Collaborative Practice, Version 3. Washington, DC: IPEC; 2023. https://www.ipecollaborative.org/ipec-core-competencies

[50]Reeves S, Pelone F, Harrison R, Goldman J, Zwarenstein M. Interprofessional collaboration to improve professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2017;6(6):CD000072. doi:10.1002/14651858.CD000072.pub3. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD000072.pub3/full

[51]Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015;520:429-431. https://www.nature.com/articles/520429a

[52]Pew Research Center. Public Trust in Scientists and Views on Their Role in Policymaking. November 14, 2024. https://www.pewresearch.org/science/2024/11/14/public-trust-in-scientists-and-views-on-their-role-in-policymaking/

[53]National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

[54]National Research Council. Enhancing the Effectiveness of Team Science. Washington, DC: National Academies Press; 2015. doi:10.17226/19007. https://www.nationalacademies.org/read/19007

[55]National Academies of Sciences, Engineering, and Medicine. The Science and Practice of Team Science. Washington, DC: National Academies Press; 2025. https://www.nationalacademies.org/read/29043/chapter/4

Measurement in science is not administrative housekeeping. It is a public covenant about what we agree to reward. If NIH measures narrowly, it will teach the next generation to work narrowly. If NIH measures well — rigor, transparency, team contribution, mentorship, honest failure, and verifiable public impact — it will restore what the h-index era eroded. That is the reform this RFI has the chance to deliver, if the new indicators are paired from day one with the humility to sunset them when they stop serving the mission.

Author note

Author. Annette Greer, PhD, MSN, RN, FNAP. Founder, Scholarly Integrity Review LLC. Faculty, Brody School of Medicine at East Carolina University.

Capacity. This SIR Briefing is submitted in personal capacity. ECU/BSOM affiliation is used only for identification; the response is not an institutional ECU or BSOM comment.

Competing interests. The author leads Scholarly Integrity Review LLC, which reviews citations, sources, methods, and manuscripts for academic clients, and Integrity Document Services. No NIH funding was used to prepare this briefing; no ECU library or database access was used to anchor claims in this document.

Method. Sources were drawn from primary NIH notices, National Academies consensus reports, peer-reviewed scholarship, and standards-body documents. Each load-bearing claim is footnoted to an identifier-resolvable source. Draft passed SIR internal audit x 3: (1) source verification, (2) bias review, (3) audience-impact assessment.

References

Consolidated reference list (AMA numbered). Each source is footnoted at each in-text use; this list is the master.

1. National Institutes of Health. Request for Information on Measuring and Rewarding Scientific Impact. Notice NOT-OD-26-087. Released June 18, 2026. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

2. Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015;520:429-431. https://www.nature.com/articles/520429a

3. DORA. San Francisco Declaration on Research Assessment. 2013 (ongoing). https://sfdora.org/read/

4. Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publishing. 2015;28(2):151-155. CRediT — Contributor Roles Taxonomy (ANSI/NISO Z39.104-2022). https://credit.niso.org/

5. National Research Council. Enhancing the Effectiveness of Team Science. Washington, DC: National Academies Press; 2015. doi:10.17226/19007. https://www.nationalacademies.org/read/19007

6. National Academies of Sciences, Engineering, and Medicine. The Science and Practice of Team Science. Washington, DC: National Academies Press; 2025. https://www.nationalacademies.org/read/29043/chapter/4

7. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

8. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature. 2012;483:531-533. doi:10.1038/483531a. https://www.nature.com/articles/483531a

9. National Academies of Sciences, Engineering, and Medicine. Fostering Integrity in Research. Washington, DC: National Academies Press; 2017. doi:10.17226/21896. https://nap.nationalacademies.org/catalog/21896/fostering-integrity-in-research

10. Bhattacharya J. NIH Strategic Vision and Plan for Strengthening Replication and Reproducibility. NIH Office of the Director briefing; 2026. https://dpcpsi.nih.gov/sites/g/files/mnhszr346/files/2026-01/1215PM-NIH-Update-Bhattacharya.pdf

11. Interprofessional Education Collaborative. IPEC Core Competencies for Interprofessional Collaborative Practice, Version 3. Washington, DC: IPEC; 2023. https://www.ipecollaborative.org/ipec-core-competencies

12. Reeves S, Pelone F, Harrison R, Goldman J, Zwarenstein M. Interprofessional collaboration to improve professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2017;6(6):CD000072. doi:10.1002/14651858.CD000072.pub3. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD000072.pub3/full

13. Pew Research Center. Public Trust in Scientists and Views on Their Role in Policymaking. November 14, 2024. https://www.pewresearch.org/science/2024/11/14/public-trust-in-scientists-and-views-on-their-role-in-policymaking/

14. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi:10.1038/sdata.2016.18. https://www.nature.com/articles/sdata201618

15. National Institutes of Health. Final NIH Policy for Data Management and Sharing. Effective January 25, 2023. https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/dms/policy-overview

AI Transparency Statement

This SIR Briefing was drafted with AI-assisted research and structure (Perplexity Computer, Claude Sonnet 4.6) under human direction. Source selection, argument, ethical framing, and all interpretive claims are human-authored by Annette Greer. AI was used for literature scanning, structural drafting, and citation formatting. All footnoted sources were verified against primary URLs. Any error is the author's.

147. Alzheimer's Association

I am responding: On behalf of an organization

Name: Laura Thornhill

Name of Organization: Alzheimer's Association

Type of Organization: Other

Type of Organization - Other: voluntary health association

Role: Other

Role - Other: Director, Regulatory Affairs

Date Submitted: 8/11/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Alz-Assn-Responses-to-NIH-RFI-on-Rewarding-Scientific-Impact.pdf

Description: Alzheimer's Association Responses to NIH RFI on Measuring and Rewarding Scientific Impact

Jay Bhattacharya, MD, PhD
Director
National Institutes of Health
9000 Rockville Pike
Bethesda, Maryland 20892

August 10, 2026

Re: Request for Information on Measuring and Rewarding Scientific Impact (Notice Number: NOT-OD-26-087)

Dear Dr. Bhattacharya,

The Alzheimer's Association and the Alzheimer's Impact Movement (AIM) appreciate the opportunity to respond to the Request for Information on Measuring and Rewarding Scientific Impact.

The Alzheimer's Association is the leading voluntary health organization in Alzheimer's care, support, and research. The Alzheimer's Impact Movement, a separately incorporated advocacy affiliate, is a nonpartisan, nonprofit organization and works to make Alzheimer's a national priority.

Today there are an estimated 8.6 million Americans with the biomarkers of Alzheimer's disease and symptoms, including 4.4 million Americans with dementia due to Alzheimer's and 4.2 million with mild cognitive impairment (MCI) due to Alzheimer's.[1] Yet only two treatments are available that change the underlying biology of Alzheimer's and slow cognitive decline, and the treatments are only for people with mild cognitive impairment or mild dementia. People with moderate or severe Alzheimer's dementia have no therapeutic options, save drugs to lessen symptoms. Furthermore, research advances have made it possible to detect Alzheimer's biomarkers in individuals without symptoms - which might be the most effective time to intervene - but they also have no therapeutic options. Therefore, the Alzheimer's Association and AIM applaud the National Institutes of Health's (NIH) efforts to drive new incentives toward scientific progress.

The Alzheimer's Association and AIM fully support the expansion of the types of non-traditional metrics proposed in concept. The exact metrics, however, may prove challenging to develop, and the evaluation of these contributions must be balanced. For example, teaching an online course is not the same as an in-person course in a medical school lab; how is that evaluated? If a metric is the number of individuals trained for the workforce and they subsequently leave the field, how is that evaluated? How are shared datasets and subsets valued? We would be pleased to work with NIH to explore these issues and ultimately drive scientific progress.

[1] Alzheimer's Association. (2026). *2026 Alzheimer's Disease Facts and Figures.*

Thank you for the opportunity to comment. The Alzheimer's Association and AIM would be glad to serve as a resource to NIH as it considers these important issues and how they relate to individuals living with Alzheimer's and related dementias. Please contact Laura Thornhill, Director, Regulatory Affairs, at 202-638-7042 or lthornhill@alz-aim.org if you have questions or if we can be of additional assistance.

Sincerely,





Robert Egge
Chief Public Policy Officer, Alzheimer's Association
President, Alzheimer's Impact Movement

Maria C. Carrillo, Ph.D.
Chief Science Officer, Alzheimer's Association

148. Jingyi Jessica Li

I am responding: On behalf of myself

Name: Jingyi Jessica Li

Name of Organization: Fred Hutchinson Cancer Center

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/11/2026

1. Rigor and Reproducibility: Rigor and reproducibility in genomics need to separate two things that current NIH metrics tend to blur: statistical reproducibility (same data, same code, same result) and biological replicability (independent data, same conclusion). Single-cell genomics is especially vulnerable here because upstream choices like normalization, batch correction, and dimensionality reduction can silently change downstream findings without anyone noticing. NIH's push for gold standard science would benefit from treating these as distinct, separately measurable indicators rather than folding them into a single "reproducibility" bucket.

On causal research specifically, rigor should be defined by whether a paper explicitly states its causal assumptions (no unmeasured confounding, faithfulness, etc.) rather than by which algorithm it uses. Work applying causal discovery to single-cell perturbation data shows that causal claims from observational genomic data require sensitivity analyses and transparent assumptions, not just methodological novelty. A concrete indicator NIH could adopt is something like an "assumption disclosure checklist," similar to CONSORT or STROBE guidelines but built for computational and causal methods.

A few other measurable indicators worth proposing:

Require papers to report whether their false discovery rate control has any theoretical guarantee, rather than relying on nominal p value thresholds. Many high throughput pipelines currently lack this, and p value free frameworks like Clipper show it's possible to validate FDR control without assuming a specific data distribution.

Track whether code and data pipelines are version pinned and containerized (tagged Git releases, fixed random seeds), since single-cell integration methods are highly sensitive to software drift.

Create a "replication readiness score" combining code and data availability, documented random seeds, benchmarking against simulated ground truth, and pre-registration for hypothesis driven studies.

As AI tools become embedded in every stage of biomedical research, from literature synthesis to code generation to result interpretation, NIH needs explicit guidelines for when and how AI outputs can be trusted rather than treating them as neutral productivity gains. A useful evaluation framework would

require researchers to disclose which parts of an analysis involved AI assistance, benchmark AI generated code or annotations against a known ground truth before relying on them, and report an error rate rather than assuming correctness. For single-cell genomics specifically, AI based cell type annotation or gene function prediction tools should be evaluated the same way any other statistical method is evaluated, meaning held out validation, calibration checks, and sensitivity to input perturbations, rather than accepted purely because a large model produced them. NIH could incentivize this by funding independent benchmarking of AI tools used in biomedical pipelines, similar to how new statistical methods are benchmarked, and by requiring grant applications and publications to state what human verification step was applied to any AI generated result before it entered the paper.

On incentives and infrastructure, NIH should fund benchmarking and replication work as its own grant category, not just novel method development. Right now, careful benchmarking papers are undervalued relative to flashy new methods, even though they're what actually let the field trust a result. Supporting shared, well characterized reference datasets with known ground truth would also help, since realistic simulators are essential for testing whether a method's reproducibility claims actually hold before publication. Tenure and grant review criteria should give explicit credit for benchmarking papers, replication studies, and negative results, which is really the core ask behind NIH moving past raw publication and citation counts.

On feasibility, early career researchers and smaller institutions often don't have the compute or data infrastructure to run large scale replication studies. NIH could offset this with shared cloud compute credits or consortium access earmarked for reproducibility verification work. Reproducibility and benchmarking work is also time intensive and mostly invisible on the tenure clock right now, so study sections and review criteria need to treat it as a real scientific contribution rather than service work.

2. Data, Software, and Model Sharing: Data, software, and model sharing in computational biology needs incentives that go beyond simply posting code to a repository, because a public GitHub link with no maintenance, documentation, or usage tracking does little to advance reproducibility. Statistical software for genomics, like FDR control frameworks or single-cell analysis tools, only delivers value if other researchers can actually install it, run it on their own data, and get comparable results. That means the measurable indicators NIH adopts should capture usability and maintenance, not just existence.

A few concrete indicators worth proposing:

Track software usage and adoption through metrics like downloads, Bioconductor or CRAN install counts, or citations of the software itself as distinct from citations of the associated paper. Right now a method paper gets citation credit while the underlying tool, which took far more sustained effort to build and maintain, gets none.

Require a minimal reproducibility package for any tool tied to a publication, meaning example data, a working vignette, and version pinned dependencies, so a new user can reproduce the paper's core result within a fixed time budget, say under an hour.

Score data sharing not just on deposit but on documentation completeness, meaning whether metadata, processing scripts, and quality control steps are included alongside raw data. A dataset dropped into a

repository without its generating code is much less useful for validation than one with a full computational trail.

For models specifically, including statistical or machine learning models trained on genomic data, require versioned model releases with reported performance on held out or independent benchmark datasets, not just performance on the training cohort.

On incentives, NIH should treat long term software maintenance as a fundable activity in its own right, separate from new method development grants. Maintaining a widely used single-cell analysis tool for five years after publication is real scientific labor, yet it is currently invisible in grant renewals and promotion packets. NIH could create a distinct award mechanism, similar to how some funders support research software engineers, that explicitly funds bug fixes, dependency updates, and user support rather than only new algorithmic contributions.

NIH should also recognize data and software curators and maintainers as contributing authors or acknowledge them through a formal contribution taxonomy, similar to CRediT, so that promotion and tenure committees can see this work reflected in a researcher's record. Institutions vary widely in how they weigh software contributions, so NIH endorsing a standard contribution statement across grant applications and biosketches would help normalize this across career stages and institution types.

A potential unintended consequence is that overly rigid sharing requirements could disadvantage early career researchers or smaller labs without dedicated engineering support, so any mandate should be paired with infrastructure funding, such as shared institutional repositories or centralized deposition services, rather than pushing the full compliance burden onto individual investigators. Large data intensive genomics studies also raise data sharing complications around patient privacy and consent scope, so indicators for data sharing need to allow for controlled access tiers rather than assuming all data can or should be fully public.

Finally, NIH could fund the creation of well characterized benchmark datasets with known ground truth for common genomic tasks, since having a shared gold standard is what actually lets independent groups verify that a tool or model works as claimed, rather than relying on each paper's own internal validation.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

149. Essex Management LLC

I am responding: On behalf of an organization

Name: Samuel Isa

Name of Organization: Essex Management LLC

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Organizational Official

Date Submitted: 8/11/2026

1. Rigor and Reproducibility: Rigor and Reproducibility

Reproducible and replicable research is the operational foundation upon which the biomedical research enterprise derives its credibility and public value. NIH's Replication and Reproducibility Initiative arrives at a critical inflection point, one in which the complexity and scale of modern biomedical research have outpaced the infrastructure, incentives, and norms required to sustain it. The following comments draw on direct operational experience supporting large-scale NCI data-generating programs, where reproducibility planning, validation frameworks, and quality assurance are contractual requirements and critical to support the NIH's mission. We offer specific, implementable recommendations across four areas: computational standards, treatment of negative and replication studies, metrics and recognition structures, and dedicated funding mechanisms.

Computational Reproducibility Standards

The reproducibility crisis in biomedical science is, in substantial part, a computational reproducibility crisis. The RFI also asks about causal rigor, including standards for causal inference and study design; these merit parallel NIH attention and are outside the scope of these four recommendations, which focus on computational reproducibility specifically. Analyses that cannot be re-executed because of undocumented software environments, proprietary dependencies, or absent version control cannot be validated, replicated, or built upon. NIH should address this directly and systematically with the four following steps.

First, NIH should require computational pipeline validation plans as a deliverable for all data-generating grants, analogous to the existing Data Management and Sharing Plan requirement. These plans should specify pre-defined acceptance thresholds for key quality metrics, document the software environment (including version-pinned dependencies), describe the validation dataset and approach, and define the human review process for automated outputs. Without such a plan, there is no systematic mechanism to verify that an analysis can be independently re-executed, that results reported at publication match what the pipeline produces, or that automated outputs have undergone human review prior to dissemination. In our experience supporting federal informatics programs, the development of a formal bioinformatic pipeline validation plan, including defined acceptance criteria, structured testing against a mock dataset, and human review of flagged outputs was essential to building confidence in novel

analysis products. This model to define criteria, validate against a gold standard or acceptance threshold, and document the process is broadly applicable and should be a standard deliverable, not an optional practice.

Second, NIH should invest in centralized validation resources and reference standards for high-priority assay types and computational methods, modeled on programs like ENCODE or Genome in a Bottle. These initiatives produced validated reference datasets and benchmarking protocols that individual research groups could adopt without recreating the validation infrastructure from scratch. The absence of such resources forces independent groups to develop their own gold standards, causing a duplication of effort that introduces variability across studies. In our own bioinformatics work, developing and validating novel pipelines for emerging technologies, such as spatial transcriptomics, required substantial investment in constructing validation benchmarks simply because the fundamental resources do not exist for this type of work. NIH-funded, community-maintained validation resources of this kind would lower the barrier for reproducible pipeline development across the field.

Third, NIH should require a documented, reproducible environment specification as a condition of data deposition and publication for computational and AI-based research. Containerized environments such as Docker or Singularity are the most reliable implementation for most work and should be the default. NIH should define approved alternatives for cases where containerization is not feasible, including work involving controlled-access data, licensed or proprietary software dependencies, export-controlled materials, or long-term maintenance constraints, provided the alternative achieves equivalent reproducibility. The environment in which an analysis was run is as much a part of the methods as the algorithm itself; without a documented specification, computational reproducibility is aspirational rather than achievable.

Finally, NIH should consider piloting concordance-based submission standards for data deposition into shared repositories, analogous to the platform concordance testing used in clinical trial settings. Under this model, research groups seeking to contribute data of a specific type to an NIH data common would validate their pipelines against a standard reference set. Concordance with defined thresholds would be a prerequisite for submission. This would establish baseline comparability across contributed datasets, a prerequisite for meaningful meta-analysis and cross-study replication.

Treatment of Negative and Replication Studies

The current incentive structure penalizes researchers who pursue replication studies or report null results, regardless of the scientific rigor with which those studies were conducted. This creates a systematic bias in published literature and undermines the self-correcting function of science. NIH should act on several fronts. NIH should explicitly recognize and reward investigators who publish rigorous replication and null-result studies during grant renewals and progress report evaluations. A well-conducted replication study that confirms or fails to confirm a high-impact finding is a valuable contribution to the field and should be treated as such. Providing equal weight to null and replication findings, rather than treating them as evidence of unproductivity, would shift researcher incentives in a meaningful direction. NIH should work with journals and preprint servers to establish structured pathways for publishing null replication findings, including standardized reporting formats that make

these studies discoverable and citable. The absence of such pathways means that failed replications frequently go unpublished, creating an invisible correction record that cannot be accessed by the broader community.

NIH should also consider funding standing replication cohorts; small, continuously funded teams whose explicit mission is to independently reproduce high-impact findings in priority disease areas. Success metrics for these teams should be defined by completion rate and concordance, not by novel publication output. This would provide a systematic, rather than ad hoc, mechanism for evaluating the replicability of landmark findings before they advance to clinical translation.

Metrics and Recognition Structures

Metrics shape behavior. If NIH wants reproducible research, it must measure and reward the practices that produce it.

NIH should develop and track a reproducibility practice index for funded investigators with structured measures of whether key reproducibility practices were followed; protocol preregistration, version-controlled code availability, standardized SOPs, complete methods reporting, and validated data QC thresholds. This index could function similarly to an H-index in concept but would track rigor practices rather than citation counts and could be factored into competitive renewal scoring.

NIH should introduce "Reproducibility Badges" on Biosketches as a formal NIH designation for investigators whose published protocols have been independently verified or whose computational pipelines have been successfully re-executed by a third party. This creates a visible, career-relevant signal that rigor is recognized and valued, particularly important for early-career investigators who face pressure to prioritize novelty over reproducibility. It is also the natural complement to the replication cohort concept and incentivizes to seek third party verification. NIH should ensure that badge criteria are attainable across institution types, including smaller institutions and disciplines without mature, shared validation infrastructure, and should provide implementation guidance and tooling support to prevent the program from inadvertently advantaging well-resourced groups.

The NIH should consider tracking independent validation of published findings, such as external reproductions, successful cross-site implementations, and secondary analyses using deposited data as a distinct positive impact metric separate from citation counts. A finding that has been independently validated across multiple sites and datasets represents a qualitatively different and higher level of scientific confidence than one that has been frequently cited but not yet reproduced. For early-career investigators, documenting independent validation of their published methods and findings during tenure review and career advancement processes would create a meaningful incentive structure at the institutional level.

Finally, NIH should incentivize the creation of reusable research infrastructure, for example, validated assays, reference datasets, and standardized workflows as a scoreable output. Grants that produce infrastructure enabling others to reproduce findings should receive scoring recognition, and dedicated funding mechanisms (e.g., R21-scale awards) should be available specifically for infrastructure

development of this kind. In our experience, SOP-driven quality frameworks such as the comprehensive suite of validated operating procedures governing clinical trial abstraction and quality assurance, provide exactly this kind of reusable infrastructure; documented, version-controlled, and subject to annual review. The value of that infrastructure compounds across the programs that adopt and adapt it.

There is established precedent within the federal ecosystem for holding recipients of government funds accountable for the quality of their work through recurring, on-the-record evaluation. Through the Contractor Performance Assessment Report (CPAR) system, federal contractors are formally rated on performance at regular intervals, and those ratings persist as part of the entity's record, following the organization across future funding opportunities. This creates a durable, portable accountability signal; consistent quality is rewarded with continued opportunity, and lapses in quality carry forward-looking consequences. NIH should consider extending an analogous principle to the reproducibility and rigor of federally funded research. A recurring, on-the-record assessment of an investigator's or institution's adherence to reproducibility practices that persist as part of their funding record rather than resetting with each new award would align incentives with long-term scientific reliability. The key features worth adopting are the periodic cadence, the persistence of the record over the life of the funded entity, and the direct linkage between demonstrated rigor and future funding competitiveness.

Dedicated Funding Mechanisms

Two funding mechanisms merit specific consideration. First, NIH should establish Independent Replication Supplements. Dedicated supplemental awards to pay independent, third-party laboratories to replicate landmark or high-impact findings before those findings advance to clinical stages or inform major funding priorities. This would create a systematic pre-translation check without requiring primary investigators to shoulder the replication burden.

Second, NIH should consider Institutional Replication Core incentives. Overhead or infrastructure funding for universities and research institutions that establish and maintain dedicated internal replication cores. These cores would audit and validate high-stakes findings internally before public dissemination, providing a first-line quality check that benefits both the investigator and the broader field, and creating institutional ownership of reproducibility as a value, not just a mandate.

2. Data, Software, and Model Sharing: Data, Software, and Model Sharing

NIH's investment in data sharing has produced extensive repository infrastructure and a policy framework that, on paper, requires investigators to make their data available. What it has not consistently produced is reuse. The gap between deposition and reuse is where the return on NIH's data investment is being lost, and closing it requires action on four fronts: measuring what actually matters, governing for reusability from the point of study design rather than the point of deposit, reducing the regulatory complexity that drives investigators toward conservatism, and funding the people and infrastructure that make shared resources usable over time.

These comments draw on operational experience building and running data sharing pipelines and harmonization infrastructure for NIAID and NCI programs and on the work established specifically to help federal clients navigate the regulatory environment governing biomedical data. We reference this

experience because several of the recommendations below arise directly from repeatedly encountering the same issues across programs.

1. Measure Reuse, Not Deposition

The single most consequential change NIH could make is to shift its measurement of data sharing success from a count of deposits to a measure of downstream reuse. A deposited dataset that no one accesses represents a compliance event, not a scientific contribution.

Develop a Data Sharing Impact Score. NIH should track the metrics that reflect actual scientific value: the number of derivative publications using a deposited dataset, the number of unique investigators who have accessed it, the number of independent replication or validation attempts it enabled, and the number of downstream studies that incorporated it into larger analyses. This score should appear on investigator profiles and be factored into competitive renewal scoring.

Weight external and independent reuse more heavily than internal reuse. A dataset used exclusively by the generating laboratory and its immediate collaborators has not entered the scientific commons in any meaningful sense. The measure of a dataset's value is whether investigators with no prior connection to the generating team can find it, understand it, and use it. Reuse metrics must distinguish clearly between these cases, for example, by factoring generators' and users' geographical location and co-authorship history.

Credit widely adopted software and models on the same basis. Investigators whose shared tools achieve broad adoption by other NIH-funded projects or clinical research networks have delivered substantial scientific value. A computational method adopted by fifty downstream studies has arguably contributed more than a well-cited but unused publication, yet the current incentive structure recognizes only the latter.

Two mechanisms make this tractable using infrastructure that already exists:

Elevate data and software citations to parity with journal citations. Persistent identifiers for datasets and open-source software are already well-supported. NIH should treat citations tracked through these identifiers with the same administrative weight as traditional publication citations in the NIH Biosketch and in review. This is a low-cost policy change with immediate signaling value.

Establish a Data Utility Index tied to future funding eligibility. This index would measure how frequently an investigator's previously funded datasets are actively used in external, independent publications. Tying a portion of renewal competitiveness to demonstrate utility of prior data outputs would create a direct and durable incentive to produce data others can use.

Anticipate gaming. Any reuse-based metric will be gamed if implemented naively. Reuse counts inflated by self-citation, by consortium partners, or by trivially derivative analyses would undermine the metric's validity. The mitigations are straightforward but must be designed from the start: weight for independence from the generating team, require that reuse be traceable to a persistent identifier rather

than self-reported, and audit a sample of claimed reuse events. NIH should also avoid disadvantaging investigators in fields where the potential for reuse, for example due to low number of researchers, is inherently small. The metric should be normalized against field-level or data-type-level baselines rather than applied as a raw count.

2. Govern for Reusability at the Point of Design, Not the Point of Deposit

The most expensive moment to harmonize data is after it has been collected. The cheapest is before it is collected at all. Current NIH policy engages data sharing primarily at the back end of the research lifecycle, through Data Management and Sharing Plans reviewed for compliance and deposition requirements enforced at publication. This is backwards from both a cost and a quality standpoint. By the time data arrives at a repository in a nonstandard format, the harmonization required to make it interoperable is labor-intensive, error-prone, and often scientifically lossy.

Move common data elements from the back end to the front end. NIH has invested substantially in CDE infrastructure through the NIH CDE Repository and, in the cancer domain, through Cancer Data Standards Registry and Repository (caDSR). That investment is not being fully realized, for two reasons. First, CDE adoption is treated as an optional best practice rather than a design requirement. Second, the practical barrier to adoption is high: with tens of thousands of CDEs available, no investigator can reasonably be expected to know which element best fits a given data capture need. Placing that navigation burden on the individual researcher guarantees inconsistent adoption.

Two changes would address this directly.

Fund tooling that makes CDE selection and mapping easy enough to become default behavior. This is a solvable technical problem. For example, sentence embedding models accurately and rapidly map source data elements to appropriate CDEs. Thus, the barrier to CDE adoption is a tooling gap, not a fundamental scientific one, and NIH is well positioned to close it.

Align clinical trial funding expectations with use of approved, community-developed CDEs where fit-for-purpose, enforced through the statistical analysis plan review process, with documented exceptions where no appropriate element exists or where domain-specific standards apply. For clinical research, the enforcement mechanism already exists. Statistical analysis plans are reviewed and approved by governing bodies, and data collection standards can be reviewed at that same gate. If a study's data capture instruments are required to use approved CDEs as a condition of protocol approval, the data arrives in a reusable format by construction and costly retrospective harmonization becomes unnecessary. NIH should pilot this requirement in a defined program area and measure the downstream reduction in harmonization cost, which we expect would be substantial.

Establish governance parity for software and model artifacts. Current DMS Plan requirements are oriented toward datasets and largely do not address model artifacts. As machine learning methods become central to biomedical discovery, this is an increasingly consequential gap. A published model with no documented training data provenance, no version-controlled code repository, and no held-out validation benchmark cannot be evaluated, reproduced, or responsibly reused. NIH should require, for

any model published with NIH support: documentation of training data provenance, a version-controlled and publicly accessible code repository, and at minimum one held-out validation benchmark reported against a defined test set. The discipline is not burdensome in our experience when adopted at the outset, but in practice, it is nearly impossible to retrofit.

Add a data sharing readiness review at grant review. Before a data-generating study is funded, reviewers should assess whether the proposed data model, formats, and deposition target are realistic for reuse by investigators outside the generating team. This is a distinct question from whether the DMS Plan is technically compliant. A study that plans to deposit a nonstandard, sparsely documented dataset into a general-purpose repository can be fully compliant and still not promote data reuse.

Verify data sharing rather than accepting attestation. Annual grant progress reviews should include data sharing progress, verified rather than self-reported. Reporting a dbGaP accession number or BioProject ID is evidence that a record was created, not that data are accessible. NIH should implement verification mechanisms, such as cross-checking repository access records and controlled-access approval lists against grantee reports, to confirm that data have actually been released and are actually reachable by qualified external requestors who have completed applicable data use agreements. We believe the gap between reported and actual accessibility is, in our operational experience, meaningful and underappreciated.

3. Reduce the Regulatory Complexity That Suppresses Sharing

Regulations governing patient-derived data are essential. They protect participant privacy, honor the terms under which participants consented, and sustain public trust in biomedical research. They are also, in aggregate, extraordinarily difficult to navigate, and that difficulty is itself suppressing the reuse of exactly the high-value datasets NIH most wants reused.

We want to be precise about the nature of the problem, because it is frequently mischaracterized. The issue is not simply that there are many rules. It is that the applicable authority stack is scenario-dependent, the scenarios are not enumerated anywhere authoritative, and the prioritization among conflicting authorities is not documented. In our experience, the correct authority stack for a given dataset depends on a set of distinctions that are not obvious and are frequently not known to the investigator:

Funding pathway. Data generated under NIH funding, data generated externally but analyzed under NIH funding, data generated under NIH funding but analyzed externally, and data generated in collaboration with NIH under a services-in-kind agreement are each governed differently. In the mixed cases, which award's expectations take precedence is a real question with a real answer, and it is not intuitive.

Generation versus analysis. An award covering analysis carries expectations that may differ from those attached to the award that generated the data, and both may apply in sequence.

Repository ownership. Whether a system is operated on behalf of NIH, qualifies as an NIH Controlled Access Data Repository, or is external to NIH determines an entirely different compliance regime, with

different security expectations and different review processes for sensitive data.

Requestor role. Access rules differ materially for senior-level PIs, non-senior investigators and lab members, students and citizen scientists, developers supporting a data system under NIH funding, and NIH program staff exercising oversight. Each of these categories has a distinct approval pathway.

Institutional and national location. Collaboration with investigators at the same institution, at a different U.S. institution, and at an international institution each trigger different requirements, with additional restrictions applying to entities based in countries of concern.

Emerging product types. Tools and AI models used for scientific management or reuse, animal models, and biospecimens raise authority questions that existing policy addresses incompletely.

The behavioral consequence of this complexity is predictable and damaging. When a researcher is uncertain whether they are permitted to share a dataset, the rational default is to not share it. Regulatory ambiguity produces conservatism, and conservatism suppresses reuse. NIH is very likely losing more data sharing to uncertainty than to actual prohibition. We would encourage NIH to treat this as a first-order priority rather than a background condition, and we offer two recommendations.

Sponsor an authoritative, maintained decision-support tool for data use authorities. A researcher should be able to describe the characteristics of their dataset and their intended use (data type, funding pathway, consent scope, repository, requestor role, recipient institution and location) and receive a clear determination of which frameworks apply and what adherence requires. This tool should be developed, it should be authoritative, and it should be centrally maintained rather than reconstructed independently by every institution and program that needs one. The alternative, which is the status quo, is that thousands of investigators each make an individual, uninformed, risk-averse judgment.

Pursue holistic, cross-agency harmonization of data sharing authorities. A decision-support tool makes a complex environment navigable, but it does not make it simpler. The deeper fix is for NIH, in coordination with other HHS components and relevant regulatory bodies, to systematically review where data sharing authorities overlap, conflict, or impose duplicative requirements without corresponding protection benefit, and to consolidate where possible. Consent-scope constraints on retroactive sharing of existing datasets deserve particular attention: a great deal of high-value legacy data is effectively stranded by consent language written before secondary reuse was contemplated, and NIH should develop clear, generalizable guidance on what is and is not permissible in these cases rather than leaving each institution to reach its own conservative conclusion. This is slow institutional work. It is also the highest-leverage action available, because every subsequent data sharing initiative is built on top of this substrate.

4. Fund the Infrastructure and the People Who Sustain It

Shared data does not maintain itself, and the current funding model consistently underwrites the work that determines whether shared resources remain usable.

Fund research software engineers and data stewards explicitly. NIH should permit, and in some cases require, grants to budget for dedicated data stewards or research software engineers whose role is to ensure that shared data, code, and models are well-documented, FAIR-compliant, and genuinely reusable. Without this, data sharing too often means depositing a raw data dump with minimal or even potentially inaccurate metadata.

Recognize data curation and harmonization as scientific work deserving career credit. Cleaning, standardizing, and contextualizing a dataset for reuse is not administrative overhead. It requires domain knowledge, methodological judgment, and a working understanding of how the data will be used downstream. Under current incentive structures, the people who do this work well receive little credit for it, and the field has a chronic shortage of them as a direct result. NIH should make curation and harmonization contributions visible and creditable through contribution taxonomies, through data citation, and through explicit recognition in review.

Create distinct funding pathways for long-term maintenance. Widely adopted software tools, APIs, repositories, and models require ongoing maintenance, hosting, dependency updates, and user support. The current model funds initial development and then withdraws support at precisely the point when adoption has made the resource most valuable to the community. NIH should establish maintenance-specific funding mechanisms and should recognize sustained stewardship of a community resource, not merely its initial release, as a creditable scientific contribution.

Establish pre-validated model repositories. NIH should incentivize deposition of AI and machine learning models into vetted, standardized repositories that run automated benchmarking to verify that a deposited model is accessible, executable, and performs as reported. This addresses a fast-growing problem: models that are nominally shared but cannot be loaded, run, or reproduced. Automated verification at deposit is far cheaper than discovery of the failure by a downstream user months later, particularly if that failure is not widely reported.

Track and publish an inventory of data sharing platforms across NIH. Institutions, intramural programs, and extramural awardees maintain data sharing platforms that are not systematically catalogued. NIH should track these platforms and make the inventory available, so that the agency and the community can understand what data exist, where they reside, and how they can be accessed. This is a prerequisite for the cross-study analysis NIH wants to enable, and it is directly analogous to what NCI's Clinical Trials Reporting Program provides for the clinical trial landscape: a single, authoritative, continuously maintained view of what exists.

Summary

The recommendation is a reorientation across four dimensions: measure reuse rather than deposition; govern for reusability at the point of study design rather than the point of deposit; reduce the regulatory complexity that drives investigators toward conservatism, beginning with authoritative documentation of the scenario-dependent authority stack; and fund the people and infrastructure that keep shared resources usable over time.

Each of these is implementable with mechanisms that largely already exist. NIH has the repositories, the CDE infrastructure, the persistent identifier systems, the review gates, and the policy authorities. What is missing is the alignment of incentives with the outcome NIH actually wants, which is not compliance but reuse. That alignment is what this RFI is positioned to create.

3. Training and Mentorship: Training and Mentorship

NIH's current indicators of training success are measured against an implicit model of the successful trainee: one who remains in academia, secures independent funding, and produces publications. That model describes a shrinking minority of the people NIH actually trains. The consequence is a systematic measurement failure. NIH trains a workforce that overwhelmingly goes on to careers in industry, government, federal contracting, public health, policy, and education. These professionals make substantial contributions to the biomedical research mission, yet their impact is largely unmeasured. Any meaningful effort to quantify training impact must begin by redefining what constitutes a successful outcome.

We offer comments in four areas: outcome measurement, mentorship quality and accountability, protected time verification, and infrastructure to support trainee career development beyond the training period. Our perspective is that of an organization performing biomedical informatics and data science work in direct support of NIH programs, staffed substantially by people with advanced scientific degrees who came through the NIH-funded “traditional” training pipeline (post-baccalaureate, graduate, and postdoctoral researchers in academic laboratories) and now advance the NIH mission from outside academia.

1. Redefine and Measure the Successful Outcome

Track long-term trainee outcomes, not trainee counts. The number of trainees supported under an award is an input measure rather than an output measure. Instead, NIH should track what happens to those individuals over time: career progression, achievement of research independence in their chosen settings, retention in scientific careers broadly defined, and whether they go on to mentor future scientists. A five-year and ten-year longitudinal assessment of these trainee outcomes would provide a far more meaningful measure of return on NIH’s training investment than simply counting the number of funded trainee positions.

Track mentorship lineage. The most durable output of excellent mentorship is not the trainee's own productivity but the future scientists they subsequently develop. NIH should track mentorship lineage as a distinct measure of impact by assessing whether trainees become mentors themselves and whether their mentees, in turn, achieve success. This captures the compounding effect that makes mentorship such a valuable investment, yet it remains entirely invisible in current metrics.

Build the metrics to recognize non-academic careers, which are the majority case. This is the central recommendation of this section. A researcher who leaves academia for industry, federal service, contracting, secondary education, policy, or the nonprofit sector has not failed the training pipeline. In many cases, they have entered precisely the role where their training delivers the most public value. Yet NIH's dominant success indicators, publication counts and independent funding, capture almost none of

this.

Consider what this invisibility costs. A researcher who completes a fellowship and accepts a role in industry, government, or federal science support does not appear in NIH's success metrics. However, a researcher who leaves academia entirely because no academic position was available appears in the same way. These are not the same outcome, and NIH currently has no way to distinguish them.

NIH should develop alternative outcome indicators that reflect the actual distribution of careers its trainees enter. Candidate indicators include years active in scientific work under a broad definition; patents and licensed technologies; technical reports, white papers, and standards contributions; invited talks and professional society leadership; software and data resources produced; and number of individuals subsequently trained or supervised. These indicators are measurable, and several are already tracked by the institutions employing these scientists.

The design of these indicators should not be done in isolation from the people they describe. NIH should convene scientists who followed non-academic paths to help define what meaningful contribution looks like in those settings, because academic reviewers may not be best positioned to evaluate them.

Anticipate the participation problem. Non-academic employers and alumni have no current obligation or incentive to report outcome data to NIH. As a result, any tracking system that depends on voluntary participation is likely to suffer from significant response bias, disproportionately capturing individuals with the strongest outcomes. NIH should design deliberate incentives for participation: small grants or supplements to institutions that maintain high-quality alumni tracking, formal recognition programs for organizations that report consistently, or integration of outcome reporting into existing federal reporting relationships where they already exist. Without an explicit incentive structure, a long-term outcomes dashboard will systematically underrepresent the very population it is designed to capture.

Create a Biomedical Informatics and Data Science pipeline indicator specifically. There is an acute and worsening workforce shortage in bioinformatics, clinical data science, and computational biology. NIH depends on this workforce to support nearly every modern data-intensive program it funds, yet it has no dedicated metric to track how many trainees enter these fields. NIH should establish such a measure and recognize placement in biomedical informatics and data science roles, in any sector, as a positively weighted training outcome.

This shortage is addressable, and one effective model is deliberately structured senior-to-junior technical pairing on real production work. In our own experience, pairing a senior analyst with a junior analyst on substantive, delivery-bound projects, rather than on training exercises, has produced both rapid skill development and genuine scientific output: original contributions across computational oncology, clinical data informatics, and multimodal analysis that the teams would not have produced in a traditional training-only structure. The mentorship was meaningful, and the outcomes were tangible, yet neither would likely have been achieved through a traditional training-only model. NIH should recognize applied, delivery-embedded mentorship models as legitimate, fundable approaches to workforce development, particularly in computational fields where the most valuable skills are acquired through hands-on experience.

2. Make Mentorship Quality Visible, Measurable, and Consequential

Mentorship quality is currently assessed almost entirely on the basis of self-reported mentoring plans. Yet the individuals best positioned to evaluate a mentor are their trainees themselves, and NIH does not systematically solicit or incorporate their feedback.

Institute anonymized, aggregate trainee feedback as a standard input to review. NIH should incorporate structured, fully anonymized feedback from graduate students and postdoctoral fellows into the evaluation of PIs and institutions. This feedback should assess key aspects of the training experience, including career development support, coaching in scientific rigor, inclusion, and the overall quality of the training environment. Results should be aggregated at the institutional level to protect trainees from retaliation while providing meaningful, actionable insights. Institutions that consistently receive poor mentorship feedback evaluations across multiple awards may have systemic issues, and NIH should have the ability to identify and address those patterns.

Pilot a Mentorship Quality Index. Building on that feedback, NIH should pilot a composite index that can be incorporated into institutional training grant renewals. We would recommend piloting this before broader implementation and being explicit about the risks. Any survey-based instrument invites manipulation, and mentorship quality is inherently difficult to capture in a single metric. However, the current alternative, in which mentorship quality is effectively unmeasured and unaccountable, is less effective. A well-designed pilot with a predefined evaluation period would allow NIH to determine whether the instrument provides a meaningful and actionable measure of mentorship quality.

Add a dedicated mentorship section to the Biosketch, and make it scoreable. PIs should document their mentoring approach, the concrete milestones achieved by past trainees, and the career trajectories those trainees followed. Making this a scored peer review criterion, rather than a narrative afterthought, would create a direct incentive for investigators to take mentorship seriously and would give reviewers something substantive to evaluate.

Establish a trainee contribution attribution standard for computational papers. Multi-author computational publications currently obscure individual contributions. NIH should require structured contribution statements that specify which analyses each trainee performed, rather than relying on author order as a proxy for contribution. This makes mentorship visible, makes trainee contribution creditable, and provides the raw material for evaluating both.

Reward mentors who produce interdisciplinary scientists and support diverse career paths. A mentor whose trainees successfully pursue careers in industry, government, and public policy, and who actively prepares them for those roles, is serving the NIH mission. Under current incentives, that mentor may be reviewed less favorable than those who retain trainees in academia. NIH should explicitly recognize mentors who support the full range of productive career destinations and should recognize collaborative mentoring models in which expertise is distributed across multiple investigators rather than concentrated in a single PI.

Reward audited mentoring excellence with tangible support. PIs with a demonstrated, verified record of exceptional mentorship should receive priority consideration for administrative supplements or extended award durations that protect dedicated mentoring time. Effective mentorship is time-intensive, and the current incentive structure expects investigators to perform it within the same fixed time budget used to produce the research outputs on which they are primarily evaluated. Until the incentive structure resolves that conflict, simply encouraging better mentorship is unlikely to produce meaningful change.

Measure participation in rigor and stewardship training directly. NIH should track trainee participation in formal mentoring plans, responsible conduct of research training, data stewardship, and reproducibility training as distinct indicators of the training experience. It should also evaluate whether these activities are effectively improving trainee knowledge, skills, and research practices rather than simply documenting participation. Reproducibility and data stewardship are learned behaviors, and the training period is when they are most efficiently instilled.

3. Verify Protected Time Rather Than Assuming It

Protected research time for trainees is currently enforced on an honor system, and the honor system is not proving effective. Trainees on T32 and K awards routinely report that clinical, teaching, or administrative obligations consume time that is nominally protected, and there is currently no mechanism for NIH to identify or monitor these situations.

Require annual reporting of actual versus required protected time. Trainees supported by protected-time awards should confidentially report, on an annual basis, the amount of protected research time they received against the time their award requires. This would represent a low-burden data collection effort with the potential to provide highly valuable insight into whether protected time commitments are being honored.

Attach consequences to persistent institutional shortfalls. Reporting without consequence changes nothing. Institutions that show consistent, verified shortfalls across multiple awards and multiple trainees should face defined consequences in future training grant competitions. The unit of accountability should be the institution rather than the individual PI, both to protect trainees from retaliation and because persistent protected-time erosion is typically an institutional resourcing problem rather than an individual failing.

Audit rather than attest. NIH should conduct periodic institutional audits of protected time protection, sampling across awards. The existence of an audit function, even at low sampling rates, changes institutional behavior in a way that self-attestation does not.

4. Support Trainee Development Beyond the Training Period

NIH's investment in a trainee currently ends when the appointment ends, at the critical point when the trainee's professional trajectory is being determined.

Invest in professional profile development during post-baccalaureate and postdoctoral fellowships. Before completing their appointments, trainees should be equipped with the professional infrastructure needed to sustain a long-term research identity, including persistent researcher identifiers, complete and maintained publication and contribution records, and a professional presence that allows their work to be found and their career trajectory to be trackable. This is a modest investment that pays for itself, both in long-term trainee outcomes and in NIH's own ability to measure them. A trainee who cannot be found five years later cannot be meaningfully included in those outcomes.

Fund publication and dissemination support for trainees after their appointment ends. In our experience, a substantial amount of work completed during training goes unpublished because the trainee departs before the manuscript is finished. This represents a direct loss of NIH-funded scientific output. NIH should establish a small, low-overhead mechanism to support peer-reviewed publication and white paper dissemination during the period following a trainee’s appointment, particularly for those entering settings without institutional publication support. The benefit is threefold: the science reaches the community, the trainee's professional development is advanced, and a durable link is maintained between NIH and a scientist now positioned in industry, government, or another sector where NIH benefits from having a relationship.

Summary

NIH cannot manage what it does not measure, and it is currently not measuring the majority of what its training investment produces. The highest-priority actions are to redefine the successful outcome to include the non-academic careers that many trainees enter, to institute anonymized trainee feedback as an accountability mechanism for mentorship quality, and to verify protected time rather than assuming it. Together these changes would give NIH an accurate picture of what its training investment produces and where it is being lost.

4. Collaboration: Collaboration

NIH has correctly identified that biomedical research has become team-based while the credit system remains individual. The standard response to this problem is to improve attribution at the point of publication, typically by adopting a contributor role taxonomy. We support that step and address it below. But we want to make a more fundamental argument first, because we think the publication-stage framing is diagnosing the problem too late.

By the time a multi-site study reaches manuscript preparation, the contribution record has already been lost. The information about who generated which samples, which laboratory ran which assay, which site enrolled which patients, which team harmonized which data elements, and which analyst produced which result exists in operational systems during the conduct of the study, and it is not carried forward. Authors then reconstruct contributions from memory, negotiation, and social convention, and a taxonomy applied at that point is formalizing a reconstruction rather than recording a fact.

The alternative is to treat contribution as provenance metadata, captured at the point of data generation and carried with the data through analysis, dissemination, and publication. This is technically achievable today, and we describe below how it works in practice in NCI programs we support. It would transform

contribution attribution from a negotiated social artifact into a queryable property of the scientific record.

Our comments address four areas: making contributions attributable through infrastructure, recognizing the enabling work that makes team science possible, structuring credit for multi-site and multi-PI science, and funding the collaboration phase itself.

1. Make Contribution Attributable at the Source

Adopt contributor role taxonomies, and adopt them as a requirement rather than an option. NIH should require use of CRediT or an equivalent contributor taxonomy for publications arising from multi-PI, center, and network awards, and should incorporate the resulting contribution records into the NIH Biosketch. Distinguishing conceptualization, methodology, software, data curation, formal analysis, validation, and project administration makes visible the work that first-and-last-author convention erases. The current system, in which position in an author list serves as a proxy for both the nature and the magnitude of contribution, systematically disadvantages methodologists, data scientists, software engineers, and clinical coordinators whose contributions do not map onto the conventional hierarchy.

This is a necessary but not a sufficient reform, for the reason stated above: a taxonomy applied at manuscript stage records what the author group agrees to record.

Build contributor attribution into data provenance. The durable solution is to tag contribution at the data element level, within the systems that operate multi-site science, and to carry that metadata forward automatically.

This is not hypothetical. In the Federal programs we support, the informatics infrastructure that enables multi-institutional research already tracks the provenance of data as an operational necessity. Data flowing from a national network of trial sites and specialized assay laboratories must be traceable to its source in order to be interpretable: an analyst needs to know which laboratory ran a given assay, under which protocol version, on samples from which trial and which site.

The insight is that the contribution record NIH wants already exists inside these systems, as a byproduct of running the science correctly. It is simply not being captured, retained, and propagated as an attribution artifact. Extending existing provenance metadata to include structured contributor information, and carrying it through harmonization into the released data package, is an incremental technical change with substantial downstream benefit:

Contribution is recorded as fact at the time of generation rather than reconstructed later.

Contributions travel with the data package at dissemination, so downstream users can see and cite the contributing sites and teams.

Contribution becomes machine-readable and therefore countable, which is precisely what NIH needs to build any credible collaboration metric.

Sites and laboratories that contribute data but do not author the primary manuscript become visible and creditable.

We would recommend NIH pursue this concretely: require that data packages released from network and consortium awards carry structured contributor metadata as a condition of deposition and fund the infrastructure work needed to generate it. The specification should be developed with the coordinating centers and informatics teams that operate these systems, because they are the ones who know what provenance is already being captured and what it would cost to extend it.

Develop a Network Contribution Index on that foundation. With structured contributor metadata in place, NIH can construct a credible index that credits each participating site and laboratory for what it contributed: samples and data generated, protocols contributed, patients enrolled, assays performed, analytic work delivered, and coordination effort supplied. Critically, this credit should be decoupled from who holds the prime award. Under the current structure, the prime awardee institution captures the institutional credit for a network's output while contributing sites, which may have generated most of the data, receive attribution only through authorship on the resulting papers, if at all.

Develop a Cross-Institutional Data Access metric. Track how frequently a dataset, tool, or resource developed under one award is used by investigators at other institutions. This operationalizes collaboration as an outcome rather than an intention, and it is measurable using persistent identifiers and repository access records that largely already exist.

Distinguish substantive interdisciplinary contribution from co-authorship inflation. A biostatistician who designed the core analytic strategy and a group that supplied one dataset have made contributions that differ in both kind and magnitude, and both currently appear as a name in a list. Structured contribution statements are the mechanism for separating them. Without this distinction, any collaboration metric based on co-authorship will reward author list length rather than collaborative substance and will be gamed accordingly within a single funding cycle.

2. Recognize the Enabling Work That Makes Team Science Possible

Multi-site science does not happen because investigators agree to collaborate. It happens because someone builds and operates the infrastructure that allows a hundred sites to produce comparable data, someone harmonizes that data into an analyzable form, someone maintains the platform on which the analysis runs, and someone coordinates the whole enterprise. That work is scientific work. It requires methodological judgment, domain knowledge, and substantial technical expertise. It is also, under current metrics, nearly invisible.

Fund collaboration infrastructure explicitly and recognize it as a scientific contribution. Shared data platforms, coordinating centers, harmonization hubs, biorepositories, and core facilities are the precondition for team science and are chronically underfunded relative to their importance. NIH should fund them as first-class scientific infrastructure and should recognize the people who build and run them accordingly.

Create recognition pathways for core facility and platform staff scientists. NIH should establish award tracks, citation mechanisms, or formal designations that credit core facility directors, coordinating center leads, staff scientists, and research software engineers whose infrastructure enables multi-center science. These individuals frequently appear as middle authors, when they appear at all, and are evaluated against productivity metrics designed for a role they do not hold.

Credit operational excellence as a collaborative contribution. Data integration, software development, project management, and reliable operational delivery are what convert a collaborative intention into a functioning multi-site program. NIH should credit investigators and teams who enable collaborative science through this work, and should evaluate collaborative outcomes, shared datasets, multicenter publications, and community resources, rather than emphasizing individual productivity alone.

Align tenure and promotion incentives with team science. NIH cannot reform university promotion committees directly, but it has considerable influence. Federal guidance urging institutions to formally recognize substantive middle-author contributions to high-impact team science as comparable to sole-PI work would give institutional advocates a lever they currently lack. Without this, NIH will continue to ask investigators to prioritize collaboration while their employers continue to promote on the basis of individual credit, and investigators will continue to respond rationally to the incentive that determines their careers.

3. Restructure Credit for Multi-PI and Multi-Site Awards

Split multi-PI credit equally by default. NIH should redesign multi-PI mechanisms so that project credit, institutional indirect cost recovery, and NIH portfolio tracking are attributed equally across PIs unless the team specifies otherwise. The current structure, in which one institution and one investigator are functionally the center of gravity regardless of the nominal equality of the multi-PI model, produces a persistent "dominant lab" dynamic that discourages genuine peer collaboration. Equal default attribution changes the negotiating baseline and any deviations from the baseline must be justified.

Recognize interdisciplinary collaboration that crosses sectors. Increasingly, the collaborations that produce translational impact span academia, healthcare delivery systems, government, and industry. NIH's metrics should recognize these collaborations rather than treating non-academic partners as adjuncts to an academic core. This is particularly relevant to informatics and data science, where the deepest operational expertise in running large-scale clinical data systems frequently resides outside academic institutions.

4. Fund the Collaboration Phase

Large team science projects fail or underperform for reasons that are established long before the award begins: the participating groups have not aligned on scientific aims, the data models are incompatible, the harmonization burden was never scoped, and the governance structure was improvised. NIH currently funds none of this preparatory work and then funds the full project on the assumption that it was somehow done.

Establish interdisciplinary bridge funding. NIH should offer small, fast, low-overhead awards specifically to fund the collaboration phase preceding a major team science project: cross-institutional scientific alignment, data model reconciliation, harmonization scoping, governance design, and pilot integration work. In our experience, the cost of failing to do this work up front is not that it goes undone; it is that it gets done during the funded project, consuming time and budget that was allocated to science, and often done under schedule pressure in ways that produce technical debt lasting the life of the program. A modest investment before the fact is dramatically cheaper than the same work performed reactively.

5. A Note on Shared Federal Knowledge Infrastructure

NIH funds an enormous volume of codebases, data resources, standard operating procedures, and technical documentation that is fragmented across institutes, programs, and awards, with no systematic mechanism for one program to discover and reuse relevant work produced under another. NIH should evaluate whether a shared, appropriately governed knowledge infrastructure built over these government-funded assets could reduce this fragmentation. The design challenges — data sensitivity, access authority, provenance, and the risk of presenting incorrect outputs confidently — are real and should be resolved before any deployment. We raise the concept as an opportunity worth evaluating, not a recommendation to implement immediately; the underlying problem, however, is real and worth solving.

Summary

The core recommendation is to move contribution attribution upstream: require that data packages released from network and consortium awards carry structured contributor metadata as a condition of deposition and fund the infrastructure to generate it. Adopting contributor taxonomies at publication is a necessary first step, but it formalizes a reconstruction of the contribution record rather than capturing the record itself. Alongside that, NIH should formally recognize the coordinating centers, data platforms, and harmonization infrastructure that make team science possible; restructure multi-PI and network credit so it is not concentrated at the prime institution; and fund the collaboration phase that precedes large team science awards, which currently goes unfunded and is therefore either skipped or performed at the expense of the science.

5. Entrepreneurship and Translation: Entrepreneurship and Translation

NIH's translation agenda is built on a pipeline model: discovery originates in academia, matures through translational funding, and passes to industry for commercialization and delivery. That model is real and it works. But it is unidirectional, and the assumption that innovation only ever flows in that one direction is costing NIH a substantial amount of innovation it could otherwise capture.

We make three arguments in this section. First, NIH's definition of a translational output is too narrow, and it systematically fails to credit the reusable tools, platforms, and data standards that enable large volumes of downstream science. Second, translation needs to be measured by adoption, not by the legal artifacts of commercialization. Third, and most substantively, NIH should build a return channel: a funding mechanism that allows innovative ideas originating in industry to be developed in academia and

matured back into practice. That channel does not currently exist, and its absence is a structural loss to the innovation ecosystem NIH is trying to foster.

1. Broaden What Counts as a Translational Output

Recognize reusable tools, platforms, and standards as primary translational outputs. A bioinformatics pipeline adopted by fifty downstream studies, a clinical data standard that makes cross-study analysis possible, or a data platform that enables a national research network has delivered translational value at a scale most individual products never reach. Under current metrics, none of this reliably registers. NIH's translational indicators are oriented toward patents, licenses, INDs, and spin-outs, and a widely adopted open research tool produces none of those artifacts while potentially advancing the mission more than any of them.

As a concrete example, NCI's Clinical Trials Reporting Program provides the authoritative national registry of NCI-supported cancer clinical trials, and its data standards, abstraction protocols, and reporting infrastructure are the substrate on which cancer trial reporting, regulatory compliance, and cross-center benchmarking depend. That infrastructure was translated in every meaningful sense: it moved from concept into operational use, it was adopted at scale, and it changed how the field works. It generated no patents. NIH should have a way to see and credit outputs of this kind, because at present it largely does not.

Recognize enabling technologies, diagnostics, and research tools, not just startups. Entrepreneurship in the biomedical context is frequently equated with company formation. Company formation is one path to impact, and NIH should continue to support it. But an investigator who produces a broadly usable diagnostic method, a validated assay, or an enabling technology adopted across the field has translated their discovery, whether or not a company resulted.

Define translation milestones by research type. A single definition of translation cannot serve basic science, clinical research, and methods development simultaneously. NIH should define appropriate milestones for each, and track investigator progress against the relevant set:

For basic research: a validated method, a licensed tool, a reagent or model system in broad community use, a mechanistic finding that enables a downstream therapeutic program.

For clinical research: a change in clinical practice guidelines, an FDA submission, an approved indication, incorporation of a finding into standard of care.

For methods and informatics research: adoption of a tool, pipeline, or standard by external groups; incorporation into a data commons or a national research infrastructure; measurable improvement in the speed or quality of downstream research.

The value of an explicit, type-specific milestone framework is that it lets NIH credit progress rather than only endpoints, and it prevents basic researchers from being penalized for being early in a pipeline that has not yet reached its terminus.

2. Measure Adoption, Not Legal Artifacts

Downstream adoption is the correct translational metric. Patents are an input to translation, not evidence of it. A large fraction of issued patents is never licensed, never practiced, and never reach a patient. NIH should measure the adoption of technologies in research, healthcare delivery, and clinical trials as the primary indicator, and treat patents, licenses, and IND filings as supporting evidence rather than as the measure itself.

Reward measurable progression toward translation. Validation studies, regulatory milestones, and advancement in technology readiness are all measurable, and all represent real progress that current metrics ignore because they precede the terminal commercialization event. Crediting progression rather than only completion changes investigator behavior at the margin, which is where the valley of death claims most projects.

Establish a Translational Impact Registry. NIH-funded discoveries that enter clinical use, receive regulatory clearance, achieve licensure, or are adopted as a research standard should be formally logged and credited back to the originating award and to the investigator's record. The absence of such a registry means that translational success, when it occurs, is often invisible in NIH's own portfolio data, particularly when it occurs years after the award closed and through intermediaries the original investigator never interacted with.

Pilot translational credit scoring in competitive renewals. Investigators who can demonstrate that prior work led to a licensed product, an initiated clinical trial, an adopted standard, or a documented change in practice should receive scoring recognition at renewal. This should be structured as a bonus rather than a requirement, so that basic researchers early in the translational arc are not penalized for the position they occupy in the pipeline.

Expand the Biosketch to reflect translational contribution. Issued patents, IND filings, licensed technologies, adopted standards, spin-out formation, and documented tool adoption should be formally counted as indicators of scientific impact alongside publications. Investigators currently have no structured place to report this work, and reviewers therefore have no structured way to weigh it.

Give positive review weight to translation training. PIs who have participated in commercialization and translation training, such as I-Corps, have made a deliberate investment in learning how to move work toward impact. Recognizing that participation in review costs NIH nothing and signals clearly that translation is valued.

Credit public-private partnerships that maintain scientific rigor. Partnerships that accelerate patient impact while preserving methodological integrity should be creditable outcomes. Our own experience supporting national-scale trial networks shows that effective public-private collaborations function only because multiple stakeholders operate as an integrated system and that integration is itself a translational achievement that is not credited anywhere.

3. Build the Return Channel: Industry-Originated, Academically Developed Innovation

This is our central recommendation, and we believe it identifies a structural gap in the innovation ecosystem that NIH is well positioned to close with the deep connections the NIH harbors.

The current flow is unidirectional. Scientists trained in NIH-funded academic environments move into industry, government, and federal contracting, carrying knowledge and methods with them. This is a large and continuous flow, and it is valuable. But there is no meaningful return channel. Ideas originating in industry, in the organizations doing applied biomedical work at scale, do not flow back into academia for development.

This is not because industry lacks ideas. Applied organizations operate against delivery commitments, and speculative development competes directly against contracted work for the same staff hours. An idea that would require six months of exploratory engineering to reach proof of concept, with genuine risk of failure, is very difficult to fund internally, regardless of its potential value. The idea is not built, published or transferred. It simply does not happen.

Academia has precisely the complementary profile. Academic environments are structured for exploratory work with uncertain outcomes. They have trainees who need substantive projects, faculty who are rewarded for novelty, and institutional tolerance for the possibility of failure. What they frequently lack is the applied problem context, the operational data, the delivery discipline, and the product maturation pathway to carry a proof of concept into practice.

These are complementary deficits, and they are not being matched. The result is a persistent, largely invisible loss: ideas that industry can conceive but cannot build, and academic capacity that could build them but never sees the problem.

We recommend NIH establish a bidirectional innovation mechanism. The structure we would propose:

A small, fast, R21-scale grant mechanism specifically for industry-originated problems developed in academic laboratories. The industry partner contributes the problem definition, the applied context, and where permissible the operational data. The academic partner performs the exploratory development. NIH funds the academic work. These awards should require minimal supporting data by design given that the value of the mechanism lies in funding work too early-stage to demonstrate feasibility.

Clear intellectual property structure established up front. This is the mechanism's central design challenge and its central risk. Industry participants will not contribute genuinely valuable ideas into an arrangement where their IP position is uncertain, and academic participants will not accept terms that suppress their ability to publish. The mechanism must resolve this explicitly: protect the originating party's core IP position while guaranteeing the academic team's right to publish the scientific results and disseminate the methods. This tension is real and cannot be papered over, but it is not novel, and workable frameworks exist in other public-private research contexts.

Career credit for the academic participants. Faculty and trainees who take on industry-originated

problems should receive recognized career credit for doing so, including promotion and tenure consideration, formal recognition in competitive renewal review, and/or with assistance with funding future publications or patents. Without this, the mechanism will attract only investigators with nothing to lose.

Milestone-linked incentives. Financial incentives tied to successful maturation milestones, such as validation, technology readiness advancement, and eventual deployment, would align both parties toward carrying the work through rather than abandoning it at proof of concept.

Deliberate resolution of the publication question. Open publication and dissemination should be a condition of participation, on a defined timeline that permits the industry partner to establish necessary protections first. NIH-funded work should enter the literature.

The mechanism would produce publishable science, train students on problems that have demonstrated real-world relevance, and generate innovations with a defined pathway into practice, because the industry partner has both the motive and the infrastructure to mature them. It also creates a durable connection between academic laboratories and the applied organizations that employ most of their graduates, which has independent value for both workforce development and translation.

A bounded pilot structure, for example a cohort of awards at the R21 scale with explicit evaluation criteria and IP protections built into the agreement terms from the outset, would allow NIH to test the mechanism and establish precedents before broader commitment. The mechanism is small-award-scale by design. And it addresses a gap that no current NIH mechanism addresses, because every existing translational mechanism assumes the idea originates in the academic laboratory.

4. Fix the Valley of Death Where It Actually Opens

Create a Phase 0 bridge mechanism. SBIR and STTR programs assume a level of technical maturity that many academic discoveries have not reached. A researcher with a promising finding and no minimum viable product is not competitive for SBIR and has no mechanism to become competitive. NIH should establish a dedicated pre-SBIR bridge award to mature a discovery to the point where commercial mechanisms become accessible. This is where a large number of viable translations quietly die.

Extend SBIR and STTR pathways for informatics and AI tools specifically. Biomedical informatics and AI tools developed under NIH awards frequently have clear clinical or commercial utility but face a regulatory and business development pathway that academic teams are unequipped to navigate. The technical work is done and the tool does not reach anyone. NIH should expand SBIR and STTR support tailored to software and data tools, including support for the regulatory pathway determination that software-as-a-medical-device products increasingly require.

Fund academic-to-industry trainee time splits. Graduate students and postdoctoral fellows should be able to split time legitimately between an academic discovery laboratory and a commercial or applied R&D environment with a transferable fellowship. This is currently difficult to arrange and frequently discouraged. It is one of the most efficient available mechanisms for transferring both knowledge and

people across the boundary, and it addresses the innovation flow problem from the workforce side.

Push institutions to standardize technology transfer. Institutional technology transfer processes vary enormously in speed, terms, and competence, and a researcher's ability to translate a discovery frequently depends more on which university employs them than on the merit of the discovery. NIH should use its considerable influence to encourage standardization and simplification of institutional tech transfer pathways, so that translation is rewarded rather than obstructed.

Summary

NIH should broaden what counts as a translational output to include the reusable tools, platforms, and standards that enable downstream science; measure translation by adoption rather than by patents and legal artifacts; establish registries and review mechanisms that make translational success visible and creditable; and address the valley of death with a pre-SBIR bridge mechanism and informatics-specific commercialization support.

Above all, NIH should recognize that the innovation pipeline it has built runs in only one direction. Industry conceives ideas it cannot afford to build; academia can build ideas it never sees. Constructing the return channel would cost relatively little, would produce publishable science and trained students, and would capture innovation that is currently being lost entirely.

6. Foundational Scientific Exploration: Foundational Scientific Exploration

NIH's difficulty in funding high-risk research is not a failure of intent. The agency has created mechanisms explicitly for this purpose and states priority clearly. The difficulty is that the surrounding system, peer review, career advancement, publication, and institutional promotion, is optimized to reward safety, and it exerts that pressure continuously on every actor in the system. Investigators, reviewers, and institutions respond rationally to the existing system. The aggregate outcome is a portfolio more conservative than anyone involved intends.

Correcting this requires acting on the specific mechanisms that produce conservatism rather than exhorting the participants to be bolder. We address four areas: the preliminary data paradox, the structural bias in expert peer review, the treatment of negative results, and the career risk borne by the investigator who takes scientific risks. We note that the barriers to bold work differ by sector, and that NIH's mechanisms should account for both.

1. Resolve the Preliminary Data Paradox

The R21 and DP2 mechanisms are the right instruments for exploratory research. In practice, they do not function as designed, for a reason that is well understood and rarely stated plainly: preliminary data is not formally required, but it is functionally required.

Reviewers, acting reasonably and in good faith, use preliminary data as the primary evidence that a proposed line of work is likely to succeed. An application with preliminary data is scored more favorably than one without. Investigators learn this quickly and respond by submitting work that is already

substantially de-risked. The mechanism intended to fund exploration ends up funding modestly extended versions of work already underway.

The paradox is structural. The riskiest proposals are, by definition, the least likely to have preliminary data. If strong preliminary evidence existed, the hypothesis would not be risky. A review process that treats preliminary data as the primary evidence of feasibility will therefore systematically select against exactly the work these mechanisms exist to fund.

We recommend NIH address this directly rather than through guidance to reviewers, which has been tried and has not worked.

Establish a distinct high-risk mechanism with its own study sections and its own criteria. A dedicated track, an R21-HR or equivalent, in which every application in the pool is speculative and preliminary data is explicitly excluded from consideration, would apply the criteria evenly across the whole pool. The essential feature is that applications compete against other speculative applications, not against de-risked ones. When a speculative proposal is scored in the same pool as a well-supported incremental one, it loses, and the reviewer has done nothing wrong.

A dedicated track has a further advantage for NIH: it makes the high-risk portion of the portfolio explicit and adjustable. NIH can set the proportion of funds allocated to speculative work deliberately, according to national scientific priorities, rather than discovering after the fact that the intended risk appetite was eroded by review dynamics.

Create a blinded review track. NIH should dedicate a defined percentage of funding to a review track that removes investigator identity and preliminary data entirely and evaluates the application solely on the significance of the question, the coherence of the reasoning, and the rigor of the proposed approach. This directly tests whether the idea can stand without the reputational and evidentiary scaffolding that normally carries an application through review. It also addresses, as a secondary benefit, the well-documented advantage that established investigators hold in conventional review.

Establish a High-Risk Research Designation with protected renewal criteria. Awards carrying this designation should be evaluated at renewal primarily on the rigor of the approach and the significance of the question pursued, explicitly not on publication productivity during the award period. An investigator who spends three years rigorously testing a bold hypothesis that proves incorrect has done what NIH asked them to do. Under current renewal criteria they will struggle, and every investigator watching them learns the relevant lesson.

Create an Early Career Exploratory Fund shielded from productivity metrics. Small awards, under $250,000, for early-career investigators pursuing genuinely novel hypotheses, evaluated at close solely on what was learned and what new questions the work opened. Early-career investigators face the sharpest version of the risk-aversion incentive, because a fallow period that a tenured investigator can absorb is career-ending for them. Shielding these awards from productivity evaluation is the goal of the mechanism.

2. Counterbalance the Conservatism of Expert Review

Peer review is simultaneously the strongest quality mechanism NIH has and a structural source of conservatism. Expert reviewers evaluate proposals using their own understanding of the field. That understanding is exactly what makes their judgment valuable, and it is also what makes them poorly positioned to evaluate work that would overturn it. Reviewers reward what they recognize as sound, and what they recognize as sound is, definitionally, the current paradigm. A proposal that would change how a field thinks will frequently be evaluated by the people most invested in how the field currently thinks. No individual reviewer needs to act in bad faith for this to produce a systematic bias toward the safe and familiar.

The counterweight to expert judgment is judgment from outside the expert community. We recommend NIH pilot a structured outsider review component, contributing a defined and limited portion of the total score, perhaps twenty percent, and scored against different criteria than the expert review:

Does the question matter? Will the answer move the field materially, independent of whether the field currently believes the answer is likely?

Is the reasoning coherent? Does the logical progression from premise to hypothesis to approach hold together?

Is there a reasonable path to the result? Not a likely positive result, but a design capable of producing an informative outcome either way.

Reviewers for this component should come from adjacent fields, from other disciplines entirely, or from applied and industry backgrounds. They will not be able to assess technical feasibility within the field, which is precisely the point: that assessment is what the expert panel provides. What outsiders can assess is whether the idea is interesting and whether the argument is sound, and those judgments are less contaminated by investment in the current paradigm.

NIH could also evaluate whether automated review support has a role in this component, given that language models can assess logical coherence and structural quality at scale and without disciplinary investment. We raise this only as an option worth studying, not as a recommendation to implement. The risks are significant, including the possibility that such systems reproduce rather than counteract existing bias, since they are trained on historical literature and would carry its assumptions forward. Any use would need to be carefully validated against human judgment before it influenced a real funding decision, and it should supplement rather than substitute for human review. The substantive recommendation stands independently: expert review needs a counterweight from outside the expert community.

Implement rapid-response speculative micro-grants. Standard grant cycles run nine to twelve months from submission to award. A researcher with a sudden, potentially paradigm-shifting idea must either abandon it or hold it for a year, during which the moment frequently passes. NIH should implement rolling, fast-tracked micro-grants that allow immediate testing of a speculative idea. The award sizes

would be small, the failure rate would be high by design, and both of those properties are appropriate to the purpose.

3. Make Negative Results a First-Class Scientific Output

Formalize negative results and productive failures as deliverables for high-risk awards. A documented account of what was attempted, what was found, why the hypothesis did not hold, and what the work rules out should be an acceptable, citable, and creditable output of an exploratory grant. NIH should define the format, require it, and count it.

The scientific logic is straightforward and we think under-appreciated. Exploratory research operates at the edge of knowledge, where hypotheses are more likely to be wrong than right. That is the definitional property of the work. A funding mechanism that treats a disproven hypothesis as a failed award is not funding exploration; it is funding confirmation. And the knowledge produced by a rigorous negative result is real: it closes off a line of inquiry, redirects the field, and prevents other groups from repeating the same effort.

Rigorously conducted negative findings can be decisive. As a clear example, a clinical trial evaluating adavosertib in heavily pretreated patients with advanced solid tumors reported that the agent did not deliver the anticipated survival benefit in that population (Kummar et al.; JCO Precision Oncology; 2026). The result was rigorous, it was published, and it was consequential. It can inform subsequent development decisions, and spare future patients an ineffective therapy. That is a valuable scientific contribution. Under conventional productivity metrics, it registers as a study failure.

Distinguish productive scientific risk from poor execution. This is the legitimate concern underlying reviewer conservatism and it deserves a serious answer rather than dismissal. Not every failed project is a valuable negative result. Some projects failed because they were badly designed, badly executed, and/or badly managed.

The distinction is assessable, and NIH should build explicit criteria for it:

Was the hypothesis rigorously tested? Was the study adequately powered, appropriately controlled, and executed with sufficient methodological quality that the negative result is trustworthy?

Was the data of high quality? A negative result from poor data is not a negative result; it is an absence of information.

Was the failure informative? Does the community now know something it did not know before?

Did the investigator pivot appropriately? When early data indicated the initial approach would not work, did the investigator recognize this and adapt, or continued to expend resources on an approach that data had already ruled out?

An investigator who rigorously disproves their own hypothesis and clearly documents what the field

should conclude has succeeded. An investigator who produces an uninterpretable result through poor design has not. Evaluating methodological rigor, data quality, and the agility of the pivot, rather than whether the hypothesis held, makes the distinction.

Support milestone-based funding. Bold ideas should be able to mature without requiring immediate positive results, and milestone-based release is the appropriate structure: funding advances at defined stages such as project plan, initial proof of concept, validation, and handoff to a larger mechanism or to a partner capable of maturing the work. This structure is well established in the startup and small-business funding world, where it has proven effective precisely because it manages risk without demanding certainty at the outset. It allows NIH to limit exposure to any single speculative bet while permitting the successful ones to continue.

Recognize foundational outputs before translational impact is visible. Foundational datasets, enabling technologies, and mechanistic insights that make future discoveries possible should be creditable outputs. Their value is frequently realized years later and by other groups, and a metric system that waits for downstream impact will fail to credit the people who made it possible.

4. Address Career Risk

Every recommendation above is undermined if the individual investigator still bears the career cost of a null result.

An investigator considering a bold project faces a clear calculation. If the speculative project fails and produces few publications, that weakens the next grant application and damages the tenure case. The safe project produces reliable output. NIH can create as many high-risk mechanisms as it likes; if this calculation holds, the strongest investigators will continue to route their most ambitious ideas into their safest proposals.

Establish institutional guidelines protecting junior faculty who take scientific risk. NIH should issue guidance urging promotion and tenure committees to formally credit rigorous high-risk work, including work that yielded a negative result, and to weigh it comparably to conventional productivity. NIH does not control tenure decisions, but it has substantial influence over institutional behavior and should use it.

Consider tying institutional incentives to this behavior. Institutions that demonstrably protect and promote faculty engaged in high-risk foundational research could receive favorable treatment in indirect cost recovery or in institutional training and infrastructure awards. This converts an exhortation into a financial incentive and institutional behavior responds to financial incentives far more reliably than guidance.

Engage the publishing ecosystem. Career risk is transmitted substantially through publication. If journals will not publish rigorous negative findings, an investigator, whose hypothesis fails, has no way to convert their work into the credential which their career requires. NIH should use its considerable leverage with publishers to establish accessible, indexed, and citable venues for rigorous negative results. Without this,

negative results as a grant deliverable produce a document that satisfies NIH and does nothing for the investigator's career.

Summary

NIH's exploratory funding is constrained by four mechanisms operating independently of anyone's intent: preliminary data functions as a de facto requirement in review, expert reviewers are structurally invested in the paradigms that bold work would overturn, negative results are treated as failure rather than knowledge, and the individual investigator bears the full career cost of scientific risk. Each is addressable through the mechanisms described above. The binding constraint is the last one: no mechanism for high-risk research will work if the investigators best positioned to do that work cannot afford to take the risk.

7. Public Impact: Public Impact

NIH's public impact is real, substantial, and largely unmeasured. NIH-funded research changes clinical practice, extends lives, generates economic activity, and produces the knowledge on which the health system operates. Evidence of this appears constantly, in news coverage, institutional communications, and the professional accounts of individual scientists. What it does not appear in is any central, structured, queryable record.

The consequence is that NIH cannot answer, with data, the question it is most often asked to answer: what has this investment produced for the American public? The agency can point to individual successes. It cannot systematically quantify them, trace them back to the awards that produced them, or use them to inform where to invest next.

This is a data infrastructure problem, and it is solvable. The information required exists. It is distributed across clinical guideline bodies, regulatory filings, publication databases, policy documents, patents, employment records, and health outcome data. Assembling it into a coherent, maintained, public-facing resource is precisely the kind of extraction, standardization, harmonization, and reporting work that NIH already funds routinely for scientific data. What is missing is the decision to apply that same discipline to NIH's own impact.

We highlight a few examples where this approach has had tangible benefits.

1. Public Impact Measurement Is a Solved Technical Problem Awaiting an Institutional Decision

During the SARS-CoV-2 pandemic, decision-makers at every level needed to know what proportion of the population had been infected. That information existed, scattered across seroprevalence studies conducted by different investigators, with different assays, different populations, and different reporting formats. Individually, no single study could answer the question, but the aggregate data could if combined in a usable way.

Seroprevalence data from studies across the country were harmonized into a common structure and surfaced through a public-facing dashboard (called SeroHub) that allowed researchers, public health

officials, and the public to see the actual state of population immunity, updated as new data arrived (Freedman et al; Sci Data; 2022). The result was that stakeholders responding to a fast-moving public health emergency could act on current, harmonized evidence rather than on outdated models or inference.

This lesson is parallel to NIH's desire to monitor public impact. The barrier to measuring NIH's public impact is not that the data does not exist, or that the technical methods are unavailable. It is that no one has been tasked with harmonizing it and no infrastructure has been funded to hold it. This is the same problem SeroHub solved, applied to a different domain.

The same is true of the infrastructure NIH already funds for other purposes. NCI's Clinical Trials Reporting Program maintains an authoritative national view of the cancer clinical trial landscape, supporting regulatory reporting compliance, accrual monitoring, and cross-center benchmarking, because the NCI knew it was necessary and funded the infrastructure to sustain it. NIH's public impact deserves the same treatment.

Recommendation: build a public-facing Research Impact Dashboard. Modeled conceptually on USASpending.gov but oriented toward health and scientific outcomes rather than obligations, this resource would allow the public, Congress, and NIH itself to see how research investments connect to health improvements, clinical practice change, economic activity, and workforce development over time. It would serve three distinct purposes simultaneously: it would build public trust through transparency, it would create genuine accountability infrastructure, and it would give NIH the evidence base to direct future investment toward what demonstrably works.

We would emphasize that this is a substantial and sustained data engineering undertaking, not a reporting exercise, and it should be resourced accordingly. It requires harmonizing heterogeneous sources, maintaining linkage between awards and downstream outcomes over multi-year lags, and sustaining the resource over time. These are known problems with known solutions, and NIH funds this class of work routinely.

2. Establish a Health Outcome Attribution Framework

The ultimate measure of NIH's mission alignment is whether funded research improved health. NIH should build a structured framework for documenting when a line of research contributed to a change in clinical guidelines, a reduction in disease burden, an improvement in diagnostic accuracy, or the availability of a new treatment, and for tracing that contribution back to the originating awards.

Attribution in biomedicine is genuinely difficult. The path from a funded study to a changed clinical outcome is typically long, indirect, and multiply determined, and any framework claiming clean causal attribution would be wrong. But the difficulty of perfect attribution is not an argument for no attribution. A framework that documents contribution transparently, with appropriate acknowledgment of shared credit and uncertainty, would be enormously more informative than the current state, which is effectively silence.

The clearest cases are traceable end to end, and NIH should start there. For example, an NCI precision oncology clinical trial network's findings in patients with BRAF V600E-mutated tumors contributed to an accelerated, tumor-agnostic FDA approval of dabrafenib plus trametinib for adult and pediatric patients with unresectable or metastatic solid tumors harboring that mutation who had progressed on prior treatment and had no satisfactory alternatives.

That is a complete chain: federal research investment, a trial network built to test biomarker-matched treatment across tumor types, an informatics infrastructure that made molecularly matched assignment possible at national scale, a rigorous clinical finding, a regulatory approval, and a treatment option now available to patients, including children, who previously had none. The scientific and clinical credit belongs to the investigators, the network, and the patients who enrolled. Our point is narrower and is about measurement: this chain is fully documentable, and NIH has no systematic mechanism for documenting it. Cases of this clarity should be captured as a matter of course, not reconstructed after the fact when someone needs an example.

Track societal and policy citation. NIH should develop metrics that capture when funded research is explicitly cited in clinical trial protocols, clinical practice guidelines, public health recommendations, regulatory submissions, federal or state policy documents, and government reports. These citations are direct evidence of real-world influence, they are machine-detectable at scale, and they are almost entirely untracked. This is among the lowest-cost, highest-yield measurement improvements available.

Measure contributions to clinical practice and healthcare delivery directly, including changes to guidelines, improvements in patient outcomes, and improvements in the efficiency or quality of care delivery.

3. Measure Equity of Impact

A finding that improves outcomes for some populations and not others has not delivered its full public value, and NIH should be able to see that in the data.

Require disaggregated reporting. Research outputs reporting clinical or population-level findings should include analyses disaggregated by relevant demographic and socioeconomic factors, and the availability of such disaggregated data should itself be counted as an enhanced impact indicator. This requires that the underlying data be captured in a structured, standardized form from the outset, which connects directly to the common data element recommendations we made in response to the data sharing question. Structured demographic capture in clinical research data, aligned to established standards such as the cancer-relevant USCDI+ data elements, is what makes population-level analysis possible at all. Where it is absent, no amount of downstream analytic effort can recover it.

Benchmark study cohorts against actual disease burden. NIH should measure how well the demographic composition of studied cohorts matches the demographic distribution of the disease in the population. A trial cohort that does not resemble the affected population produces findings of uncertain applicability to the people who most need them. This is measurable: trial accrual data is collected, and disease burden data exists. Reporting infrastructure of the kind NCI maintains for cancer trial accrual

demonstrates that the data required for this comparison is already being captured and simply needs to be analyzed and reported accordingly.

4. Recognize Public Goods and Community Engagement as Impact

Credit the creation of community resources. Public datasets, educational materials, open research infrastructure, and widely adopted community tools deliver diffuse public value that no publication metric captures. NIH should recognize these as impact outputs in their own right.

Reward genuine engagement across the research lifecycle. NIH should evaluate whether research involved patient advocates in study design, whether results were communicated back to affected communities in accessible language, and whether the research question itself emerged from community partnership. That last item is the most consequential and the most rarely achieved: research that answers a question the affected community actually asked is qualitatively different in public value from research that arrives at a community as a finished product.

Require plain-language summaries. Every funded publication should carry a plain-language abstract, on the order of 300 words, genuinely accessible to a non-specialist reader. Compliance should be tracked at the institutional level. This is a small requirement with a large cumulative effect on public accessibility and trust, and it is currently encouraged rather than required, which means it is inconsistently done and frequently done badly.

Add a public engagement section to progress reports. A dedicated section documenting outreach to non-academic stakeholders, patient advocacy organizations, community health settings, and the public would make this work visible and creditable. What is not reported is not valued, and what is not valued does not happen.

5. Measure Economic Impact Systematically

Standardize economic indicators. NIH should develop consistent metrics tracking commercialization outcomes, workforce development and job creation, cost savings to the health system, regional supplier and biotech ecosystem growth, and the enabling effect of NIH investment on subsequent private-sector research activity.

Count workforce development as economic impact. As we noted in response to the training question, NIH trains a scientific workforce that overwhelmingly enters industry, government, and applied settings, where it performs work of direct economic and scientific value. That workforce is an NIH product. It is also, at present, essentially uncounted in any assessment of NIH's economic return.

NIH should also evaluate creating a structured mechanism through which meritorious but unfunded proposals, with the originator's consent, become visible to other Institutes, federal agencies, foundations, and private-sector partners who may wish to support them. NIH generates far more scientifically vetted ideas than it can fund. At present, an idea that does not find support through its initial mechanism simply disappears. A public venue of this kind would both reduce the loss of good

ideas and make the breadth and seriousness of NIH-funded scientific thinking visible to the public. The design challenges are real and would need resolution before implementation, but the concept merits evaluation

Summary

The recommendations in this section reduce to a single argument. NIH's public impact is not primarily a communication problem. It is a measurement problem.

The impact is real. It is happening continuously. It is visible in fragments, in news coverage, in individual accounts, in the occasional traceable success story like a tumor-agnostic approval that gives children with a specific mutation a treatment they did not previously have. What is missing is the infrastructure to capture it systematically, attribute it credibly, and present it in a form that the public, Congress, and NIH itself can act on.

That infrastructure is buildable. NIH funds work of exactly this character routinely, for scientific data, for clinical trial reporting, and, when a public health emergency made it urgent, for national seroprevalence surveillance. The methods are established. The data sources exist. What is required is the decision to treat NIH's own impact as a dataset worth building, and to fund it accordingly. Anything that is not measured cannot be improved and cannot be defended. NIH's public impact deserves both.

8. Other Areas: About Essex Management. Essex Management is a federal contractor and member of the Emmes Group with over 15 years of experience supporting NIH and HHS agencies. Our work spans three integrated practice areas: Biomedical Informatics and Data Science (BIDS), which supports complex NIH data platforms, bioinformatics pipelines, and clinical informatics programs; Modern Technology and Engineering, which delivers cloud-native software development, data engineering, and FISMA-compliant infrastructure; and Program Strategy and Delivery, which supports scientific program operations, stakeholder engagement, and translational research coordination. Together, these capabilities have supported NIH programs including precision medicine clinical trials, immuno-oncology data commons, pediatric cancer registries, clinical trial management and reporting systems, real-world data platforms, and AI-enabled health informatics. We offer the following perspective to help inform the NIH’s initiative for Measuring and Rewarding Scientific Impact.

150. Brian Bauer

I am responding: On behalf of myself

Name: Brian Bauer

Name of Organization: University of Georgia

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/12/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Funders already track policy citations as an impact indicator. The UK's National Institute for Health and Care Research uses the Overton database, which indexes more than 13 million policy documents from over 43,000 organizations, to monitor the policy impact of the research it funds. Altmetric indexes citations in policy as a separate category.

However, what these platforms currently report is an unweighted count. Tunn et al. (2025), evaluated both tools against 110 articles from an NIHR policy research unit, and found that the way citing documents were used as evidence varied considerably. At one end, a finding provided general background. At the other, it supplied the stated rationale for a specific clinical recommendation. The authors concluded that citation analysis alone cannot provide the complete picture. A count records both cases as one citation.

I recommend that NIH develop a weighted measure of policy citation, classifying each citation event on two dimensions. The first is role, or how load-bearing the citation is in the document, ranging from background mention through direct support for a stated recommendation. The second is fidelity, based on if the claim as stated in the policy document is supported by the claim as stated in the cited source. For example, a finding that supplies the rationale for a clinical guideline recommendation represents a different kind of uptake than the same finding listed among background references, and a misstated finding in that position is different again. A count cannot distinguish any of these.

Both dimensions are computable from Overton, GovInfo, and Federal Register text using large language model extraction of the citing passage and the source claim, followed by human adjudication of a

validation subsample to estimate error rates. Building the coding scheme and the labeled corpus could be agency-level work. Because the measure is defined per citation event, it aggregates to whatever unit is under evaluation. The same extraction would also show whether retracted findings continue to appear in policy documents after retraction, by joining the Retraction Watch database to the same corpora.

8. Other Areas: N/A

151. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Research Participant/Patient Advocacy Organization/Patient Advocate/Caregiver

Role: Research Participant/Patient/Advocate/Caregiver

Date Submitted: 8/12/2026

1. Rigor and Reproducibility: * Enforce data sharing policies
* Encourage data depository use and maintenance
* Provide training and incentives for data re-use/secondary data analysis
* Deprioritize funding of incremental improvements to evidence-based interventions. We don't need another study on the Diabetes Prevention Program in 50-59 women in rural Wyoming for instance--just reproduce and spread established findings from the DPP, as one example.

2. Data, Software, and Model Sharing: * Enforce current data sharing policies
* Encourage data depository use and maintenance
* Provide training and incentives for data re-use/secondary data analysis
* Utilize known repositories like GitHub, bioLINCC, dbGaP, and MIMIC rather than creating de novo or bespoke data repositories
* Use and promote the standards and resources promulgated by the Mobilizing Computable Biomedical Knowledge initiative

It does take substantial work by "data janitors" to clean and maintain data sets and help ensure that secondary users understand provenance, mode of collection, metadata, and other nuances. This work largely goes unfunded because it's not included in research grants and the grants run out when the data are ready for sharing and reuse

3. Training and Mentorship: In the same way that primary care providers should be paid an equivalent rate to specialty care providers, those who train the next generation of scientists should be compensated and recognized at a rate comparable to those doing the science.

4. Collaboration: Collaboration requires robust infrastructure that needs to be built, maintained and sustained. In the 2000s, Several NIH agencies supported research infrastructures for cancer, mental health, and cardiovascular health for years, but eventually paused these and in some cases, restarted similar endeavors in the 2020s. What if we just acquiesced to the need for ongoing, publicly funded research infrastructures so that we don't need to build new roads every time the car needs to go somewhere?

The peer review process offers an opportunity to identify overlapping/similar initiatives and could actually be a source for brokering collaboration rather than competition

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: I once asked an NIH program officer what it would take for NIH to start looking at research impact vs. length of CV or prominence of publications as a yardstick for research success. If we have a successful intervention to help bolster participation in cancer clinical trials (a longstanding deficit of biomedical research), the intervention should be promulgated and the researchers who built the intervention and fostered its implementation would deserve merit (tenure credits internally). We have an R01 mindset with less attention to the ROI. Particularly for embedded researchers (i.e., in health systems), there should be a mechanism to support dissemination and implementation efforts within the health system for a successful intervention. PCORI does something similar to this with its D&I and HSII mechanisms and could serve as a blueprint.

8. Other Areas: Impact is hastened when study populations are sufficiently large and representative. Research recruitment remains a chokepoint for many studies. What could be done via a public-private campaign to reinforce the value of research to society and encourage people to participate? Altruism, but also financial incentives, could go a long way to accelerate the impact of taxpayer funded research.

152. Chaitanya Jain

I am responding: On behalf of myself

Name: Chaitanya Jain

Name of Organization: University of Miami

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/12/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: The following addresses how to ensure that funds spent on competitive renewals are best distributed.

I have long felt that there has not been any good way to measure the effectiveness of NIH funding when grants come up for competitive renewal. Many application reviews tend to be biased towards an investigator's reputation or a long list of publications in which the primary contribution was mostly by others. As a result, investigators who are not as well established or have fewer publications, albeit with with greater effort, are disproportionately left out. To provide a quantitative metric for evaluating applications, I would measure the impact of funding of an investigator or laboratory using the following equation:

Impact = Sum(publications x lab contribution percentage x Impact score)/Dollars awarded.

In this model, I am evaluating the funding impact by measuring the number of research publications generated by a laboratory over the funding duration, each multiplied by the percentage contribution of the laboratory and the impact of the publication (e.g., impact factor). The numerator should then be divided by the total funds received by the laboratory. The final number will help to rank the impact of each grant, and hopefully, create a more equitable basis to determine how NIH could distribute its funds to deliver the most impact.

A few notes on the % contribution metric. For any manuscript, the contribution of all authors in any

paper should add up to 100%. Then, in the equation above, the total contribution of all members of the laboratory who were funded by a grant should be used. Only research publications that generate new findings should be considered, with reviews excluded.

To illustrate this point, I would like to use some hypothetical examples. Let's say that a publication has five authors, of which only the first and the fifth have been funded by an application that is being considered for re-funding. In this example, it would be appropriate to assume that project was initiated by the lab whose PI has submitted a renewal application and that the sum of the contributions of the first and fifth authors might be 60-80%. In a second example, let’s assume there are eight authors, and the persons from the lab that is being considered for re-funding are the third and sixth authors. In that example, I will assume that the work was initiated by a different lab and the bulk of the work was done by the other authors, so the contribution of the third and the sixth authors might only be 5-10%.

I acknowledge that determining the contribution of each author will not be simple and it might need other opinions to decide on the best way to do this. Moreover, basing a grant's impact on publications alone may not be appropriate for all applications. Nonetheless, I hope that whatever I have written, however, imperfect, will help to stimulate a discussion on the best way to distribute NIH funds in a manner that quantifies research impact normalized by the amount of funds received.

153. Institute for Progress

I am responding: On behalf of an organization

Name: Daniel Turner-Evans

Name of Organization: Institute for Progress

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: Replicating research findings to confirm their legitimacy is an important part of the scientific process. When determining how much to spend on replication, NIH should carefully consider the return on investment from each dollar spent. Under realistic assumptions, NIH could productively spend about 1.4% of its annual budget on targeted replication studies before hitting negative returns relative to funding new science.

2. Data, Software, and Model Sharing: NIH recently concluded a competition to develop a metric to quantify data sharing, the S-Index, and should now encourage the scientific community to adopt it. As a first step to make PIs aware of the S-Index and to build their comfort with the metric, NIH should pick one of the S-Index metrics from the competition and require that PIs list it in the Data, Management, and Sharing plan for any of their grants.

NIH should also offer incentives for PIs to improve the quality and accessibility of the data, software, and models that they share. PIs can already request funds for Data Management and Sharing, and additional funding opportunities should be made available for personnel or services, including AI tools, to perform these functions

3. Training and Mentorship: Academic graduate programs rarely focus on preparing students for careers in the biomedical industry. NIH should incentivize programs that fill this gap by, for example, incentivizing Ph.D. programs that offer industry internships, similar to the University-Industry Demonstration Partnership Industry-Integrated Ph.D. Scholars Program just launched at NSF.

NIH should also incentivize students to learn how to use state-of-the-art tools, including LLMs and field-specific AI models. The agency should expand support for summer schools and courses that teach students these methods, like the Marine Biological Labs Courses. These courses also offer an excellent opportunity for students to develop budding collaborations and inspire interdisciplinary work.

NIH should quantify how well students learn how to pursue independent research. Many Ph.D. programs already ask students to write an NIH-style grant proposal for their qualifying exam, and NIH should develop guiding rubrics for these exams and dissertation committee assessments and then collect evaluations from faculty who serve on those committees.

NIH should also determine if students are being given sufficient opportunities to collaborate. The number of co-PI publications could serve as a starting point for this evaluation.

4. Collaboration: To incentivize collaboration, NIH could offer additional funding opportunities for multiple PIs, such as the U19 mechanism. The agency could also work to develop a new citation metric that more heavily weights publications with multiple PIs than those with a single PI.

5. Entrepreneurship and Translation: NA

6. Foundational Scientific Exploration: NIH should support a portfolio of mechanisms that can promote both more novel / higher-risk research and more incremental research. Some funding mechanisms are more appropriate for funding conventional research (such as the R01), whereas others are designed to allow for flexibility. For example, person-based grants like the NIH R35 grant provide longer-term funding and flexibility that can allow for innovative research. NIH could also pilot smaller, fast-timeline grants for nascent ideas, as the National Science Foundation does through its EArly-Concept Grants for Exploratory Research (EAGER) program. NIH institutes and centers can also experiment with integrating "golden tickets" into conventional award processes, which increases the novelty of awards by giving reviewers a "golden ticket" to select a project even when the other reviewers disagree.

7. Public Impact: NA

8. Other Areas: Clear metrics are essential for evaluating the success of NIH funding mechanisms and the impact of federally funded science, and carefully structured incentives are necessary for aligning research efforts with national goals. However, when considering how to develop and apply new metrics, NIH should be aware that applying new metrics can add administrative burden on staff or grantees and should therefore perform careful cost-benefit analysis for each new metric before implementing it. We therefore primarily recommend incentives that NIH can use to stimulate more novel, open, and collaborative work and suggest a balance of options that NIH can pursue to shape the research enterprise.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/IFP-Response-to-Metrics-and-Incentives-RFI.pdf

About IFP

The Institute for Progress (IFP) is a non-partisan think tank focused on innovation policy. Our organization works to accelerate and shape the direction of scientific, technological, and industrial progress. Headquartered in Washington D.C., IFP works with policymakers across the political spectrum to make it easier to build the future in the United States.

Introduction

The National Institutes of Health requested input on metrics and incentives that reflect the collaborative and interdisciplinary nature of the modern biomedical research enterprise. Clear metrics are essential for evaluating the success of NIH funding mechanisms and the impact of federally funded science, and carefully structured incentives are necessary for aligning research efforts with national goals. However, when considering how to develop and apply new metrics, NIH should be aware that applying new metrics can add administrative burden on staff or grantees and should therefore perform careful cost-benefit analysis for each new metric before implementing it. We therefore primarily recommend incentives that NIH can use to stimulate more novel, open, and collaborative work and suggest a balance of options that NIH can pursue to shape the research enterprise.

Rigor and Reproducibility

Replicating research findings to confirm their legitimacy is an important part of the scientific process. When determining how much to spend on replication, NIH should carefully consider the return on investment from each dollar spent. Under realistic assumptions, NIH could productively spend about 1.4% of its annual budget on targeted replication studies before hitting negative returns relative to funding new science.

Data, Software, and Model Sharing

NIH recently concluded a competition to develop a metric to quantify data sharing, the S-Index, and should now encourage the scientific community to adopt it. As a first step to make PIs aware of the S-Index and to build their comfort with the metric, NIH should pick one of the S-Index metrics from the competition and require that PIs list it in the Data, Management, and Sharing plan for any of their grants.

NIH should also offer incentives for PIs to improve the quality and accessibility of the data, software, and models that they share. PIs can already request funds for Data Management and Sharing, and additional funding opportunities should be made available for personnel or services, including AI tools, to perform these functions

Training and mentorship

Academic graduate programs rarely focus on preparing students for careers in the biomedical industry. NIH should incentivize programs that fill this gap by, for example, incentivizing Ph.D.

programs that offer industry internships, similar to the University-Industry Demonstration Partnership Industry-Integrated Ph.D. Scholars Program just launched at NSF.

NIH should also incentivize students to learn how to use state-of-the-art tools, including LLMs and field-specific AI models. The agency should expand support for summer schools and courses that teach students these methods, like the Marine Biological Labs Courses. These courses also offer an excellent opportunity for students to develop budding collaborations and inspire interdisciplinary work.

NIH should quantify how well students learn how to pursue independent research. Many Ph.D. programs already ask students to write an NIH-style grant proposal for their qualifying exam, and NIH should develop guiding rubrics for these exams and dissertation committee assessments and then collect evaluations from faculty who serve on those committees.

NIH should also determine if students are being given sufficient opportunities to collaborate. The number of co-PI publications could serve as a starting point for this evaluation.

Collaboration

To incentivize collaboration, NIH could offer additional funding opportunities for multiple PIs, such as the U19 mechanism. The agency could also work to develop a new citation metric that more heavily weights publications with multiple PIs than those with a single PI.

Foundational Scientific Exploration

NIH should support a portfolio of mechanisms that can promote both more novel / higher-risk research and more incremental research. Some funding mechanisms are more appropriate for funding conventional research (such as the R01), whereas others are designed to allow for flexibility. For example, person-based grants like the NIH R35 grant provide longer-term funding and flexibility that can allow for innovative research. NIH could also pilot smaller, fast-timeline grants for nascent ideas, as the National Science Foundation does through its EArly-Concept Grants for Exploratory Research (EAGER) program. NIH institutes and centers can also experiment with integrating “golden tickets” into conventional award processes, which increases the novelty of awards by giving reviewers a “golden ticket” to select a project even when the other reviewers disagree.

154. David Piston

I am responding: On behalf of myself

Name: David Piston

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: To improve reproducibility, replication studies should be reviewed and funded on their own terms and not compete against novel-science R01s for the same reviewers' attention. We would also need to count replication studies legitimate scientific outputs, with dedicated paths to publication in the same place that the original study was published. If replication studies are underfunded and siloed, they will become a career backwater where labs will do more when they do not have novel science grants – leading to weaker replication studies.

2. Data, Software, and Model Sharing: Data/software sharing is one place where simple metrics, such as was it deposited, does the code run, is it used by others? This check-box stuff can be verified by an NIH program officer, but trying to rank sharing "quality" would invite gaming.

3. Training and Mentorship: Mentorship is of utmost importance, yet it resists quantification almost entirely. Output measures such as job placement mainly reflect trainee talent, not mentoring effort, and metrics for this area can punish people who take on harder cases. A structured and qualitative approach would be confidential trainee letters read by people who've actually mentored. This could be informed by the approach used by HHMI, where lab staff are surveyed extensively and the data are taken seriously. Mentorship evaluation is a category where people should be punished for coming up short but not rewarded for pumping out more PhDs.

4. Collaboration: Given the current interdisciplinary requirements for biomedical research, collaboration and team science should be a top priority. We have not found a clean fix for individual attribution in team science, and perhaps it would be better for NIH to accept this rather than pretend a formula can solve it. CRediT-style contribution statements are a useful input, but turning this into a “point system” could be a real deterrent from collaboration. Instead, it takes careful evaluation by experts to assess whether scientists are embracing collaborative science by reading the papers and evaluating the data from the collaborative teams.

5. Entrepreneurship and Translation: For basic biomedical research, entrepreneurship and translation should sit adjacent to the core reward system, not inside it. If these areas become necessary for judging the impact of basic science research, folks are likely to start chasing patentable results rather than focusing on rigorous and reproducible biological results.

6. Foundational Scientific Exploration: We should place substantially greater emphasis on the quality of science rather than the quantity or perceived prestige of publications. This should include evaluating the

rigor of the work, reproducibility, methodological innovation, the quality and independence of peer review, and the significance of the questions being addressed. Of course, the biggest issue with evaluating impact is that any given quality is a judgment call and figuring out if something/someone is actually good takes experts spending real time, which is at a premium. Any efforts to produce simple indicators to replace publications and citations as measures used to evaluate funding research will ultimately come up somewhat short. No qualitative-judgment-over-metrics system can work unless reviewers actually have time to read things carefully and think about it deeply. An effective system would need to account for limited reviewer bandwidth and not just add new review categories to already overloaded study sections. Similarly, the scientific community should also rethink the peer-review system itself. High-quality reviewing is an essential scholarly contribution, yet it is largely invisible in current academic evaluation systems. Recognizing meaningful reviewing, and particularly sustained, thoughtful, and expert reviewing, could help create incentives for scientists to invest the time necessary to maintain the quality of scientific literature.

7. Public Impact: I am skeptical about any attempt to simplify measurement of scientific impact of NIH-funded research. Public impact is almost impossible to assign to any one set of work. Building scientific progress is like building a pyramid – adding one block at the base doesn't change the height of the pyramid, but the bigger the base, the higher the pyramid can rise. Is the one block at the bottom less impactful than the one at the top, even though the impact on human health will come from the top block? Health impacts play out on wildly different timescales, and evaluating everything by a single metric ends up favoring flashy short-horizon work that will end up killing future advances. Because it builds the base of our research to health impact pyramid, foundational and high-risk science is poorly suited to qualitative evaluation over traditional metrics here. Basic biomedical science would benefit from longer award periods, real freedom to change direction, and renewal evaluation that asks, "was this a good question, and was it pursued rigorously" instead of "did you hit your milestones." The NIH R35/MIRA model is successful proof of this concept.

8. Other Areas: There are significant problems exist in the current publication landscape and find a strong consensus in our department that publications have become a dubious measure of impact. There has been an astronomical ever-growing number of journals, driven by an expansion of low-quality publications, coupled with unreliable citation-based metrics and a review system in which scientists are asked to evaluate an enormous number of manuscripts without compensation. Inevitably, this compromises the depth and quality of peer review, and inevitably, the quality of science in general. The expansion of OA publishing model, coupled in many cases with very high OA charges, has created incentives for publishers to maximize the number of papers processed rather than necessarily the number of papers that represent truly important advances. This proliferation of journals and the focus of these journals to "game" citation metrics makes it difficult to distinguish scientific impact from sheer publication volume. We agree that to improve the quality of science, we need to move away from asking "how much has this person published, and where?" towards asking "what is the quality of their scientific contribution, how rigorously has it been evaluated, and how much has it advanced the field?"

155. Cleveland Clinic

I am responding: On behalf of an organization

Name: blair barnhart hinkle

Name of Organization: Cleveland Clinic

Type of Organization: Healthcare Institution

Role: Organizational Official

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: Attached please find the comments of Cleveland Clinic

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/8_13_2026-NIH-Comment-letter_Gold-Standards.pdf

Description: Comment letter from Cleveland Clinic

Serpil C. Erzurum, M.D
Chief Research and Academic Officer

August 11, 2026

National Institutes of Health
6705 Rockledge Drive
Bethesda, MD 20892

Re: Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Submitted via: https://osp.od.nih.gov/comment-form-measuring-and-rewarding-scientific-impact/

Cleveland Clinic is a not-for-profit, integrated healthcare system dedicated to patient-centered care, teaching, and research. With a footprint in Northeast Ohio, Florida, and Nevada, Cleveland Clinic Health System operates 23 hospitals, including a main campus near downtown Cleveland, with more than 6,690 beds, and 276 outpatient locations. Cleveland Clinic employs over 5,700 physicians and researchers, and 16,500 nurses. Last year, our system cared for 3.3 million patients, including 13.7 million outpatient visits and 323,000 hospital admissions and observations. Below are our comments on the draft guidance.

Cleveland Clinic is home for laboratory-based, translational, computational, clinical and implementation science across Cleveland Clinic's global health system. Cleveland Clinic's research footprint, which includes facilities in Cleveland and Florida, spans nearly one million square feet. Our research enterprise includes over 2,400 researchers and support personnel, and in 2025 our annual research funding exceeded $479 million.

Cleveland Clinic Research encompasses ten scientific departments of Biomedical Engineering; Cancer Sciences; Computational Life Sciences; Heart, Blood and Kidney Research; Genomic Sciences and Systems Biology; Inflammation and Immunity; Microbial Sciences in Health; Neurosciences; Ophthalmic Research; and Quantitative Health Sciences. Cleveland Clinic Research is also committed to the training and career development of young research scientists and physician-scientists. In 2017, the Cleveland Clinic BioRepository was established with a vision to fuel translational research and innovation that leads to improved clinical outcomes and the betterment of patients' lives. The mission of the BioRepository is to safely maximize the value of human samples by supplying high-quality, well-annotated specimens to researchers whose discoveries will be essential for the evolution of personalized medicine.

We appreciate the opportunity to provide comments in response to the National Institutes of Health (NIH) Request for Information (RFI) on Measuring and Rewarding Scientific Impact (NOT-OD-26-087). We commend NIH's efforts through its Replication and Reproducibility Initiative and Plan to Drive Gold Standard Science to identify approaches that more accurately recognize the diverse activities that contribute to scientific advancement and public benefit. NIH has specifically requested feedback regarding measurable indicators, implementation approaches, potential benefits, and unintended consequences related to modern biomedical research. We offer the following

recommendations based on perspectives from active investigators engaged in basic, translational, and clinical research.

Rigor and Reproducibility

Scientific rigor and reproducibility are foundational elements of impactful biomedical research. While investigators currently address rigor and reproducibility requirements in grant applications, we encourage NIH to focus on approaches that influence research culture and daily scientific practices rather than solely emphasizing application requirements.

Strong training environments remain among the most effective mechanisms for promoting rigorous and reproducible research. NIH should consider indicators that evaluate the quality and extent of training in experimental design, statistical rigor, transparency, and reproducibility practices. Additionally, NIH should explore mechanisms to support and reward replication studies, which often provide substantial scientific value but receive limited recognition in traditional research assessment frameworks.

Importantly, NIH should recognize that uncertainty regarding funding availability may create pressure that unintentionally discourages careful, deliberate scientific investigation. Clear and consistent communication regarding research priorities and funding expectations can help reduce these pressures and reinforce rigorous scientific practices.

Data, Software, and Model Sharing

Cleveland Clinic strongly supports data sharing as a critical component of scientific advancement. Shared datasets, software tools, and analytical models enable validation of results, promote transparency, facilitate collaboration, and accelerate discovery.

To encourage meaningful participation, NIH should recognize that data sharing requires substantial investments of investigator time and institutional resources. Measurable indicators could include the number of datasets deposited, dataset utilization rates, adoption of common data elements, and documented reuse of shared resources. However, these expectations should be accompanied by appropriate support.

We recommend that NIH permit dedicated budget allocations for data-sharing activities, including personnel effort required to curate, prepare, and submit data. NIH should also continue efforts to simplify repository submission processes and support development of common data elements across clinical and basic research disciplines. Without these improvements, administrative complexity may become a significant unintended barrier to compliance and meaningful participation.

Training and Mentorship

The future of biomedical research depends upon a stable and confident workforce. We are increasingly hearing concerns from trainees and early-career investigators regarding the long-term stability of federal research support. This uncertainty may discourage talented individuals from pursuing research careers at a time when biomedical innovation depends on attracting and retaining highly qualified investigators.

NIH should consider indicators such as trainee retention, career progression, fellowship participation, and application rates for career development and training awards. Such measures would provide meaningful insight into the workforce and the effectiveness of NIH training investments.

In addition, NIH should recognize high-quality mentorship as a significant scientific contribution. Effective mentors help ensure rigor, foster collaboration, develop future scientific leaders, and sustain the biomedical research enterprise. Institutions and investigators who demonstrate excellence in mentorship should receive appropriate recognition within NIH evaluation frameworks.

Collaboration

Modern biomedical research increasingly depends upon collaborative, interdisciplinary, and multi-institutional teams. Traditional measures of scientific productivity often fail to capture the value of these contributions.

NIH should develop assessment approaches that recognize meaningful participation in team science. Potential indicators could include participation in multi-principal investigator grants, leadership or coordination roles within collaborative initiatives, contributions to shared scientific resources, and evidence of successful multi-institutional partnerships.

We also encourage NIH to consider potential unintended consequences of policies that may inadvertently discourage collaborative applications, including domestic and international collaborations. For example, policies that limit grant submission opportunities, create uncertainty around eligibility or administrative requirements, or increase the complexity of cross-border partnerships could reduce participation in multi-principal investigator proposals and ultimately diminish incentives for team science. NIH should consider addressing challenges specific to international collaborations, including differences in regulatory requirements, data-sharing expectations, institutional approvals, and timelines that may affect investigators' ability to participate fully in collaborative research. Monitoring trends in collaborative grant submissions, including applications involving international partners, may help NIH identify such impacts and ensure that policies continue to support productive scientific collaboration.

Entrepreneurship and Translation

Translation of scientific discoveries into therapies, diagnostics, technologies, and clinical applications represents a critical component of NIH's mission to improve human health.

Artificial intelligence and advanced computational approaches are increasingly creating opportunities to accelerate the movement of discoveries from laboratory investigations to real-world clinical implementation. NIH should consider indicators that measure successful translation, including development of clinical decision-support tools, commercialization activities, implementation of research findings into clinical practice, and responsible application of emerging technologies.

Future NIH assessment frameworks should reward investigators who successfully bridge basic science and clinical application while maintaining rigorous standards for scientific validity, safety, and ethical implementation.

Foundational Scientific Exploration

Many transformative discoveries arise from basic science investigations and high-risk, high-reward research that may not produce immediate outcomes. However, these efforts remain essential to long-term scientific progress.

NIH should continue to support and recognize foundational scientific exploration by valuing scientific creativity, innovation, and willingness to pursue challenging research questions. Stable funding environments and transparent expectations will encourage investigators to undertake ambitious projects that may ultimately yield paradigm-shifting advances.

Assessment frameworks should avoid overemphasizing short-term productivity metrics that may inadvertently discourage innovative or exploratory research.

Public Impact

Biomedical research generates value extending far beyond publications and citations. NIH should continue exploring approaches to assess public impact, including contributions to health outcomes, patient education, community engagement, public trust, accessibility of scientific knowledge, and societal benefit.

Effective communication strategies increasingly rely on both traditional community engagement and modern communication platforms, including digital content, webinars, public education resources, and social media. Meaningful indicators could include dissemination activities, public engagement efforts, adoption of evidence-based health information, and demonstrated influence on health behaviors or clinical practice.

Evaluation systems should recognize investigators and institutions that successfully translate scientific knowledge into tangible public benefit.

Conclusion

As NIH continues to develop approaches for measuring and rewarding scientific impact, it is important that future frameworks reflect the collaborative, rigorous, and mission-driven nature of modern biomedical research. The most meaningful indicators will be those that encourage behaviors that strengthen the biomedical research enterprise, including mentorship, transparency, collaboration, data sharing, innovation, and public engagement.

We appreciate NIH's commitment to ensuring that scientific incentives align with activities that ultimately improve human health and advance discovery. Thank you for your consideration of our feedback. Should you need any further information, please contact me.

Sincerely,



Serpil C. Erzurum, MD
Alfred Lerner Memorial Chair in Innovative Biomedical Research
Cleveland Clinic Chief Research & Academic Officer

156. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: The NIH's decision to collapse qualities of the Investigator - prior productivity, accomplishments, ability to deliver etc. - into a simple binary of acceptable or unacceptable undermines the mission to support impactful science. I cannot think of other endeavors - in business, the arts, or anything else - where the abilities of the grantee are not evaluated along with the proposed work. (Not to mention the double standard of having the investigator track record rank highly for new investigators in mechanisms such as the DP2. If the qualities of an early career investigator be a matter worth considering in funding decisions, why should it be different for a mid-career or senior investigator?) I understand the fairness argument - that it is unacceptable to simply give a researcher the "benefit of the doubt" - but the pendulum has swung too far in terms of the restrictions placed on what grant reviewers are "allowed" to consider. I think this results in a different kind of unfairness.

157. Ian Hagemann, MD, PhD

I am responding: On behalf of myself

Name: Ian Hagemann, MD, PhD

Name of Organization: Washington University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: I am concerned about the erosive effect of AI on data and result integrity. There could be counterintuitive adverse effects of data sharing. If a lab has expensively acquired data, it will have a better understanding of its limitations. I'm concerned that AI chatbots will make it possible to repeatedly reuse public data, exposing the public to the risks of false discovery due to repeated hypothesis testing. I see this in peer review requests that I get, where unknown research groups are grinding against public data (typically TCGA or UKBB). Even if there is internal correction for multiple hypothesis testing, the reduplicative use of the data will eventually lead to type I errors. AI overall makes it possible to iteratively create grants, models, papers and presentations. This can be both good and bad but is currently completely uncontrolled. People use AI to write their papers, then reviewers use AI to read them, then hiring committees use AI to evaluate these scientists' CVs... what if the AI is just wrong about any of this? How will we know?

3. Training and Mentorship: Realistically an assistant professor needs to secure at least two R01 equivalents. As it starts to become unrealistic to maintain this cadence, talented individuals will drop out of the tenure track pool. The training and recruitment pipeline into the tenure track is at risk currently due to low paylines and uncertain funding environment/priorities.

With regard to paylines, when 5% or less of grants are funded, the probability of successfully securing enough funding to keep a lab running becomes impossibly low. At such low paylines, even meritorious grants can fail by chance. NIH scientific review is effective and improves the quality of science but it definitely has a component of noise. There is subjectivity in NIH scoring (intra/inter-observer variability) and so a grant that might have been funded in one round, or with a given set of reviewers, can fail, not because it needed to be improved, but simply by chance. Or grants that meet all standards of scientific review ("fundable" on scientific merit) may be denied repeatedly simply because there is not enough money. What are these investigators supposed to do? Keep trying in case they get lucky a different time? Reduplicatively submit more and more grants in order to get more darts to throw? It is not efficient to expect applicants to invest resources not only in their best ideas, but also in their 2nd, 3rd, 4th, 5th best ideas.

With regard to funding priorities, the vast majority of scientists are not ideologically driven and are pursuing a combination of pure truth, societal benefit and personal career stability. They rely on the

principle that their field will not experience major shifts in fundability over their career (or only on a timeframe where they can reasonably pivot). In the current environment it seems entire areas of research are at risk of being de-emphasized on short notice. Regardless of the societal merit or legal defensibility of this approach, there will be a chilling effect on recruitment of talented scientists into the tenure-track pipeline, if there is a risk that the rug will be pulled out from under their research area at a moment's notice.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

158. Suresh Palaniyandi

I am responding: On behalf of myself

Name: Suresh Palaniyandi

Name of Organization: Henry Ford Health System/Wayne State University

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: NIH should reward researchers who demonstrate that their findings are reproducible, not simply those who publish the greatest number of papers. Practical indicators could include successful replication by independent laboratories, publication of negative or confirmatory results, use of validated methods, and transparent reporting of experimental conditions. NIH could provide small supplemental awards for independent replication studies and allow investigators to identify their most reproducible or independently validated findings in biosketches and progress reports. Importantly, these measures should recognize rigorous work without creating additional paperwork for investigators.

2. Data, Software, and Model Sharing: NIH should recognize and reward the act of making scientific data, software, models, and protocols openly available, since sharing itself requires time, effort, documentation, and resources and is the essential first step toward reproducibility and reuse. Additional indicators could then capture the broader value of these resources, such as downloads, reuse of datasets, requests for animal models or reagents, adoption of protocols or software, and publications generated using shared resources. Investigators who create and maintain useful shared resources should be able to list these as major scientific contributions comparable to publications. NIH could also provide small supplements or dedicated support for preparing, documenting, and maintaining high-value resources so that data sharing does not become an unfunded administrative burden.

3. Training and Mentorship: NIH should recognize mentorship across the entire biomedical research pipeline, including high school, undergraduate, master's, PhD and MD students, postdoctoral fellows, and other degree holders entering or advancing in research. The number of individuals trained is an important indicator of a researcher's commitment to developing the workforce, but it should be considered together with the quality, duration, and depth of mentorship. Useful outcomes could include trainee publications and presentations, research skills gained, fellowship or grant applications, continued education, transition to independent positions, and career advancement in biomedical research or related fields. Importantly, dedicated mentorship of a smaller number of trainees can have substantial impact and should not be undervalued simply because another investigator has trained a larger number. NIH should recognize both the reach of training and the quality and sustained commitment of mentorship, without creating substantial additional reporting burden.

4. Collaboration: Modern biomedical research increasingly depends on collaboration, yet current measures often give disproportionate credit to the PI, corresponding author, or senior author. NIH

should recognize building productive collaborations as a scientific achievement itself, particularly when those collaborations lead to joint grant applications, funded projects, publications, abstracts, presentations, shared resources, new methods, or multidisciplinary research directions. Contributions such as providing specialized expertise, developing methods, generating unique models or datasets, and performing critical experiments should also receive meaningful recognition regardless of authorship position.

Investigators should be able to document important collaborative accomplishments in their biosketches and progress reports, including collaborative grant submissions and scientific outputs. NIH could further incentivize collaboration through supplemental funding, recognition of sustained productive partnerships, or consideration of collaborative accomplishments during grant review. This would encourage investigators to share expertise and build strong research teams rather than compete primarily for individual credit, while still recognizing leadership by PIs and senior investigators.

5. Entrepreneurship and Translation: Patents, licensing, startup companies, and commercialization are important and measurable indicators of successful translation and should continue to be recognized and rewarded. NIH should also recognize additional steps that move discoveries toward practical application, including development of biomarkers, diagnostics, therapeutic targets, assays, research tools, clinical protocols, and technologies; industry partnerships; progression to preclinical or clinical testing; and adoption by other researchers or healthcare systems. Importantly, translation often requires years and contributions from multiple investigators. Researchers should therefore receive credit for meaningful milestones along the translational pathway, not only for the final commercial product. NIH could also provide targeted supplemental funding to help promising discoveries cross the difficult gap between an academic finding and the proof-of-concept work needed for commercialization or clinical development.

6. Foundational Scientific Exploration: NIH should recognize that positive, negative, nonsignificant, and unexpected results can all contribute to scientific knowledge when the research is rigorous and well designed. Negative or nonsignificant findings can prevent unnecessary duplication, challenge existing assumptions, and guide future research, yet these results are often difficult to publish and therefore receive little recognition. This strong emphasis on positive and novel findings may also unintentionally contribute to publication bias and create pressure that can undermine research rigor and integrity. NIH should encourage and give credit for communicating rigorous negative and nonsignificant findings through publications, brief reports, data repositories, or other appropriate scientific forums.

Publications should remain an important measure of scientific productivity, but they should not be treated as the sole measure of scientific achievement. Building collaborations, training and mentoring researchers, developing methods and models, sharing data and resources, presenting findings, contributing reviews or scientific perspectives, and other scholarly activities all contribute to knowledge building. NIH evaluation should recognize this broader range of contributions so that investigators are rewarded for advancing science, not simply for producing a larger number of positive-result publications.

7. Public Impact: Public impact should be measured broadly and over an appropriate time frame. Useful indicators could include changes in clinical practice, incorporation into guidelines, development of diagnostics or treatments, use of research findings by healthcare systems, public-health applications, patents or licensing, and educational or community outreach. NIH should recognize that basic

discoveries may require many years before producing measurable health outcomes; therefore, investigators should not be penalized when immediate clinical impact is unrealistic.

8. Other Areas: NIH should consider measuring scientific value created per research dollar, rather than relying heavily on publication quantity. Investigators at smaller institutions or with limited resources may produce highly rigorous and impactful science without generating large numbers of publications. Indicators could include discoveries produced, resources shared, collaborations enabled, trainees developed, subsequent research stimulated, and progress toward clinical or public-health applications relative to the resources available.

NIH should also recognize research that saves the scientific community time and money. Demonstrating that a widely assumed mechanism does not reproduce, validating an inexpensive alternative method, sharing a model that prevents other laboratories from recreating it, or identifying why an experimental approach fails can have substantial value even if these contributions generate fewer citations. Preventing unnecessary research expenditure is itself an important return on NIH investment.

Finally, any new system should remain simple. NIH should avoid replacing publication counts with dozens of new metrics that increase investigators' administrative workload. A practical approach could allow researchers to select three to five major contributions from the previous funding period and briefly document their measurable impact—such as independent validation, data or model reuse, collaboration, trainee development, translation, or public benefit. Reviewers could then evaluate the quality and impact of those contributions rather than relying primarily on publication and citation counts.

I especially recommend including that last paragraph. It addresses a major risk in this initiative: NIH could create 20 new metrics and inadvertently make life harder for the very "average researcher" it is trying to recognize. Your proposal—“show me your 3–5 most valuable contributions and evidence of their impact”—is simple, feasible, and gives researchers multiple ways to demonstrate success.

159. American Society of Human Genetics

I am responding: On behalf of an organization

Name: Karina Miller

Name of Organization: American Society of Human Genetics

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: The American Society of Human Genetics (ASHG) appreciates the opportunity to respond to NIH's Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087). We offer these comments in our capacity as the world's largest professional organization for human genetics research. Our membership includes researchers, academics, physicians, technicians, genetic counselors, students, and others with an interest in human genetics and genomics.

Human genetics and genomics provide a strong example of how rigorous research and team science work together to produce impactful discoveries. Advances in our field often result from contributions by large multi-institutional teams, longitudinal data generation and stewardship, engagement with research participants and patient communities, and translation of discoveries into clinical care. ASHG welcomes NIH's efforts to develop scientific assessment frameworks that better recognize and incentivize the collaborative, rigorous, and mission-driven nature of modern biomedical research.

At the same time, ASHG encourages NIH to consider changes to indicators and incentives through a coordinated framework that benefits from the expertise of NIH's advisory councils and advisory committees. Because research evaluation policies influence funding decisions, workforce development, institutional investments, and scientific culture, they should be developed holistically rather than through a series of discrete policy changes that may create unintended or conflicting incentives. Engagement with NIH peer advisory groups can help ensure that proposed updates are evaluated in the broader context of the agency's mission and the long-term health of the biomedical research enterprise.

We also emphasize the importance of transparent peer review and peer advice in assessing scientific quality and impact. While alternative indicators can complement traditional evaluation methods, they should not replace expert scientific judgment. Maintaining the integrity, independence, and credibility of peer review is essential to ensuring public trust in federally funded research and to maximizing the value and impact of research investments.

ASHG appreciates NIH's consideration of these comments and looks forward to ongoing dialogue as NIH continues this work. For additional information, please contact Karina Miller, PhD, at kmiller@ashg.org.

Description: The American Society of Human Genetics (ASHG)'s response to NOT-OD-26-087.

160. Kirsten Davison

I am responding: On behalf of myself

Name: Kirsten Davison

Name of Organization: Boston College

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Response to NIH Request for Information NOT-OD-26-087: Measuring and Rewarding Scientific Impact

Respondent: Kirsten K. Davison, PhD, Associate Dean for Research, Boston College School of Social Work. I respond as an individual investigator and research administrator; the views are my own.

Sub-topic addressed: Public Impact.

CONTEXT

Research impact, the benefit science delivers to health, the economy, public policy, the environment, and quality of life, is an established field of practice [1] and, more recently, of research [2,3]. Our counterparts in Canada, the Netherlands and, especially, the United Kingdom (UK) have prioritized and measured research impact at a national scale for more than a decade. Because the US does not utilize a national impact framework, many US academics and institutions lack an impact vocabulary and culture [1].

NIH has requested input on a number of areas of modern-day science including "approaches for assessing and recognizing the broader public impact of biomedical research," with a premium placed on "specific and measurable indicators." In the case of societal research impact, which emerges through many hands, over long and uneven timeframes, and in forms that are hard to attribute to any single study or summarize with a specific metric, this framing invites a long and unwinnable debate [3,4].

Bodies from the UK's Research Excellence Framework (REF) to the San Francisco Declaration on Research Assessment have long recognized that research impact resists simple quantitative metrics.

NIH can distinguish between measuring impact, the downstream outcomes that are slow to appear and hard to attribute, and rewarding the practices that reliably produce impact, which we can see today. It can expect and reward knowledge translation, community engagement, open science, and technology transfer. The currency should not be a new citation-focused metric; rather, it should be evidence that a researcher planned and implemented the translational work alongside their high-quality science: the policy brief, the community report, or the licensed technology.

Not all science should be judged this way. Discovery research, curiosity-driven research that is often decades ahead of its application, is the foundation of scientific progress, and forcing it into near-term translation would be a mistake. Such research should be explicitly exempt, not scored as weak.

Recognizing these challenges, and drawing on my personal experiences (as an NIH-funded researcher and associate dean of research) along with the empirical literature, I provide four recommendations to shift researcher behavior toward prioritizing the societal impact of research. Alongside these recommendations, I provide specifics of their implementation, candidate indicators, and the conditions under which they would work across career stages, disciplines and institutions with varying impact capacity.

RECOMMENDATIONS

1. BUILD ON NSF'S "BROADER IMPACTS" CRITERION

In my experience, NSF's emphasis on societal impact (i.e., broader impacts criterion), and the way reviewers weigh it, pushes research markedly closer to impact than the process adopted by NIH [5]. Adapting a proven, domestic model rather than inventing one from scratch is the most immediately achievable step.

1.1 Implementation. Integrate a scored translation and public benefit element into the existing review criteria for applicable activity codes, with an exemption for discovery research. It is critical to provide reviewer guidance and worked examples before the first receipt date; inconsistent reviewer interpretation, not the criterion itself, has been the main failure of impact criteria elsewhere [6].

1.2 Indicators. At the application stage, indicators could include specificity of the named intended users, evidence that the resulting outputs are valuable to said users, and personnel time and funds budgeted for translation.

2. LEARN FROM INTERNATIONAL FRAMEWORKS AND FROM THEIR FAILURES

The UK's REF, Horizon Europe, and Canada's knowledge-translation framework offer hard-won lessons in implementing impact frameworks, including the administrative burden and gaming of the exercise [4,7,8] and the risk of rewarding polished storytelling over real change [9,10]. Australia is a case in point:

its national Engagement and Impact Assessment ran once, in 2018, and was discontinued in 2023 with no replacement as yet identified. NIH should study the implementation of existing and retired impact frameworks, and talk directly with their sponsors, before importing them.

3. BUILD IMPACT LITERACY AND CAPACITY ACROSS RESEARCHERS, INSTITUTIONS AND THE WIDER RESEARCH SECTOR

Impact does not happen by chance. It requires the knowledge, skills, systems and culture to anticipate, design for, support and document change from the outset [11]. Established frameworks and practical resources already exist to strengthen individual and institutional impact literacy [12,13]. Requiring translational planning without building this capacity would undermine the objective.

3.1 Implementation. Incorporate activities to build impact literacy into institutional and career development training awards, fund administrative supplements for translational planning and implementation, and provide shared or regional expertise for institutions with limited translation infrastructure.

3.2 Indicators. Trainee completion of translation or engagement training, presence of institutional support structures, and translational output of trained cohorts relative to comparison groups.

4. REQUIRE INVESTIGATORS TO REPORT TRANSLATIONAL ACTIVITIES ANNUALLY, AND MAKE THOSE REPORTS PUBLIC

Today there are few expectations for what to include in progress reports, so meaningful work goes unrecorded. NIH should instruct researchers to document the knowledge- and technology-transfer practices (open science, co-creation) and products (op-eds, policy briefs) that move science beyond academia. These are not achievements to save for the final progress report; they are real outcomes along the pathway to impact. When shared openly, the research community can study them and build the science of impact alongside its practice.

4.1 Implementation. NIH should collect this information through a structured module in the Research Performance Progress Report using a controlled vocabulary rather than narrative, with persistent identifiers where they exist and a brief free-text field. The resulting data should be released as a public, machine-readable resource linked to RePORTER, as NIH already does for publication linkages.

4.2 Indicators of activity. The module should capture non-journal outputs with persistent identifiers, including policy briefs, community reports, curricula, decision aids, protocols, datasets, software, invention disclosures, and licenses. It should also record engagement roles, such as practice or policy advisory bodies, co-authorship with non-academic partners, and formal comment or testimony to public bodies.

4.3 Indicators of early uptake. Early uptake can be tracked using existing data sources. These include citation of award-linked publications in policy documents and clinical practice guidelines, documented reuse of deposited data and software, and downstream licensing.

4.4 Portfolio-level indicators. NIH can compute a set of indicators from its own records including the share of awards reporting at least one translational output, time from award to first non-journal output, share of award-linked publications cited in guidelines or policy documents, and data reuse rates by repository.

These are indicators of practice and early uptake, not of health, economic, or social outcomes. They should be reported as a profile rather than combined into a composite score, and NIH should state that they are unsuitable for ranking individual investigators.

CLOSING

Changes like these would ripple well beyond NIH, reshaping how universities evaluate and promote their faculty and, over time, helping to build an impact culture. Two strategies could improve NIH's impact strategy. First, treat the initial indicator set as provisional, evaluate the protocol and its outcomes, and disseminate the findings. Second, release the structured reporting data publicly.

REFERENCES

1. Bayley, J. E., & Phipps, D. (2019). Building the concept of research impact literacy. Evidence & Policy, 15(4), 597-606. https://doi.org/10.1332/174426417X15034894876108

2. Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. Journal of the American Society for Information Science and Technology, 64(2), 217-233. https://doi.org/10.1002/asi.22803

3. Penfield, T., Baker, M. J., Scoble, R., & Wykes, M. C. (2014). Assessment, evaluations, and definitions of research impact: A review. Research Evaluation, 23(1), 21-32. https://doi.org/10.1093/reseval/rvt021

4. Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. HEFCE. https://www.ukri.org/wp-content/uploads/2021/12/RE-151221-TheMetricTideFullReport2015.pdf

5. Holbrook, J. B., & Frodeman, R. (2011). Peer review and the ex ante assessment of societal impacts. Research Evaluation, 20(3), 239-246. https://doi.org/10.3152/095820211X12941371876788

6. Derrick, G. (2018). The Evaluators' Eye: Impact Assessment and Academic Peer Review. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-63627-6

7. Stern, N. (2016). Building on Success and Learning from Experience: An Independent Review of the Research Excellence Framework. Department for Business, Energy & Industrial Strategy. https://assets.publishing.service.gov.uk/media/5a803df4e5274a2e8ab4f03d/ind-16-9-ref-stern-review.pdf

8. Martin, B. R. (2011). The Research Excellence Framework and the "impact agenda": are we creating a Frankenstein monster? Research Evaluation, 20(3), 247-254. https://doi.org/10.3152/095820211X13118583635693

9. Watermeyer, R. (2016). Impact in the REF: issues and obstacles. Studies in Higher Education, 41(2), 199-214. https://doi.org/10.1080/03075079.2014.915303

10. Watermeyer, R. (2014). Issues in the articulation of "impact": the responses of UK academics to "impact" as a new measure of research assessment. Studies in Higher Education, 39(2), 359-377. https://doi.org/10.1080/03075079.2012.709490

11. Bayley, J., & Phipps, D. (2019). Extending the concept of research impact literacy: levels of literacy, institutional role and ethical considerations. Emerald Open Research, 1:14. https://doi.org/10.12688/emeraldopenres.13140.1

12. Bayley, J., & Phipps, D. Impact Literacy Workbook. Emerald Publishing. https://www.emeraldgrouppublishing.com/sites/default/files/2020-06/Impact%20Literacy%20Workbook%20Final.pdf

13. Research Impact Canada. Institutional Healthcheck Workbook. https://researchimpact.ca/resources/institutional-healthcheck-workbook/

PROGRAMS AND FRAMEWORKS REFERENCED BY HYPERLINK IN THE TEXT

Canada, Making an Impact: A Shared Framework for Assessing the Impact of Health Services and Policy Research on Decision-Making: https://researchimpact.ca/archived/making-an-impact-a-shared-framework-for-assessing-the-impact-of-health-services-and-policy-research-on-decision-making/

Netherlands, NWO Impact Plan Approach: https://www.nwo.nl/en/impact-plan-approach

United Kingdom, REF impact: https://www.ukri.org/who-we-are/research-england/research-excellence/ref-impact/

Research Excellence Framework 2029: https://2029.ref.ac.uk/

San Francisco Declaration on Research Assessment: https://sfdora.org/

US National Science Foundation, Merit Review (broader impacts criterion): https://www.nsf.gov/funding/merit-review

European Commission, Monitoring, evaluation and impact assessment of Horizon Europe: https://research-and-innovation.ec.europa.eu/strategy/support-policy-making/shaping-eu-research-and-innovation-policy/monitoring-evaluation-and-impact-assessment-horizon-europe_en

Canadian Institutes of Health Research, Guide to Knowledge Translation Planning at CIHR: https://cihr-

irsc.gc.ca/e/45321.html

Australian Research Council, Engagement and Impact Assessment: https://www.arc.gov.au/evaluating-research/engagement-and-impact-ei

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/2026-08-13_NIH_RFI_NOT-OD-26-087_Davison_FINAL.pdf

Description: Davison submission_PDF version with hyperlinks

Response to NIH Request for Information NOT-OD-26-087: Measuring and Rewarding Scientific Impact

Respondent: Kirsten K. Davison, PhD, Associate Dean for Research, Boston College School of Social Work. I respond as an individual investigator and research administrator; the views are my own.
Sub-topic addressed: Public Impact.

The Context
Research impact, the benefit science delivers to health, the economy, public policy, the environment, and quality of life, is an established field of practice [1] and, more recently, of research [2,3]. Our counterparts in Canada, the Netherlands and, especially, the United Kingdom (UK) have prioritized and measured research impact at a national scale for more than a decade. Because the US does not utilize a national impact framework, many US academics and institutions lack an impact vocabulary and culture [1].

NIH has requested input on a number of areas of modern-day science including “approaches for assessing and recognizing the broader public impact of biomedical research,” with a premium placed on “specific and measurable indicators.” In the case of societal research impact, which emerges through many hands, over long and uneven timeframes, and in forms that are hard to attribute to any single study or summarize with a specific metric, this framing invites a long and unwinnable debate [3,4]. Bodies from the UK's Research Excellence Framework (REF) to the San Francisco Declaration on Research Assessment have long recognized that research impact resists simple quantitative metrics.

NIH can distinguish between measuring impact, the downstream outcomes that are slow to appear and hard to attribute, and rewarding the practices that reliably produce impact, which we can see today. It can expect and reward knowledge translation, community engagement, open science, and technology transfer. The currency should not be a new citation-focused metric; rather, it should be evidence that a researcher planned and implemented the translational work alongside their high-quality science: the policy brief, the community report, or the licensed technology.

Not all science should be judged this way. Discovery research, curiosity-driven research that is often decades ahead of its application, is the foundation of scientific progress, and forcing it into near-term translation would be a mistake. Such research should be explicitly exempt, not scored as weak.

Recognizing these challenges, and drawing on my personal experiences (as an NIH-funded researcher and associate dean of research) along with the empirical literature, I provide four recommendations to shift researcher behavior toward prioritizing the societal impact of research. Alongside these recommendations, I provide specifics of their implementation, candidate indicators, and the conditions under which they would work across career stages, disciplines and institutions with varying impact capacity.

Recommendations

1. Build on NSF’s "broader impacts" criterion

In my experience, NSF's emphasis on societal impact (i.e., broader impacts criterion), and the way reviewers weigh it, pushes research markedly closer to impact than the process adopted by

NIH [5]. Adapting a proven, domestic model rather than inventing one from scratch is the most immediately achievable step.

1.1 *Implementation.* Integrate a scored translation and public benefit element into the existing review criteria for applicable activity codes, with an exemption for discovery research. It is critical to provide reviewer guidance and worked examples before the first receipt date; inconsistent reviewer interpretation, not the criterion itself, has been the main failure of impact criteria elsewhere [6].

1.2 *Indicators.* At the application stage, indicators could include specificity of the named intended users, evidence that the resulting outputs are valuable to said users, and personnel time and funds budgeted for translation.

2. Learn from international frameworks and from their failures

The UK's REF, Horizon Europe, and Canada's knowledge-translation framework offer hard-won lessons in implementing impact frameworks, including the administrative burden and gaming of the exercise [4,7,8] and the risk of rewarding polished storytelling over real change [9,10]. Australia is a case in point: its national Engagement and Impact Assessment ran once, in 2018, and was discontinued in 2023 with no replacement as yet identified. NIH should study the implementation of existing and retired impact frameworks, and talk directly with their sponsors, before importing them.

3. Build impact literacy and capacity across researchers, institutions and the wider research sector

Impact does not happen by chance. It requires the knowledge, skills, systems and culture to anticipate, design for, support and document change from the outset [11]. Established frameworks and practical resources already exist to strengthen individual and institutional impact literacy [12,13]. Requiring translational planning without building this capacity would undermine the objective.

3.1 *Implementation.* Incorporate activities to build impact literacy into institutional and career development training awards, fund administrative supplements for translational planning and implementation, and provide shared or regional expertise for institutions with limited translation infrastructure.

3.2 *Indicators.* Trainee completion of translation or engagement training, presence of institutional support structures, and translational output of trained cohorts relative to comparison groups.

4. Require investigators to report translational activities annually, and make those reports public

Today there are few expectations for what to include in progress reports, so meaningful work goes unrecorded. NIH should instruct researchers to document the knowledge- and technology-transfer practices (open science, co-creation) and products (op-eds, policy briefs) that move science beyond academia. These are not achievements to save for the final progress report; they are real outcomes along the pathway to impact. When shared openly, the research community can study them and build the science of impact alongside its practice.

4.1 *Implementation.* NIH should collect this information through a structured module in the Research Performance Progress Report using a controlled vocabulary rather than narrative, with persistent identifiers where they exist and a brief free-text field. The

resulting data should be released as a public, machine-readable resource linked to RePORTER, as NIH already does for publication linkages.

4.2 *Indicators of activity.* The module should capture non-journal outputs with persistent identifiers, including policy briefs, community reports, curricula, decision aids, protocols, datasets, software, invention disclosures, and licenses. It should also record engagement roles, such as practice or policy advisory bodies, co-authorship with non-academic partners, and formal comment or testimony to public bodies.

4.3 *Indicators of early uptake*. Early uptake can be tracked using existing data sources. These include citation of award-linked publications in policy documents and clinical practice guidelines, documented reuse of deposited data and software, and downstream licensing.

4.4 *Portfolio-level indicators.* NIH can compute a set of indicators from its own records including the share of awards reporting at least one translational output, time from award to first non-journal output, share of award-linked publications cited in guidelines or policy documents, and data reuse rates by repository.

These are indicators of practice and early uptake, not of health, economic, or social outcomes. They should be reported as a profile rather than combined into a composite score, and NIH should state that they are unsuitable for ranking individual investigators.

Closing

Changes like these would ripple well beyond NIH, reshaping how universities evaluate and promote their faculty and, over time, helping to build an impact culture. Two strategies could improve NIH's impact strategy. First, treat the initial indicator set as provisional, evaluate the protocol and its outcomes, and disseminate the findings. Second, release the structured reporting data publicly.

References

1. Bayley, J. E., & Phipps, D. (2019). Building the concept of research impact literacy. *Evidence & Policy*, 15(4), 597-606. https://doi.org/10.1332/174426417X15034894876108
2. Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology*, 64(2), 217-233. https://doi.org/10.1002/asi.22803
3. Penfield, T., Baker, M. J., Scoble, R., & Wykes, M. C. (2014). Assessment, evaluations, and definitions of research impact: A review. *Research Evaluation*, 23(1), 21-32. https://doi.org/10.1093/reseval/rvt021
4. Wilsdon, J., et al. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. HEFCE. https://www.ukri.org/wp-content/uploads/2021/12/RE-151221-TheMetricTideFullReport2015.pdf
5. Holbrook, J. B., & Frodeman, R. (2011). Peer review and the ex ante assessment of societal impacts. *Research Evaluation*, 20(3), 239-246. https://doi.org/10.3152/095820211X12941371876788
6. Derrick, G. (2018). *The Evaluators' Eye: Impact Assessment and Academic Peer Review*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-63627-6
7. Stern, N. (2016). *Building on Success and Learning from Experience: An Independent Review of the Research Excellence Framework*. Department for Business, Energy & Industrial Strategy.

https://assets.publishing.service.gov.uk/media/5a803df4e5274a2e8ab4f03d/ind-16-9-ref-stern-review.pdf

8. Martin, B. R. (2011). The Research Excellence Framework and the "impact agenda": are we creating a Frankenstein monster? *Research Evaluation*, 20(3), 247-254. https://doi.org/10.3152/095820211X13118583635693
9. Watermeyer, R. (2016). Impact in the REF: issues and obstacles. *Studies in Higher Education*, 41(2), 199-214. https://doi.org/10.1080/03075079.2014.915303
10. Watermeyer, R. (2014). Issues in the articulation of "impact": the responses of UK academics to "impact" as a new measure of research assessment. *Studies in Higher Education*, 39(2), 359-377. https://doi.org/10.1080/03075079.2012.709490
11. Bayley, J., & Phipps, D. (2019). Extending the concept of research impact literacy: levels of literacy, institutional role and ethical considerations. *Emerald Open Research*, 1:14. https://doi.org/10.12688/emeraldopenres.13140.1
12. Bayley, J., & Phipps, D. *Impact Literacy Workbook*. Emerald Publishing. https://www.emeraldgrouppublishing.com/sites/default/files/2020-06/Impact%20Literacy%20Workbook%20Final.pdf
13. Research Impact Canada. *Institutional Healthcheck Workbook*. https://researchimpact.ca/resources/institutional-healthcheck-workbook/

161. Michael Frumkin

I am responding: On behalf of myself

Name: Michael Frumkin

Name of Organization:

Type of Organization: Not Applicable

Role: Other

Role - Other: Retired head of the Google Accelerated Science lab (Corporate AI/bio research)

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: Instead of asking people to create datasets we need to ask them to create reference implementations of the experimental systems that can produce such data. We need incentives for people to build, run, and help create instances of these systems. With AI we need to exit the era of "datasets" and usher in the era of "reference experimental implementations" that can create multiple related but different datasets repeatedly and reproducibly. In building AI models you need experimental infrastructure to both provide training data, but even more importantly, to provide ground truth verification data. In most cases you need to train and differentiate the performance of dozens if not hundreds of models with different parameters to figure out what works best. Even with data holdouts, doing this with a single data set becomes extremely limiting.

We need to create incentives for people to create systems that generate data, not standalone datasets. These systems need to be fully open, replicable, and made widely available.

2. Data, Software, and Model Sharing: This is similar and tied in to the reproducibility and rigor problem. Again, we need to incentivize the production of systems that produce experimental data, not datasets. In the age of AI models the key to success is an iterative process between building and testing a model, and producing new experimental data specifically to eliminate noise in the model (as opposed to illuminate a specific hypothesis). By doing this repeatedly the model becomes truly predictive. This cannot be accomplished with a dataset because it is impossible to predict ahead of time exactly what data will be needed for the model to converge. By incentivizing the creation of datasets we are actually holding back the effectiveness of AI in biology and chemistry. We need to incentivize the production of reference implementations of experimental infrastructure that produce datasets, not a particular dataset.

3. Training and Mentorship: N/A

4. Collaboration: Collaboration is crucial for making progress. Especially with AI techniques revolutionizing scientific capabilities. For the first time ever, something other than a human brain can make sense of complex patterns in scientific data, and do it in ways that are different and frequently complimentary to human abilities. I spent 15 years creating and running such collaborations. The main problem is that it takes significant time for very smart people from different disciplines to truly

understand each other. Collaborations without a long term commitment are fundamentally limited. Biologists naturally gravitate to having AI solve one of their current problems. Computer Scientists naturally gravitate to the biology problem that best matches the abilities of their current model or algorithm. In almost all cases both of these approaches are suboptimal. The right thing to do is to ask "what scientific question could we answer together that was previously not approachable". This is a difficult thought exercise, but crucial. Otherwise a lot of effort is spent on the wrong (or at least suboptimal) problems. This is an extreme case of only looking for your keys under the streetlight. I would argue that this is currently the worst problem we have in interdisciplinary collaborations. The main way to fix this is to tailor incentives for longer collaborations and bigger impact. This will both allow both side to understand each other to ask the right big questions, and discourage short term but lower impact work.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: The right approach it to create longer term incentives with a higher bar. Often PIs answer the biggest question they can fit into the timeline of a grant and not the question they truly think is important. Incentivizing PIs to answer questions that are not actually very important is a waste. It makes more sense to create bigger and longer term grants but hold PIs to a higher standard for productivity. Publication should be encouraged but should not be used as a real measure of impact. We need to incentivize the building of reusable experimental systems and the production of reusable models. This is harder than publishing a paper, but much more valuable. If other scientists actually use the actual models or experimental systems, that is a much much higher accomplishment than referencing a paper and should be rewarded as such.

7. Public Impact: N/A

8. Other Areas: N/A

162. Nigel Paneth

I am responding: On behalf of myself

Name: Nigel Paneth

Name of Organization: Michigan State University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/13/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: I will only comment on this category, which I believe is the single most important and yet the single most ignored issue in grant evaluation. And in this category I single out, to use your terminology "contributions to health outcomes". Useful as they are, none of the considerations in the six categories listed above have any long-lasting significance for NIH unless they lead to measurable improvement in the health of the population. Health is the purpose of the NIH, just as science is the purpose of the NSF.
So how do we assess this impact? Simply by asking of studies, how was the health of the population affected by the findings?
I am a pediatrician and it is sobering to consider that some of the most important recent advances in child health did not emerge from NIH research. The finding that neural tube defects can be prevented by periconceptional folic acid is owed to trials conducted in the UK and in Hungary. Notably, the NIH hesitated to undertake a trial of folic acid supplementation in the 1980s preferring to continue with observational research on the hypothesis. Supine sleeping in infants, which saves some 4,000 US lives annually, emerged from studies done in Holland , New Zealand and Australia.
The development of components of newborn intensive care and the demonstration of their impact on child health, one the other hand, did emerge from NIH research. What I think we are missing is a systematic inventory of major advances in health and an attempt to trace back their origins in research. If we did that, we might better understand how research eventuates in measurable improvements in population health parameters. And, in parallel with the above, we need to ask of studies what impact they had on population health after enough time has passed to make a judgment, The lack of immediacy in the result should not preclude examining outcomes a decade after a study has been completed. It might be interesting and useful to ask investigators to provide an assessment of the improvements in

health that have resulted from their work - a question that could be asked 2, 5 and ten years after study completion. I would be happy to provide such an assessment of the several NIH studies I have undertaken as a PI since my first R01 that began in 1978.

8. Other Areas: N/A

163. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/14/2026

1. Rigor and Reproducibility: Incentives: failures to replicate need to be publishable. Built-in replications need to be rewarded in both the publishing and granting process. For example, it many cases it is trivial to build a replication into a new study and find (or not find) the prior results along with generating new results. Such a process should be rewarded; however, often it is penalized. Journals reject manuscripts that include such replications along with new results, citing lack of novelty, which actively de-incentivizes researchers from taking this approach. This approach should be easy and low cost. Infrastructure: researchers need training in null hypothesis testing, Bayesian inference, and overall replication best-practices. Changing every variable in a study and failing to replicate prevents interpretation of what factor may have driven the original effect. Combining both: registered-reports for replication studies is a good idea on paper; however, in practice it is insanely cumbersome. The months that it can take a basic replication to pass through multiple rounds of peer review is unacceptable and untenable.

2. Data, Software, and Model Sharing: Then why won't the NIH fund the Center for Open Science (COS)? This open-access platform was used for a decade by many researchers and due to lack of funding, will no longer host data in 2027. This is a major loss to the research community. To incentivize sharing, it must be made easy and free. The level of burden placed on researchers is incredible: researchers do not have the time or money to share their data, even if they are amenable to the practice. It is also not just the time it takes to organize the data to comply with policies. It is also the time it takes to format the "data/code availability" sections and tables of manuscripts since the rules and expectations are different for nearly every academic journal.

3. Training and Mentorship: Indicators - how many F31 recipients proceed to a post-doc position, other academic position, industry position, or leave science relative to those who do not? Same for F32, but transition to faculty positions. If the goals of the training is career potential and preparedness, measure the careers obtained as the outcome measure.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Fund more basic research. Do not evaluate basic science proposals on their translational/clinical potential, but instead on the knowledge to be gained from the work itself. No one, both basic and applied science researchers, can operate to their maximum potential

in this climate of uncertainty. Ensure support for basic science by removing threats to cancel grants at will, without cause, and without recourse to appeal.

7. Public Impact: N/A

8. Other Areas: N/A

164. American Psychological Association

I am responding: On behalf of an organization

Name: Angela Sharpe for Adrienne Stith, PhD

Name of Organization: American Psychological Association

Type of Organization: Non-Profit Research Organization

Role: Organizational Official

Date Submitted: 8/14/2026

1. Rigor and Reproducibility: APA supports NIH's efforts to recognize rigor, reproducibility, replicability, robustness, and credibility as important dimensions of scientific impact. Psychological science has contributed significantly to the study of research credibility and has demonstrated that confidence in findings can be strengthened through complementary approaches such as transparent reporting, independent replication, reproducibility checks, robustness analyses, preregistration, and structured expert assessment.

Drawing on the literature, APA recommends that NIH distinguish between reproducibility, robustness, replicability, and trustworthiness when developing assessment frameworks. This distinction is important for implementation. Treating these concepts as interchangeable may create confusion and unintended incentives. In addition, different research questions may require different forms of assessment. Assessment systems should encourage investigators and reviewers to identify the most appropriate methods for evaluating confidence in findings and reward contributions that make research more transparent, interpretable, and verifiable.

Nosek, B. A., et al., propose a concise glossary to improve precision and communication about research quality and evidence assessment. The glossary is directly relevant because assessment systems can create unclear incentives when key concepts are treated as interchangeable. Accordingly, APA recommends that NIH use precise terminology in application instructions, review criteria, reviewer guidance, progress reports, and renewal decisions so that investigators and reviewers understand whether they are assessing reproducibility of analyses, robustness of conclusions across defensible analytic choices, replicability with independent data, or broader credibility and trustworthiness.

APA recommends that NIH invest in the science of credibility assessment itself. Nosek and colleagues' work on replicability, robustness, and reproducibility demonstrates that scientific impact depends not only on producing new findings, but also determining whether findings are reliable, interpretable, reusable, and capable of informing cumulative knowledge. SCORE shows that scalable indicators of replicability, robustness, reproducibility, transparency, and other credibility dimensions must be validated before they are used as incentives. NIH should support research, infrastructure, and pilot programs that test whether proposed indicators are reliable, explainable, field-sensitive, and useful to reviewers and decision-makers. Vazire's work reinforces this point by showing that quality uncertainty erodes trust when consumers of science lack the information needed to evaluate claims, and that the

credibility revolution may reduce apparent individual productivity while increasing genuine scientific progress. Policies for measuring and rewarding impact should be evaluated empirically, implemented flexibly, and designed to encourage ongoing scrutiny of whether new metrics actually improve science.

APA recommends that NIH incentivize practices that increase confidence in findings, including well-powered designs, transparent analytic decisions, preregistration where appropriate, clear reporting of methods and limitations, independent replication, and evaluation of whether findings hold across analytic choices, populations, measures, and contexts. These practices are especially important in fields that inform clinical care, public health, education, and policy, where premature or overstated claims have significant downstream consequences, despite the frequent need for rapid decision-making and response.

At the same time, APA encourages NIH to implement rigor and reproducibility indicators in ways that are appropriate. Not all research questions are suited to the same forms of replication or preregistration, and exploratory, qualitative, community-engaged, longitudinal, and observational research may require different standards of transparency and robustness. NIH should support discipline-specific guidance and avoid rigid requirements that could unintentionally discourage innovation, exploratory discovery, and more research with hard-to-reach populations.

APA recommends that NIH incorporate these indicators into grant review, progress reporting, and career advancement mechanisms through narrative contribution statements and portfolio-based assessments that allow investigators to show the significance of their most meaningful work. In psychology and behavioral science, especially valuable contributions include demonstrating that findings replicate across diverse populations and settings, developing culturally valid measures, creating reusable research resources, advancing methods for studying behavior and mental health, and generating evidence that informs clinical practice, prevention, education, workplace well-being, and public policy.

The association further recommends that NIH also recognize contributions that strengthen the research enterprise itself, such as leadership in multi-site collaborations, training the next generation of scientists, and advancing open and transparent research practices. Rewarding the practices that produce trustworthy, reproducible, and actionable knowledge would strengthen stewardship of public research investments, accelerate scientific progress, and enhance public trust in biomedical and behavioral research.

Psychology-specific research indicators include:

- Independent replications of influential findings and interventions.
- Evidence that findings are robust across populations, contexts, measures, and analytic approaches.
- Preregistration and registered reports, when appropriate to research design.
- Transparent reporting of methods, limitations, analytic decisions, and effect sizes.
- Development and use of rigorous standards for causal inference in behavioral and social science research.
- Clear differentiation among reproducibility of analyses, replicability of findings, and robustness across analytic choices, samples, measures, and contexts.
- Use of repeatability as an umbrella term while separately assessing reproducibility, robustness, and

replicability.
- Validated, explainable credibility indicators that assess research features directly rather than relying only on journal prestige, citation counts, or publication status.

2. Data, Software, and Model Sharing: APA recommends that NIH recognize open science practices as meaningful contributions to scientific impact. Sharing data, code, software, research materials, protocols, measures, models, and analytic plans can accelerate discovery, enable independent verification, reduce duplication, and increase the public return on federal research investments.

The Score initiative is one of the largest efforts ever to evaluate the credibility of research findings in social and behavioral science. The project was developed through a DARPA-funded collaboration involving the Center for Open Science, academic researchers, and technology partners. The SCORE program underscores that scientific trustworthiness does not require certainty. It depends on the research enterprise's capacity to reduce uncertainty through credible evidence, transparent critique, and ongoing assessment. Common proxies for credibility, such as peer-reviewed publication, journal prestige, and citation counts are easy to measure but weak indicators of research quality because they do not examine the research itself. According to Nosek, B. A. et al, credibility assessment should instead ask whether the findings deserve serious consideration and whether the evidence is strong enough to guide future research, resource allocation, policy, practice, or translation.

SCORE tested whether human judgements and automated methods could predict replication outcomes and provide explainable assessments. The results show modes of progress but also indicate that neither human nor machine predictions are yet accurate enough to replace replication studies. For NIH, this shows that credibility indicators require sustained investment, validation, transparency, and caution before they are used to guide incentives or decision-making.

According to Nosek, B. A., et al., replicability is only part of a broader credibility landscape that includes transparent peer evaluation, engagement with prior knowledge, bias control, measurement precision and reliability, accountability, and whether research claims are proportionate to the evidence. Emerging modes such as Lifecycle Journal, an open-science scholarly publishing platform designed to evaluate research across the entire research lifecyle, allow multiple independent review services to assess discrete features of credibility, such as code execution, data sharing quality, consistency between research plans and outcomes, reporting standards, and statistical errors. A more diversified assessment system could complement peer review and help researchers, reviewers, policymakers, journalists, and the public evaluate research quality more effectively.

3. Training and Mentorship: APA strongly supports NIH's inclusion of mentorship, training, and team science as dimensions of scientific impact. These activities contribute directly to the integrity, sustainability, and the effectiveness of the biomedical and behavioral research enterprise but are often undervalued in publication-based assessments.

NIH should recognize evidence of effective mentorship, trainee development, inclusive research environments, methodological and ethical training, collaborative leadership, shared infrastructure development, and contributions to large interdisciplinary projects. Evaluation criteria should remain flexible enough to accommodate differences in career stage, institutional context, and research settings.

Psychology-specific research indicators include:
- Development of rigorous methodological and ethical training programs.
- Contributions to building and sustaining the behavioral and social science workforce.
- Successful mentorship of trainees and early-career investigators.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Public Impact Should Include Health, Trust, Accessibility, and Community Benefit
APA strongly agrees that scientific impact should ultimately be connected to NIH's mission to generate knowledge that enhances health, lengthens life, and reduces illness and disability. Public impact should be understood broadly to improvements in health outcomes, clinical practice, prevention, education, public policy, community well-being, accessibility of scientific knowledge, and public trust in science.

For psychological and behavioral science, public impact may include validated interventions, improved mental and behavioral health services, better measurement tools, evidence-based prevention strategies, reduced stigma, improved adherence to care, enhanced learning and development, strong community resilience, and more effective health communication. NIH should recognize that translation may occur through clinical guidelines, public health programs, community partnerships, education, workforce training, policy implementation, and dissemination of accessible evidence—not only through commercialization or patents. It often requires sustained engagement, adaptation, partnership, and communication beyond publication.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/APA-Response-to-NIH-Request-for-Information-on-Measuring-and-Rewarding-Scientific-Impact-8.19.2026.pdf

Description: The complete APA response to RFI.

AMERICAN PSYCHOLOGICAL ASSOCIATION

August 14, 2026

Lyric Jorgenson, PhD
NIH Associate Director for Science Policy
NIH Office of Science Policy
6705 Rockledge Drive
Suite 630, MSC-7985
Bethesda, MD 20892-7985

Re: Request for Information on Measuring and Rewarding Scientific Impact (Notice Number: NOT-OD-26-087)

Via email: SciencePolicy@od.nih.gov

Dear Dr. Jorgenson:

The American Psychological Association (APA) appreciates the opportunity to provide input to the National Institutes of Health (NIH) regarding how scientific impact should be measured, recognized, and incentivized in modern biomedical and behavioral research via the Request for Information (RFI): Measuring and Rewarding Scientific Impact, Notice Number: NOT-OD-26-087. The APA is the nation's largest scientific and professional nonprofit organization representing the discipline and profession of psychology, as well as 190,000 members and affiliates who are clinicians, researchers, educators, consultants, and students in psychological science.

APA welcomes NIH's efforts to update how scientific impact is measured and rewarded. Research assessment should reflect the practices that support credible science, including rigorous methods, transparency, collaboration, mentorship, and public benefit. As behavioral, cognitive, developmental, mental health, and social sciences research continue to inform health and well-being, NIH has an opportunity to reward not only productivity, but quality, credibility, openness, mentorship, collaboration, and real-world value.

Academia's reliance on quantitative indicators, including publication counts, journal impact factors, and citation metrics, creates incentives that can work against research quality. Assessment systems should move from asking how much research is produced to asking how well research is conducted. APA supports NIH's interests in responsible research assessment that prioritizes research quality, rigor, transparency, credibility, and societal value over simple measures of productivity. Psychology has been a leader in developing methods that support this shift. Assessment and incentive systems should recognize contributions that strengthen the credibility and usefulness of science, including transparent reporting, data and code sharing, preregistration where appropriate, replication efforts, methodological

750 First Street, NE
Washington, DC 20002-4242
202.336.5800
202.336.6123 TDD

innovation, team science, and shared research resources. Several initiatives currently exist that aim to address this challenge.

For behavioral and social science research, quality indicators should recognize contributions that improve credibility, reproducibility, and cumulative nature of scientific knowledge, including the development of validated measures, replication studies, open research resources, methodological advances, and efforts that improve research transparency and accessibility. Psychological science has been at the forefront of developing and evaluating approaches to replication, reproducibility, research transparency, and credibility assessment, and these advances demonstrate that scientific quality can be measured in ways that are more closely aligned with NIH's mission than traditional publication-based metrics alone.[1]

Accordingly, APA recommends that NIH take a multidimensional approach to scientific impact that balances traditional indicators, such as publications and citations, with evidence of rigor, reproducibility, responsible data and software sharing, effective training and mentorship, meaningful collaboration, translation, and contributions to public trust. No single metric can capture the full value of scientific work. NIH should therefore avoid replacing one narrow set of incentives with another and instead develop flexible, field sensitive approaches that recognize diverse forms of contributions across career stages, institution types, research methods, and communities served.

Systematizing Confidence in Open Research and Evidence (SCORE) Credibility Assessment

APA recommends that NIH recognize open science practices as meaningful contributions to scientific impact. Sharing data, code, software, research materials, protocols, measures, models, and analytic plans can accelerate discovery, enable independent verification, reduce duplication, and increase the public return on federal research investments.[2]

The Score initiative is one of the largest efforts ever to evaluate the credibility of research findings in social and behavioral science. The project was developed through a DARPA-funded collaboration involving the Center for Open Science, academic researchers, and technology partners. The SCORE program underscores that scientific trustworthiness does not require certainty. It depends on the research enterprise's capacity to reduce uncertainty through credible evidence, transparent critique, and ongoing assessment. Common proxies for credibility, such as peer-reviewed publication, journal prestige, and citation counts are easy to measure but weak indicators of research quality because they do not examine the research itself. According to Nosek, B. A. et al, credibility assessment should instead ask whether the findings deserve serious consideration and whether the evidence is strong enough to guide future research, resource allocation, policy, practice, or translation.[3]

---

[1] Gärtner A, Leising D, Schönbrodt FD (2024) Towards responsible research assessment: How to reward research quality. PLoS Biol 22(2): e3002553. https://doi.org/10.1371/journal.pbio.3002553

[2] Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. Science, 348(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

[3] Nosek, B. A., et al. (2026). *Reimagining and diversifying assessment of the credibility of research findings* (abjur_v1). MetaArXiv. https://osf.io/preprints/metaarxiv/abjur_v1

APA.ORG

750 First Street, NE
Washington, DC 20002-4242

202.336.5500
202.336.6123 TDD

SCORE tested whether human judgements and automated methods could predict replication outcomes and provide explainable assessments. The results show modes of progress but also indicate that neither human nor machine predictions are yet accurate enough to replace replication studies. For NIH, this shows that credibility indicators require sustained investment, validation, transparency, and caution before they are used to guide incentives or decision-making.

According to Nosek, B. A., et al., replicability is only part of a broader credibility landscape that includes transparent peer evaluation, engagement with prior knowledge, bias control, measurement precision and reliability, accountability, and whether research claims are proportionate to the evidence. Emerging modes such as Lifecycle Journal, an open-science scholarly publishing platform designed to evaluate research across the entire research lifecyle, allow multiple independent review services to assess discrete features of credibility, such as code execution, data sharing quality, consistency between research plans and outcomes, reporting standards, and statistical errors. A more diversified assessment system could complement peer review and help researchers, reviewers, policymakers, journalists, and the public evaluate research quality more effectively.[4]

Core Indicators of Scientific Impact

APA supports NIH's efforts to recognize rigor, reproducibility, replicability, robustness, and credibility as important dimensions of scientific impact. Psychological science has contributed significantly to the study of research credibility and has demonstrated that confidence in findings can be strengthened through complementary approaches such as transparent reporting, independent replication, reproducibility checks, robustness analyses, preregistration, and structured expert assessment.

Drawing on the literature, APA recommends that NIH distinguish between reproducibility, robustness, replicability, and trustworthiness when developing assessment frameworks. This distinction is important for implementation.[5] Treating these concepts as interchangeable may create confusion and unintended incentives. In addition, different research questions may require different forms of assessment. Assessment systems should encourage investigators and reviewers to identify the most appropriate methods for evaluating confidence in findings and reward contributions that make research more transparent, interpretable, and verifiable.[6]

Nosek, B. A., et al., propose a concise glossary to improve precision and communication about research quality and evidence assessment.[7] The glossary is directly relevant because assessment systems can create unclear incentives when key concepts are treated as interchangeable. Accordingly, APA recommends that NIH use precise terminology in application instructions, review criteria, reviewer

[4] Nosek, B. A., et al. (2026). *Reimagining and diversifying assessment of the credibility of research findings* (abjur_v1). MetaArXiv. https://osf.io/preprints/metaarxiv/abjur_v1

[5] Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. Annual Review of Psychology, 73, 719-748. https://doi.org/10.1146/annurev-psych-020821-114157

[6] Nosek, B. A., et al. (2025). A brief glossary of terms about repeatability: Replicability, robustness, and reproducibility (mqfp4_v1). MetaArXiv. https://doi.org/10.31222/osf.io/mqfp4_v1

[7] Nosek et al. (2025). A Brief Glossary of Terms about Repeatability: Replicability, Robustness, and Reproducibility (No. mqfp4_v1). *MetaArXiv*. https://doi.org/10.31222/osf.io/mqfp4_v1

APA.ORG

750 First Street, NE
Washington, DC 20002-4242

202.336.5500
202.336.6123 TDD

guidance, progress reports, and renewal decisions so that investigators and reviewers understand whether they are assessing reproducibility of analyses, robustness of conclusions across defensible analytic choices, replicability with independent data, or broader credibility and trustworthiness.

APA recommends that NIH invest in the science of credibility assessment itself. Nosek and colleagues' work on replicability, robustness, and reproducibility demonstrates that scientific impact depends not only on producing new findings, but also determining whether findings are reliable, interpretable, reusable, and capable of informing cumulative knowledge. SCORE shows that scalable indicators of replicability, robustness, reproducibility, transparency, and other credibility dimensions must be validated before they are used as incentives. NIH should support research, infrastructure, and pilot programs that test whether proposed indicators are reliable, explainable, field-sensitive, and useful to reviewers and decision-makers. Vazire's work reinforces this point by showing that quality uncertainty erodes trust when consumers of science lack the information needed to evaluate claims, and that the credibility revolution may reduce apparent individual productivity while increasing genuine scientific progress.[8] [9] Policies for measuring and rewarding impact should be evaluated empirically, implemented flexibly, and designed to encourage ongoing scrutiny of whether new metrics actually improve science.

APA recommends that NIH incentivize practices that increase confidence in findings, including well-powered designs, transparent analytic decisions, preregistration where appropriate, clear reporting of methods and limitations, independent replication, and evaluation of whether findings hold across analytic choices, populations, measures, and contexts. These practices are especially important in fields that inform clinical care, public health, education, and policy, where premature or overstated claims have significant downstream consequences, despite the frequent need for rapid decision-making and response.

At the same time, APA encourages NIH to implement rigor and reproducibility indicators in ways that are appropriate. Not all research questions are suited to the same forms of replication or preregistration, and exploratory, qualitative, community-engaged, longitudinal, and observational research may require different standards of transparency and robustness. NIH should support discipline-specific guidance and avoid rigid requirements that could unintentionally discourage innovation, exploratory discovery, and more research with hard-to-reach populations.

APA recommends that NIH incorporate these indicators into grant review, progress reporting, and career advancement mechanisms through narrative contribution statements and portfolio-based assessments that allow investigators to show the significance of their most meaningful work. In psychology and behavioral science, especially valuable contributions include demonstrating that findings replicate across diverse populations and settings, developing culturally valid measures, creating reusable research resources, advancing methods for studying behavior and mental health, and generating evidence that informs clinical practice, prevention, education, workplace well-being, and public policy.

[8] Vazire, S. (2017). Quality uncertainty erodes trust in science. Collabra: Psychology, 3(1), Article 1. https://doi.org/10.1525/collabra.74

[9] Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. Perspectives on Psychological Science, 13(4), 411-417. https://doi.org/10.1177/1745691617751884

APA.ORG

750 First Street, NE
Washington, DC 20002-4242

202.336.5500
202.336.6123 TDD

The association further recommends that NIH also recognize contributions that strengthen the research enterprise itself, such as leadership in multi-site collaborations, training the next generation of scientists, and advancing open and transparent research practices. Rewarding the practices that produce trustworthy, reproducible, and actionable knowledge would strengthen stewardship of public research investments, accelerate scientific progress, and enhance public trust in biomedical and behavioral research.

Psychology-specific research indicators include:

- Independent replications of influential findings and interventions.
- Evidence that findings are robust across populations, contexts, measures, and analytic approaches.
- Preregistration and registered reports, when appropriate to research design.
- Transparent reporting of methods, limitations, analytic decisions, and effect sizes.
- Development and use of rigorous standards for causal inference in behavioral and social science research.
- Clear differentiation among reproducibility of analyses, replicability of findings, and robustness across analytic choices, samples, measures, and contexts.
- Use of repeatability as an umbrella term while separately assessing reproducibility, robustness, and replicability.
- Validated, explainable credibility indicators that assess research features directly rather than relying only on journal prestige, citation counts, or publication status.

Reward Open, Transparent, and Reusable Research Outputs

APA supports NIH's interest in recognizing open and reusable research outputs as meaningful scientific contributions. Sharing data, code, software, measures, protocols, materials, and analytic workflows increase the value of public research investments by enabling verification, reuse, collaboration, and cumulative knowledge building.

NIH should evaluate both the availability and usability of these resources, including documentation quality, accessibility, responsible governance, and evidence of reuse by others. At the same time, sharing requirements should remain flexible enough to accommodate privacy concerns, Tribal sovereignty, community governance, consent limitations, and other ethical obligations.

Consistent with the Transparency and Openness Promotion (TOP) framework, a set of guidelines developed by the Center for Open Science and a coalition of journal editors, researchers, funders, and scientific organizations to improve transparency, credibility, and reproducibility of research, NIH should adopted a tiered approach that encourages transparency while recognizing the differences across research designs and disciplines. Tiered standards would allow NIH to recognize meaningful progress toward transparency across diverse research designs, including experimental, observational, qualitative, longitudinal, clinical, and community-engaged research. This approach would also help ensure that data, code materials, protocols, and analysis plans are shared when appropriate, while preserving necessary protections for privacy, consent, Tribal sovereignty, community governance, and sensitive human subjects data.

NIH should also treat open research outputs as substantive scholarly contributions. Investigators should be able to document the quality, accessibility, reuse, and scientific value of shared datasets, measures,

APA.ORG

750 First Street, NE
Washington, DC 20002-4242

202.336.5500
202.336.6123 TDD

protocols, software, code, analytic workflows, and other research materials in bio-sketches, progress reports, grant renewals, and career-stage assessments. Recognizing these outputs would align individual incentives with the public value of research that is verifiable, reusable, and cumulative.

Psychology-specific research indicators include:

- Public sharing of validated measures, intervention manuals, study protocols, code, and datasets where appropriate.
- Development of research infrastructure, repositories, and community resources that benefit the field.
- Demonstrated reuse of shared resources by other investigators.
- Use of tiered transparency standards that recognize appropriate sharing of data, code, materials, protocols, analysis plans, and documentation across different research designs.

Value Quality Over Quantity

As noted above, APA recommends that NIH reduce reliance on publication counts, journal prestige, and citation-based indicators when evaluating scientific impact. Although such measures can provide useful contextual information, they should not serve as the only proxy for scientific value.

Assessment systems should instead emphasize the quality, credibility, transparency, rigor, and cumulative contribution of scientific work. Activities such as replication, correction of errors, development of methods and measures, open science practices, development of community resources, and meaningful translation often create substantial scientific value when they do not generate large numbers of publications.

Narrative contributions statements and portfolio-based assessment approaches may provide a more accurate picture of scientific impact by allowing investigators to describe their most significant contributions rather than simply listing outputs.[10]

APA recommends that NIH also protect creativity and foundational exploration while strengthening rigor. Assessment frameworks should recognize exploratory and high-risk research when investigators are transparent about uncertainty, clearly distinguish exploratory from confirmatory analysis, and describe how early-stage findings will be tested, replicated, or refined. This would help avoid a narrow compliance culture in which researchers pursue only low-risk, easily preregistered questions, while still advancing NIH's goals of rigor, reproducibility, and public trust.[11]

NIH should reward research practices that make scientific claims easier to evaluate. Impact indicators should recognize transparent disclosure of study design, measures, sampling decisions, exclusions, analytic flexibility, limitations, and departures from preregistered plans when applicable. These practices would help reviewers, other scientists, policymakers, and the public distinguish well-supported claims from more tentative findings and would strengthen trust in NIH-supported science.[12]

---

[10] Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. Perspectives on Psychological Science, 13(4), 411-417. https://doi.org/10.1177/1745691617751884

[11] Vazire, S. (2017). Quality uncertainty erodes trust in science. Collabra: Psychology, 3(1), Article 1. https://doi.org/10.1525/collabra.74

[12] Vazire, S. (2017). Quality uncertainty erodes trust in science. Collabra: Psychology, 3(1), Article 1. https://doi.org/10.1525/collabra.74

APA.ORG

750 First Street, NE
Washington, DC 20002-4242

202.336.5500
202.336.6123 TDD

Recognize Mentorship, Training, and Team Science as Scientific Contributions
APA strongly supports NIH's inclusion of mentorship, training, and team science as dimensions of scientific impact. These activities contribute directly to the integrity, sustainability, and the effectiveness of the biomedical and behavioral research enterprise but are often undervalued in publication-based assessments.

NIH should recognize evidence of effective mentorship, trainee development, inclusive research environments, methodological and ethical training, collaborative leadership, shared infrastructure development, and contributions to large interdisciplinary projects. Evaluation criteria should remain flexible enough to accommodate differences in career stage, institutional context, and research settings.[13]

Psychology-specific research indicators include:

- Development of rigorous methodological and ethical training programs.
- Contributions to building and sustaining the behavioral and social science workforce.
- Successful mentorship of trainees and early-career investigators.

Public Impact Should Include Health, Trust, Accessibility, and Community Benefit
APA strongly agrees that scientific impact should ultimately be connected to NIH's mission to generate knowledge that enhances health, lengthens life, and reduces illness and disability. Public impact should be understood broadly to improvements in health outcomes, clinical practice, prevention, education, public policy, community well-being, accessibility of scientific knowledge, and public trust in science.

For psychological and behavioral science, public impact may include validated interventions, improved mental and behavioral health services, better measurement tools, evidence-based prevention strategies, reduced stigma, improved adherence to care, enhanced learning and development, strong community resilience, and more effective health communication. NIH should recognize that translation may occur through clinical guidelines, public health programs, community partnerships, education, workforce training, policy implementation, and dissemination of accessible evidence—not only through commercialization or patents. It often requires sustained engagement, adaptation, partnership, and communication beyond publication.

Implementation Should Minimize Unintended Consequences
APA encourages NIH to implement new impact measures cautiously and evaluate them regularly for unintended consequences. Indicators should be piloted across disciplines, institution types, and career stages before broad adoption. NIH should monitor whether proposed measures improve the assessment of rigor, transparency, and scientific value without creating excessive administrative burden, discouraging innovation, or disadvantaging researchers working in resource-constrained settings or with sensitive data.

[13] Gärtner, A., Leising, D., & Schönbrodt, F. D. (2024). Towards responsible research assessment: How to reward research quality. *PLoS Biology, 22*(2), e3002553. https://doi.org/10.1371/journal.pbio.3002553

APA.ORG
750 First Street, NE
Washington, DC 20002-4242
202.336.5500
202.336.6123 TDD

Successful implementation will require reviewer guidance, stakeholder feedback, and investments in the infrastructure needed to support replication, data stewardship, software maintenance, training, and methodological support.[14]

Recommended Indicators and Incentives
APA urges NIH to elevate rigor, transparency, collaboration, and public impact by operationalizing new incentives through application instructions, review criteria, reviewer guidance, progress reports, renewals, bio-sketches, final reports, and career-stage assessments.

Responsible Research Assessment
- Reduce reliance on publication counts and journal-level metrics.
- Use narrative contributions statements and portfolio-based evaluations emphasizing rigor, transparency, and cumulative value.

Rigor and Credibility
- Reward reproducibility, robustness analyses, replication, transparent reporting, appropriate preregistration, adequate methods and analyses,
- and evidence of generalizability.

Open and Reusable Research
- Recognize sharing of data, code, measures, software, protocols, and research materials, with appropriate privacy and governance protections.

Mentorship and Team Science
- Recognize mentorship quality, trainee and workforce development, methodological and ethical training, inclusive research environments, collaborative leadership, shared infrastructure, and interdisciplinary teamwork.

Innovation and Translation
- Support credible high-risk research and recognize contributions to policy, practice, implementation, and public health impact.

Evaluation and Continuous Improvement
- Pilot, validate, and periodically reassess indicators to ensure they are reliable, fair, field sensitive, and resistant to unintended consequences.

APA commends NIH for seeking input on how to better measure and reward scientific impact. The biomedical and behavioral research enterprise depends on incentives that support credible, transparent, collaborative, and socially beneficial science. By broadening assessment beyond narrow productivity measures and recognizing rigor, reproducibility, openness, mentorship, collaboration, and public benefit, NIH can strengthen the reliability of scientific findings, improve stewardship of public research investments, and reinforce public trust in science. Should you have any questions, please contact Angela L. Sharpe, MG, Senior Director, Congressional & Federal Relations at asharpe@apa.org.

[14] Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

APA.ORG
750 First Street, NE
Washington, DC 20002-4242
202.336.5500
202.336.6123 TDD

Sincerely,



Adrienne Stith, PhD
Executive Lead Psychologist, Science
American Psychological Association

References

Abatayo, A. L., et al. (2026). Assessments of credibility in the social and behavioural sciences (7u58q_v1). MetaArXiv. https://osf.io/preprints/metaarxiv/7u58q_v1

Aczel, B., Szaszi, B., Clelland, H. T., et al. (2026). Investigating the analytical robustness of the social and behavioural sciences. *Nature, 652*(8108), 135-142. https://doi.org/10.1038/s41586-025-09844-9

Alipourfard, N., et al. (2021). *Systematizing Confidence in Open Research and Evidence (SCORE)* (46mnb_v1). SocArXiv. https://doi.org/10.31235/osf.io/46mnb

Miske, O., Abatayo, A. L., Daley, M., et al. (2026). Investigating the reproducibility of the social and behavioural sciences. *Nature, 652*(8108), 126-134. https://doi.org/10.1038/s41586-026-10203-5

National Academies of Sciences, Engineering, and Medicine. 2025. *The Science and Practice of Team Science*. Washington, DC: The National Academies Press DOI: https://doi.org/10.17226/29043.

Nosek, B. A., et al. (2026). *Reimagining and diversifying assessment of the credibility of research findings* (abjur_v1). MetaArXiv. https://osf.io/preprints/metaarxiv/abjur_v1

Tyner, A. H., Abatayo, A. L., Daley, M., et al. (2026). Investigating the replicability of the social and behavioural sciences. *Nature, 652*(8108), 143-150. https://doi.org/10.1038/s41586-025-10078-y

APA.ORG

750 First Street, NE
Washington, DC 20002-4242

202.336.5500
202.336.6123 TDD

165. American Dental Association

I am responding: On behalf of an organization

Name: David Linn

Name of Organization: American Dental Association

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Director, Federal Affairs

Date Submitted: 8/14/2026

1. Rigor and Reproducibility: NIH should establish dedicated incentives and funding opportunities for replication and reproducibility, including access to small-grant mechanisms such as R03 awards, supplements, or other limited-duration mechanisms where appropriate, while preserving robust support for R01-scale projects that feature rigorous replication and reproducibility.
Replication studies, pre-registration, transparent reporting of methods and analyses, and publication of well-designed negative or null findings should be recognized as meaningful scientific contributions. NIH could incorporate these activities into appropriate application and progress-report materials and provide reviewers and program staff with guidance for assessing them separately from the production of novel findings.
Recognition should be based on the quality and informativeness of the work rather than the simple presence of a replication label. Funding such rigorous replication projects through a range of award sizes would allow early-career investigators, clinician-scientists, and researchers at dental schools and smaller institutions to participate without requiring the resources or preliminary development associated with a large research project.
A diversified funding structure would increase the efficiency of NIH investments and broaden participation without creating unnecessary competition between replication and discovery. NIH should also ensure that replication incentives do not unintentionally reduce support for exploratory or hypothesis-generating research. Review criteria should distinguish causal conclusions from associational findings based on study design, methodology, and analytic limitations, while recognizing the value of research that appropriately identifies associations warranting further study.

2. Data, Software, and Model Sharing: NIH should recognize persistent identifiers, including dataset and software digital object identifiers and Research Resource Identifiers, as verifiable evidence of contributions to data, software, model, and research-resource sharing, while assessing scientific impact based on the quality, stewardship, adoption, and reuse of those resources rather than on identifiers or usage counts alone.
Investigators should be able to document these contributions in appropriate application, review and progress-report materials alongside publications and other research outputs. NIH should assess not only whether a dataset, model, software tool, or research resource has been shared, but also whether it is adequately documented, accessible, reusable, appropriately curated, and capable of supporting validation, replication, or additional research. Relevant indicators of impact could include independent

scientific reuse, citations or other documented attribution, incorporation into external analyses or resources, adoption by the research community, and evidence that a shared resource has enabled subsequent scientific discoveries.

The Human Oral Microbiome Database (HOMD), developed and sustained at the ADA Forsyth Institute with NIDCR support, provides an example of the long-term scientific value that can result from investing in openly accessible and expertly curated research resources. Since its public launch, HOMD has provided the scientific community with curated taxonomic, genomic, phylogenetic, and bibliographic information and associated bioinformatics resources for microorganisms of the human oral and aerodigestive tract. Its continuing expansion and use across the scientific community illustrate how an NIH-supported research program can produce both new scientific knowledge and a durable shared resource that enables research well beyond the originating institution. NIH should therefore recognize not only the initial act of sharing but also the scientific stewardship required to sustain high-value community resources over time, including expert curation, quality control, annotation, updating, interoperability, documentation, and development of tools that make shared information scientifically useful.

Consistent recognition of these contributions could expand the value of oral health datasets, practice-based research network data, software and other resources beyond the institutions that initially developed them. NIH should, however, account for the technical and administrative burden associated with preparing materials for responsible sharing, particularly for smaller laboratories, clinical practices and community-based research networks. Phased requirements, technical assistance and funding for data management and documentation could improve feasibility and reduce disparities in institutional capacity.

3. Training and Mentorship: NIH should incorporate measurable indicators of mentorship quality and trainee development into the assessment of institutional and individual awards.

Relevant indicators could include trainee retention, scholarly contributions, completion of research or career-development milestones, placement across a range of professional settings, and continued participation in research, teaching, clinical investigation, public health or related scientific activities. These measures should recognize that successful trainees may pursue varied career paths and should not presume that placement in a traditional academic research position is the only evidence of effective mentorship.

NIH should also support cohort-based and cross-disciplinary mentorship programs that develop research, teaching, leadership, and other professional competencies. HRSA's Primary Care Dental Faculty Development Center Program provides one model for supporting the development of junior dental faculty through structured training and institutional collaboration. Similar NIH approaches could help dental schools and academic health centers that may not have enough research-active faculty in a single discipline to sustain a traditional mentorship pipeline.

Implementation should remain flexible. Cross-disciplinary programs should complement, rather than replace, discipline-specific scientific mentorship and should not impose substantial new coordination burdens on faculty or institutions with limited administrative resources.

4. Collaboration: NIH should use standardized contributor-role frameworks, such as the Contributor Roles Taxonomy, to recognize the full range of contributions to interdisciplinary and team-based science. Authorship position alone may not adequately capture the work of biostatisticians, data scientists, clinicians, community partners, research coordinators, software developers and other contributors. NIH

should allow these roles to be documented consistently in appropriate application and reporting materials and should provide clear guidance so their recognition does not depend on the practices of individual reviewers or institutions.
NIH should also ensure that community-based research partners are appropriately involved and resourced. Funding structures should account for the direct and indirect costs attributable to community partners' roles, including staffing, data collection, patient engagement, and administrative responsibilities. Direct funding, subawards or other arrangements may be appropriate when they allow community organizations to participate meaningfully in study design, implementation, interpretation and dissemination rather than serving solely as recruitment or data-collection sites.
To improve feasibility for smaller institutions and early-career investigators, NIH should provide model language, templates and technical assistance for contributor-role documentation and community-partnership agreements.

5. Entrepreneurship and Translation: NIH should recognize a broad range of translational outputs, including patents, licensing agreements, technology-transfer milestones, clinical tools, guidelines, validated measures, implementation models and public-health interventions, while recognizing that commercialization is one of several pathways through which research may achieve translational impact. Patents and licensing activity may provide useful evidence that a discovery is progressing toward application, particularly in fields such as dental devices, diagnostics, and biomaterials. These indicators should not, however, be treated as the sole or preferred measure of translation. Many important research outputs advance clinical care or public health without resulting in a commercial product, while other discoveries may require appropriate intellectual-property protection to support further development and enable convergence to eventual translation.
NIH should allow investigators to document commercial and noncommercial translational milestones using appropriate, verifiable materials. Review guidance should also account for differences in institutional scale, infrastructure, and access to technology-transfer and commercialization resources, so that investigators are evaluated on the significance and progress of their translational work rather than on the size of the institutional support structure available to them. Small, independent research institutes such as the ADA Forsyth Institute demonstrate that relatively small institutions without an affiliated medical center can nevertheless achieve excellence across the continuum from foundational discovery to translational research and should not be disadvantaged by impact measures that implicitly favor the infrastructure of large academic health centers.
NIH should further guard against incentives that could encourage unnecessary proprietary restrictions on data, methods, or research materials. Measures of entrepreneurial impact should be balanced with NIH's goals for transparency, validation, and responsible sharing. Evaluation should recognize that intellectual property protection and research sharing can serve complementary purposes and should avoid creating incentives either to restrict research resources unnecessarily or to discourage appropriate protection of discoveries needed for successful development and translation.

6. Foundational Scientific Exploration: NIH should protect support for foundational, high-risk, and high-reward research and assess its impact using criteria appropriate to the scientific question, stage of discovery, and time required for important advances to emerge.
Foundational research should not be disadvantaged because it lacks an immediate clinical, commercial or economic application. NIH should recognize rigorous methods, the development of new scientific concepts or tools, as well as the generation of important negative findings that are derived from well-

designed, appropriately controlled, carefully executed, and adequately powered studies. Rigorous investigation that lead to the resolution of significant scientific uncertainties, and the opening of productive new lines of inquiry are meaningful forms of progress that should be supported.
These questions should be investigated using study designs and analytical approaches appropriate to the claims being tested, without presuming causation or effects that have not been firmly established. Their scientific value may include determining the strength and consistency of observed associations, identifying plausible mechanisms, defining populations in which effects may differ, and generating the evidence needed to inform larger prospective, mechanistic, or interventional studies where appropriate.
NIH should also encourage cross-institute and multidisciplinary research examining potential relationships among oral health interventions, oral disease burden, systemic health outcomes, health care utilization, and quality of life. Such research could include studies evaluating oral hygiene interventions and respiratory outcomes or examining how visual, motor, or other functional limitations affect oral disease burden and access to care.
Their value lies in determining whether relationships exist, identifying relevant mechanisms and establishing the evidence needed to guide clinical care and health-system decision-making. Clear, field-appropriate evaluation standards would help preserve accountability without imposing conventional short-term productivity expectations on research whose impact may take years to emerge.

7. Public Impact: NIH should develop standardized field-appropriate measures of public impact that recognize contributions to clinical practice, health outcomes, public policy, accessibility, public trust and community benefit, where such outcomes are relevant to the scientific aims and stage of research. Potential indicators could include citations in clinical guidelines or public policy, adoption of validated clinical or public-health tools, incorporation of findings into care-delivery models, meaningful community engagement and measurable effects on patient or population outcomes. These indicators should supplement, rather than replace, rigorous scientific assessment and should not be applied in ways that disadvantage foundational, mechanistic, exploratory, or early-stage research for which downstream public impact may not yet be measurable by current methods within the timeframe of the research.
For oral health research, NIH should consider validated patient-reported and health-outcome measures, including the Oral Health Impact Profile where appropriate, as potential secondary endpoints in studies that are not primarily dental. Procedure-based measures remain important, but may not independently capture quality of life, functional status, prevention, access or the effects of interventions that are not themselves dental procedures.
Because nonbibliometric measures may be more difficult to standardize and verify, NIH should develop clear definitions, documentation requirements, and review rubrics. These safeguards would reduce inconsistent application across institutes and discourage superficial activities undertaken solely to satisfy an impact metric.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/101226_ADA_Recommendations_to_NIH.pdf

ADA American Dental Association®

400 C Street N.E.
Washington, D.C. 20002
202.898.2400 • ADA.org

August 12, 2026

NIH Office of Science Policy
Office of the Director
National Institutes of Health
6705 Rockledge Drive, Suite 630
Bethesda, MD 20892

Re: Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Dear NIH Office of Science Policy:

As the leading authority on oral health in the United States, the American Dental Association (ADA) appreciates the opportunity to comment on NIH Notice NOT-OD-26-087, "Request for Information on Measuring and Rewarding Scientific Impact."

Dental, oral and craniofacial research is an integral part of the biomedical research enterprise. It encompasses foundational, clinical, behavioral, translational and population-based research and contributes to the development of evidence-based prevention and treatment, improved health outcomes and a better understanding of the relationship between oral health and overall health. NIH policies governing how scientific impact is measured and rewarded directly affect dental schools, academic health centers, community-based research partners, clinician-scientists and interdisciplinary research teams.

The ADA supports broadening the assessment of scientific impact beyond publication counts and citation metrics. At the same time, NIH should avoid replacing one overly narrow set of measures with another. New indicators should be transparent, verifiable, appropriate to the relevant field and research stage, and feasible for investigators and institutions with differing missions and resources. They should also recognize that scientific impact may emerge over different timelines and through different outputs, including replication, data and resource sharing, mentorship, collaboration, clinical translation and public benefit.

Rigor and Reproducibility

NIH should establish dedicated incentives and funding opportunities for replication and reproducibility, including access to small-grant mechanisms such as R03 awards, supplements, or other limited-duration mechanisms where appropriate, while preserving robust support for R01-scale projects that feature rigorous replication and reproducibility.

Replication studies, pre-registration, transparent reporting of methods and analyses, and publication of well-designed negative or null findings should be recognized as meaningful scientific contributions. NIH could incorporate these activities into appropriate application and

progress-report materials and provide reviewers and program staff with guidance for assessing them separately from the production of novel findings.

Recognition should be based on the quality and informativeness of the work rather than the simple presence of a replication label. Funding such rigorous replication projects through a range of award sizes would allow early-career investigators, clinician-scientists, and researchers at dental schools and smaller institutions to participate without requiring the resources or preliminary development associated with a large research project.

A diversified funding structure would increase the efficiency of NIH investments and broaden participation without creating unnecessary competition between replication and discovery. NIH should also ensure that replication incentives do not unintentionally reduce support for exploratory or hypothesis-generating research. Review criteria should distinguish causal conclusions from associational findings based on study design, methodology, and analytic limitations, while recognizing the value of research that appropriately identifies associations warranting further study.

Sharing Data, Software, and Models

NIH should recognize persistent identifiers, including dataset and software digital object identifiers and Research Resource Identifiers, as verifiable evidence of contributions to data, software, model, and research-resource sharing, while assessing scientific impact based on the quality, stewardship, adoption, and reuse of those resources rather than on identifiers or usage counts alone.

Investigators should be able to document these contributions in appropriate application, review and progress-report materials alongside publications and other research outputs. NIH should assess not only whether a dataset, model, software tool, or research resource has been shared, but also whether it is adequately documented, accessible, reusable, appropriately curated, and capable of supporting validation, replication, or additional research. Relevant indicators of impact could include independent scientific reuse, citations or other documented attribution, incorporation into external analyses or resources, adoption by the research community, and evidence that a shared resource has enabled subsequent scientific discoveries.

The Human Oral Microbiome Database (HOMD), developed and sustained at the ADA Forsyth Institute with NIDCR support, provides an example of the long-term scientific value that can result from investing in openly accessible and expertly curated research resources. Since its public launch, HOMD has provided the scientific community with curated taxonomic, genomic, phylogenetic, and bibliographic information and associated bioinformatics resources for microorganisms of the human oral and aerodigestive tract. Its continuing expansion and use across the scientific community illustrate how an NIH-supported research program can produce both new scientific knowledge and a durable shared resource that enables research well beyond the originating institution. NIH should therefore recognize not only the initial act of sharing but also the scientific stewardship required to sustain high-value community resources over time, including expert curation, quality control, annotation, updating, interoperability, documentation, and development of tools that make shared information scientifically useful.

Consistent recognition of these contributions could expand the value of oral health datasets, practice-based research network data, software and other resources beyond the institutions that initially developed them. NIH should, however, account for the technical and administrative burden associated with preparing materials for responsible sharing, particularly for smaller laboratories, clinical practices and community-based research networks. Phased requirements, technical assistance and funding for data management and documentation could improve feasibility and reduce disparities in institutional capacity.

Training and Mentorship

NIH should incorporate measurable indicators of mentorship quality and trainee development into the assessment of institutional and individual awards.

Relevant indicators could include trainee retention, scholarly contributions, completion of research or career-development milestones, placement across a range of professional settings, and continued participation in research, teaching, clinical investigation, public health or related scientific activities. These measures should recognize that successful trainees may pursue varied career paths and should not presume that placement in a traditional academic research position is the only evidence of effective mentorship.

NIH should also support cohort-based and cross-disciplinary mentorship programs that develop research, teaching, leadership, and other professional competencies. HRSA's Primary Care Dental Faculty Development Center Program provides one model for supporting the development of junior dental faculty through structured training and institutional collaboration. Similar NIH approaches could help dental schools and academic health centers that may not have enough research-active faculty in a single discipline to sustain a traditional mentorship pipeline.

Implementation should remain flexible. Cross-disciplinary programs should complement, rather than replace, discipline-specific scientific mentorship and should not impose substantial new coordination burdens on faculty or institutions with limited administrative resources.

Collaboration

NIH should use standardized contributor-role frameworks, such as the Contributor Roles Taxonomy, to recognize the full range of contributions to interdisciplinary and team-based science.

Authorship position alone may not adequately capture the work of biostatisticians, data scientists, clinicians, community partners, research coordinators, software developers and other contributors. NIH should allow these roles to be documented consistently in appropriate application and reporting materials and should provide clear guidance so their recognition does not depend on the practices of individual reviewers or institutions.

NIH should also ensure that community-based research partners are appropriately involved and resourced. Funding structures should account for the direct and indirect costs attributable to community partners' roles, including staffing, data collection, patient engagement, and

administrative responsibilities. Direct funding, subawards or other arrangements may be appropriate when they allow community organizations to participate meaningfully in study design, implementation, interpretation and dissemination rather than serving solely as recruitment or data-collection sites.

To improve feasibility for smaller institutions and early-career investigators, NIH should provide model language, templates and technical assistance for contributor-role documentation and community-partnership agreements.

Entrepreneurship and Translation

NIH should recognize a broad range of translational outputs, including patents, licensing agreements, technology-transfer milestones, clinical tools, guidelines, validated measures, implementation models and public-health interventions, while recognizing that commercialization is one of several pathways through which research may achieve translational impact.

Patents and licensing activity may provide useful evidence that a discovery is progressing toward application, particularly in fields such as dental devices, diagnostics, and biomaterials. These indicators should not, however, be treated as the sole or preferred measure of translation. Many important research outputs advance clinical care or public health without resulting in a commercial product, while other discoveries may require appropriate intellectual-property protection to support further development and enable convergence to eventual translation.

NIH should allow investigators to document commercial and noncommercial translational milestones using appropriate, verifiable materials. Review guidance should also account for differences in institutional scale, infrastructure, and access to technology-transfer and commercialization resources, so that investigators are evaluated on the significance and progress of their translational work rather than on the size of the institutional support structure available to them. Small, independent research institutes such as the ADA Forsyth Institute demonstrate that relatively small institutions without an affiliated medical center can nevertheless achieve excellence across the continuum from foundational discovery to translational research and should not be disadvantaged by impact measures that implicitly favor the infrastructure of large academic health centers.

NIH should further guard against incentives that could encourage unnecessary proprietary restrictions on data, methods, or research materials. Measures of entrepreneurial impact should be balanced with NIH's goals for transparency, validation, and responsible sharing. Evaluation should recognize that intellectual property protection and research sharing can serve complementary purposes and should avoid creating incentives either to restrict research resources unnecessarily or to discourage appropriate protection of discoveries needed for successful development and translation.

Foundational Scientific Exploration

NIH should protect support for foundational, high-risk, and high-reward research and assess its impact using criteria appropriate to the scientific question, stage of discovery, and time required for important advances to emerge.

Foundational research should not be disadvantaged because it lacks an immediate clinical, commercial or economic application. NIH should recognize rigorous methods, the development of new scientific concepts or tools, as well as the generation of important negative findings that are derived from well-designed, appropriately controlled, carefully executed, and adequately powered studies. Rigorous investigation that lead to the resolution of significant scientific uncertainties, and the opening of productive new lines of inquiry are meaningful forms of progress that should be supported.

These questions should be investigated using study designs and analytical approaches appropriate to the claims being tested, without presuming causation or effects that have not been firmly established. Their scientific value may include determining the strength and consistency of observed associations, identifying plausible mechanisms, defining populations in which effects may differ, and generating the evidence needed to inform larger prospective, mechanistic, or interventional studies where appropriate.

NIH should also encourage cross-institute and multidisciplinary research examining potential relationships among oral health interventions, oral disease burden, systemic health outcomes, health care utilization, and quality of life. Such research could include studies evaluating oral hygiene interventions and respiratory outcomes or examining how visual, motor, or other functional limitations affect oral disease burden and access to care.

Their value lies in determining whether relationships exist, identifying relevant mechanisms and establishing the evidence needed to guide clinical care and health-system decision-making. Clear, field-appropriate evaluation standards would help preserve accountability without imposing conventional short-term productivity expectations on research whose impact may take years to emerge.

Public Impact

NIH should develop standardized field-appropriate measures of public impact that recognize contributions to clinical practice, health outcomes, public policy, accessibility, public trust and community benefit, where such outcomes are relevant to the scientific aims and stage of research.

Potential indicators could include citations in clinical guidelines or public policy, adoption of validated clinical or public-health tools, incorporation of findings into care-delivery models, meaningful community engagement and measurable effects on patient or population outcomes. These indicators should supplement, rather than replace, rigorous scientific assessment and should not be applied in ways that disadvantage foundational, mechanistic, exploratory, or early-stage research for which downstream public impact may not yet be measurable by current methods within the timeframe of the research.

For oral health research, NIH should consider validated patient-reported and health-outcome measures, including the Oral Health Impact Profile where appropriate, as potential secondary endpoints in studies that are not primarily dental. Procedure-based measures remain important,

but may not independently capture quality of life, functional status, prevention, access or the effects of interventions that are not themselves dental procedures.

Because no bibliometric measures may be more difficult to standardize and verify, NIH should develop clear definitions, documentation requirements, and review rubrics. These safeguards would reduce inconsistent application across institutes and discourage superficial activities undertaken solely to satisfy an impact metric.

****

The ADA appreciates NIH's efforts to develop measures that better reflect the collaborative, rigorous, and mission-driven nature of modern biomedical science. A balanced framework should recognize varied forms of scientific contribution while remaining feasible across career stages, disciplines, and institution types. The ADA welcomes continued engagement with NIH as it develops and implements these measures.

For questions regarding these comments, please contact Jorge E. Rojas, DDS, MDS, Director of Scientific Affairs, at rojasj@ada.org.

Sincerely,

/s/

Richard J. Rosato, D.M.D.
President

/s/

Nader A. Nadershahi, D.D.A., M.B.A., Ed.D.
Executive Director

JJS:jer

166. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 8/14/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: I encourage NIH to think expansively about "training and mentorship" to include undergraduate research mentorship specifically, not just formally funded trainees (T32s, F-mechanisms, postdocs). My lab runs on a substantial undergraduate research assistant program — students commit meaningful weekly hours across a full academic term, and take on real scientific responsibilities: stimulus design, data collection, data analysis, and increasingly, contributions to manuscript preparation and shared code repositories. This work is currently almost entirely invisible to NIH's own measures of a PI's impact, despite representing a substantial and sustained investment in workforce development.

A useful model to consider may be NSF's Broader Impacts criterion, which formally evaluates a "how does this advance training" component alongside intellectual merit. NIH has no close equivalent — training impact shows up informally in biosketches at best, and isn't a scored review criterion. Concrete possibilities:

- A dedicated, actually-scored training narrative in RPPRs and renewal applications, rather than a largely pro forma section
- Counting undergraduate-authored outputs (posters, papers, code contributions) as legitimate PI-level training impact, distinct from — but connected to — the PI's own publication record

4. Collaboration: Pure quantitative counts — publication counts, citation rates, author position — miss important nuance about how contributions to collaborative science actually happen. As the junior half of a co-PI lab, I've experienced firsthand how my contributions can be undervalued when credit is assumed to be due to the senior person. Substantial intellectual and technical work often doesn't map cleanly onto first-authorship or grant-PI-count.

Some recent changes have helped. Preprints now counting toward productivity, and the cultural shift toward citing open datasets as first-class scholarly contributions, both better reflect where real effort goes. Building on that momentum, I'd encourage NIH to consider:

- Formalizing contribution statements (e.g., CRediT taxonomy) as a recognized, review-relevant part of the biosketch or progress report, rather than leaving them buried in journal front-matter where they aren't weighted in funding decisions. This would let a junior collaborator document "I led the analysis pipeline" or "I co-designed the shared paradigm" as a citable, evaluable contribution distinct from authorship order.
- Recognizing MPI/Co-PI leadership explicitly, not just PI-count, in career-stage criteria (e.g., Early Stage Investigator status, K-award eligibility) — the current framing can implicitly penalize junior faculty who choose collaborative, multi-PI structures over solo R01s early in their careers.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

167. Aimee Edinger

I am responding: On behalf of myself

Name: Aimee Edinger

Name of Organization: UC Irvine

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/14/2026

1. Rigor and Reproducibility: As a regular and diligent reviewer for NIH, I have seen large discrepancies in how the rigor of grants is assessed and graded by reviewers. Merit measured by publishing in high impact journals does not reward rigorous and well controlled research as splash and reputation is prioritized over rigor and controls and a balanced presentation of ALL of the results. Of course, publishing in lower impact journals does not ensure rigor either! Clear and specific standards are needed for how reviewers (and grant writers) should apply standards for replication, statistical analysis, and critical controls to preliminary data and proposed studies. There need to be very specific standards regarding expectations for specific kinds of experiments, not just broad and general statements about rigor and reproducibility that are interpreted very differently by different people. This is also a problem for publications, work that is not rigorous is published even in the "best" journals. From personal experience I can attest that critical controls are often missing, and statistical practices can be suspect, but if the author has sufficient prestige, the work gets published in science, nature, and cell even when these things are pointed out to the editors during review. No wonder there is a replication problem! On the other hand, reviewers for these journals can make dogmatic but factually incorrect statements without editorial pushback at these journals. Who better than NIH to take the lead in providing clear standards for rigor and replication when performing specific kinds of experiments, incorporating feedback from and fostering open discussions within the scientific community? This kind of guidance would also help shape grant budgets to allow rigor and replication and pilot studies to be appropriately planned for and funded.

2. Data, Software, and Model Sharing: current rules for data sharing (e.g. providing ALL primary data in public repositories) is very burdensome without sufficient benefit to justify the effort. Sharing of large datasets in ways that allow reanalysis by others make sense, but putting all pieces of every kind of data in a repository does not make sense and is not likely to be accurately or sufficiently annotated anyway. Sharing large datasets, software, animal models, cell models, and plasmids (or similar reagents) is CRITICAL and should be REQUIRED (many people do not share models and reagents at present and there is no cost to them for not sharing). Asking researchers to place ALL data in a repository is a waste of resources. Make them deposit plasmids and models and cell lines so people can reproduce their work and evaluate these reagents independently makes more sense. This would need a mechanism for enforcement, or at least an incentive mechanism, or people will continue to fail to make these reagents widely available to others. There is no mechanism to report an ignored request to share reagents from NIH funded research or consequence for doing so.

3. Training and Mentorship: The incentive system currently does not fairly reward those who invest time and effort in mentoring and training. How good mentoring is assessed and rewarded definitely needs discussion. There is not an easy answer for how to do this. it is also a problem that people at undergraduate institutions who have significant teaching commitments have their research productivity directly compared to people at medical schools or institutes where significant teaching is not required. This is not a fair comparison for impact on the scientific enterprise. Good teachers make good researchers, but good teaching also takes a lot of time and effort.

4. Collaboration: The policy considering capping the number of grants an investigator can participate in is BAD for collaborations and misses the point. Any restriction should be based on funding $$. Furthermore, NIH should take into account the increased cost of paying personnel and buying supplies over the last 10 years when setting limits on funding. Team science should not just be a bunch of "famous" scientists at "prestige" universities - teams of people from diverse backgrounds who work together, particularly at the same university, can bring humility and collegiality to a collaboration that allows truly new ideas to incubate and develop. Some products of these collaborations will be publications, but it can take a very long time for these collaborations to advance health - publications (number of impact factor) should not be the only metric as they do not reflect the decades it can take to make an idea benefit health (see comments under Entrepreneurship and Translation) and the many years it takes to accumulate citations.

5. Entrepreneurship and Translation: The paper about modified RNA bases that provided a foundation for the development of the COVID vaccine was rejected from Nature. Dr. Kariko (Nobel laureate!) never had an NIH grant of her own! Something is wrong here. The grant system needs a separate track for innovative grants beyond the DP1 - innovation can take a long slog and really SHOULD require preliminary data and the "outstanding track record" part is not rational. It takes years of dedication and diligent work to go bench to bedside - the COVID vaccine science story should shape how you think about funding innovative work and innovative scientists who are not already famous. Moreover, keeping work patentable/developable is not always compatible with publication and this is not recognized. Entrepreneurship is tremendous work, and often fails at first - so what is strong track record really? A commitment to innovation and entrepreneurship is not appropriately recognized and rewarded with the present review structure. Translational research that could benefit patients is reviewed by people who simply do not understand this process - SBIR review is a mess (I have been on both ends), and the entrepreneurial process is too difficult to participate in while maintaining an academic appointment.

6. Foundational Scientific Exploration: you need to change the peer review system if you want to fund innovative research. The reviewers often lack entrepreneurial experience and/or are not diligent in their assessment of grants - there is little reward for participating in peer review and no cost for failing to participate or serving but failing to provide diligent, accurate, EVIDENCE BASED reviews. Funding famous people to do innovative work with no preliminary data and very large budgets is not a good strategy (DP1). Neither is funding people with many publications in journals that do not require rigor. R21 awards are too small and too short to do the work that is needed to open new areas of investigation. The target researchers and key features of the DP1 (giving large amount of money to researchers that publish in science, nature, and cell without preliminary data needed) is not a foundation that supports the reality of new therapeutic advances with truly new ideas at their root. I think you need a realistic entrepreneurial track award path focused on translational science from everyone and everywhere, not just the prestige crowd, that is founded on preliminary data and that assesses the risk reward balance

more accurately and with more weight on REWARD. Innovation to market takes DECADES in biology, and funding people based on a "track record of success" simply does not make sense.

7. Public Impact: You need stakeholders - the public - to have a say in NIH's broad goals and priorities. The American Cancer Society and CDMRP involves stakeholders in review, the NIH should too. This could potentially help the public understand and appreciate science more as well. Science that addresses the most important unmet needs to be prioritized for the limited funding available while scientists clearly need to assess if the approach is sound. A 2-3 sentence narrative statement is not an effective way to make researchers accountable for taxpayer priorities being considered in research.

8. Other Areas: N/A

168. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/14/2026

1. Rigor and Reproducibility: Create an NIH non-profit electronic journal explicitly to publish reproducibility findings. Include short format micropublications and longer formats as well. This could be an excellent venue to support rigor and reproducibility while also supporting training and publications from faculty at resource limited institutions. For example, an undergraduate or junior graduate student could do a reproducibility study on an important experiment as part of their training/foundational work on their thesis. If this was a small publishable unit at the NIH Journal for RIGOR and Reproducibility, they could learn about the data standards for publication, and how to author a small publication, and they and their mentor would generate citable, valuable (but likely not high impact) work. This kind of work is important for the community and could be valuable to these participants but is likely not going to be taken on by large labs at the edge of their fields; but it could also provide a check on those labs that are leading their fields.

2. Data, Software, and Model Sharing: Provide federally funding, secure, open-access databases for these data, and work with ORCID to provide a way of tracking dataset deposition to facilitate it becoming a productivity and impact metric since the goal it to reward the creation and sharing of these datasets.

3. Training and Mentorship: I am the PI of a lab in a department that awards MS and BS degrees. I do not have access to PhD students or Postdocs. The bulk of my research effort is in training individuals that can only conduct research for 10-20 hours per week. Many of these people join my group having read less than 10 scientific papers in their career to date. I have placed 6 into biomedical/technical manufacturing, 8 into health related doctoral programs, 10 into biomedical MS degrees, and 7 into R1 PhD programs since I joined the faculty in 2018. An additional 8 are still working on applications to graduate or professional programs. I would like to believe that this is valuable work at the base of the biomedical workforce pipeline. I am able to report these outcomes at my institutional level, and I do report in my RPPR each year about trainee outcomes and contributions, but to my knowledge, that is it. This is considerable, and I would argue valuable, work. Maybe there could be a "trainee tracker" program and PI's could get a mentor score component based on the outcomes of their trainees. This could be algorithmically integrated with ORCID with users needing to add a field for the name of their undergraduate, MS, PhD and PostDoc mentors, for example. Maybe?

4. Collaboration: Make more awards that are specifically for research teams with multiple Co-Is

5. Entrepreneurship and Translation: I think the bottleneck between IP ownership by academic institutions and private companies is most of the bottleneck, and so focusing efforts to smooth that out, as well as increased trainings on how to navigate it, might make it more accessible. Many people don't want to bother because it is typically very time consuming, difficult, and prone to failure at late stages.

6. Foundational Scientific Exploration: Make more funding available for basic and lower model organism research; potentially consider specific categories earmarked for that. Stop creating narratives that only translational and clinical work is fundable. Promote basic findings more and have a stronger media presence valorizing and contextualizing basic work to increase public trust. Make social media shorts reflecting on how, for example, basic work in understanding how a fruit fly eye forms was critical to understanding EGFR signaling, which empowers our breakthroughs in cancer research today. Often times the value of basic work takes time to fully be clear, so retrospective "connections" type videos could be really great on social media: How can studying a fruit fly eye help cure cancer? How does watching a worm develop unlock the secrets of cell death? How does studying bacterial immune systems lead to a personalized genome editing cure for sickle cell anemia?
There are so many options and I think if it was well done it could be a great way to increase public trust in science.

7. Public Impact: Promote basic findings more and have a stronger media presence valorizing and contextualizing basic work to increase public trust. Make social media shorts reflecting on how, for example, basic work in understanding how a fruit fly eye forms was critical to understanding EGFR signaling, which empowers our breakthroughs in cancer research today. Often times the value of basic work takes time to fully be clear, so retrospective "connections" type videos could be really great on social media: How can studying a fruit fly eye help cure cancer? How does watching a worm develop unlock the secrets of cell death? How does studying bacterial immune systems lead to a personalized genome editing cure for sickle cell anemia?
There are so many options and I think if it was well done it could be a great way to increase public trust in science.

8. Other Areas: N/A

169. Good Science Project

I am responding: On behalf of an organization

Name: Stuart Buck

Name of Organization: Good Science Project

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/14/2026

1. Rigor and Reproducibility: The NIH Director's Office should put serious resources from the Common Fund (e.g., at least $100m a year) into large-scale replication efforts. These can take a number of forms:

1) Randomly sampling across some field, domain, or IC, so as to provide systematic and representative evidence on the rate of replicability in that area.

2) Sampling from the most highly-funded and/or highly-cited scientists, so as to see the rate of replicability as to labs that are getting the most money and attention.

3) Sampling from some high-impact disease area that has stalled (e.g., Alzheimer's), so as to help us tell whether the replication problem is part of the reason for the lack of progress.

All three forms of sampling make sense, and should be pursued in parallel.

As well, due to the length of time it takes to perform the normal peer review process at NIH, plus the length of time it takes to set up replication experiments, we highly recommend that NIH use OTA contracts to engage with outside organizations, such as the Center for Open Science (COS). COS led a huge DARPA contract that performed the most extensive replication effort in the world to date, involving 865 researchers and resulting in 9 papers as a collection in Nature: https://www.cos.io/score They could spin up an NIH replication project quickly and could start showing impact soon.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: High-risk, high-reward science is something of a misnomer. Many scientific breakthroughs come not from doing anything that was identified as "high risk" ahead of time, but from exploring areas thought to be irrelevant--such as mRNA research in the 1990s, or the early CRISPR research on archaebacteria, or the early research on taq polymerase by Thomas Brock, or Hamilton Smith's early research on restriction enzymes.

If we want to encourage more breakthroughs, we need a system that tolerates a little more curiosity-driven research in obscure bacteria, non-model organisms, botanical specimens, and more.
Moreover, we need to systematically reward scientists who are honest about failures and null results. Specifically, NIH should either require scientists to publish failures and null results on a regular basis, or should give an extra "thumb on the scale" in reviewing future grant applications if the scientist has published failures and null results.
Nobel winner Robert Lefkowitz has said, "Science is 99 percent failure, and that's an optimistic view." https://today.duke.edu/2012/11/lefkowitzdukeidea Any scientist who is actually pushing the frontier of human knowledge cannot truthfully claim a record of 100% success as to every experiment, every idea, etc. When scientists are discouraged from talking openly about failure, that leads to many systematic inefficiencies: 1) the rest of the field doesn't have a chance to learn about dead ends, mistakes, etc.; 2) scientists are incentivized to study marginal questions where they already have a high chance of getting positive results; and 3) scientists are incentivized to exaggerate or p-hack their results to avoid failures.

7. Public Impact: N/A

8. Other Areas: N/A

170. Esmeralda Casas-Silva. Ph.D.; Kanakadurga Addepalli, Ph.D.

I am responding: On behalf of myself

Name: Esmeralda Casas-Silva. Ph.D.; Kanakadurga Addepalli, Ph.D.

Name of Organization:

Type of Organization: Government Agency

Role: Other

Role - Other: Health Services Administrator

Date Submitted: 8/14/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: In considering approaches to incentivize and reward data, software, and model sharing, NIH could recognize contributions across the research product lifecycle, including responsible and timely sharing; fitness for reuse; discoverability, accessibility, and interoperability; independent reuse; and downstream scientific impact. Indicators and incentives could reflect contributions across this continuum while accounting for differences in disciplines, data types, research products, and research communities.

Potential indicators

Examples of measurable indicators that could be considered include:

Responsible and timely sharing: Whether data, software, and models are shared through appropriate repositories or other mechanisms and within timeframes appropriate to the research context. Observable indicators could include use of appropriate repositories, assignment of persistent identifiers, timing of release, and documentation of access conditions. Measures should recognize legitimate restrictions related to participant privacy and consent, security, and other applicable ethical, legal, or governance considerations. Controlled access or other appropriate sharing mechanisms should not be interpreted as a lack of sharing when open access is not appropriate.

Fitness for reuse: Whether shared research products contain sufficient information and documentation for others to understand, evaluate, and appropriately reuse them. Observable indicators could include the presence and completeness of metadata, data dictionaries, provenance and processing documentation, quality-control information, software or model version information, and clear terms of reuse. For software, models, and computational workflows, relevant evidence could also include documentation sufficient for use or implementation and information necessary to reproduce or extend the resource. Where appropriate, alignment with relevant community standards and FAIR principles could provide additional evidence of fitness for reuse.

Discoverability, accessibility, and interoperability: Whether research products can be found,

appropriately accessed, and integrated by the intended research community. Observable indicators could include indexing in relevant catalogs or search systems, availability through appropriate access mechanisms, and use of community standards, formats, terminologies, common data elements, or computational interfaces that facilitate integration with other research products.

Independent reuse: Evidence that data, software, or models are used beyond the originating project or team, regardless of whether that use has yet resulted in a subsequent scientific output. Observable indicators could include documented use in independent research studies, application to new research questions, integration with other datasets, use in secondary or validation analyses, or incorporation into new analytic workflows, software, or models.

Downstream scientific impact: Evidence that sharing and reuse contribute to subsequent scientific outputs, findings, methods, or applications. Indicators could include citations or other attribution in publications and research products; derived datasets; new software, models, methods, or workflows; subsequent research building on the shared resource; or documented instances in which reuse contributed to new findings or applications.

NIH's ongoing Data Sharing Index (S-index) initiative offers an opportunity to test and evaluate several of these concepts in practice. The S-index Challenge considers factors including FAIR practices, timeliness, completeness of annotation, subsequent data use, and related scientific outputs, which align with several of the indicators described above. Lessons from this initiative could inform how indicators are applied across research contexts and how they might be used to recognize and incentivize high-quality data sharing.

These indicators capture related but distinct dimensions of scientific contribution. Investigators can control the quality and timeliness of their sharing practices but have less control over whether or when others reuse their work. Independent reuse and downstream impact, in turn, may provide additional evidence of scientific value. Considering these dimensions together could provide a more complete picture of the scientific contribution of sharing.

It may also be useful to distinguish compliance with applicable sharing requirements from indicators of scientific contribution or impact. While compliance is essential, incentives could additionally recognize practices that increase the usability, interoperability, and scientific value of shared research products.

Potential approaches to incentivize and recognize contributions

Recognize shared research products as scientific contributions. NIH could consider mechanisms that allow investigators and teams to identify significant datasets, software, models, workflows, and other shared research products in applications and relevant reporting mechanisms. Information about an individual's or team's role in developing, curating, documenting, maintaining, or sharing these products, together with available evidence of their use or impact, could complement traditional publication-based measures of scientific contribution.

Strengthen attribution of shared research products and their reuse. Consistent citation practices,

persistent identifiers, and standardized attribution mechanisms could make it easier to connect research products to their creators and contributors and to subsequent research that uses them. Where feasible, automated capture of attribution and reuse information from repositories and other research systems could improve recognition while minimizing additional reporting burden.

Incentivize reuse as well as sharing. Recognition of rigorous reuse of existing research products could complement incentives for original investigators to make their work reusable. Relevant activities could include secondary analysis, data integration, validation and replication studies, development or evaluation of models, or enhancement of existing resources through curation, annotation, or harmonization.

Recognize stewardship and sustainability. High-value data resources, software, models, and computational tools may require ongoing curation, maintenance, documentation, harmonization, security updates, or user support to remain scientifically useful. NIH could consider ways to recognize these activities as substantive scientific contributions and, where appropriate, support the personnel and infrastructure necessary to sustain valuable shared resources.

Potential benefits and considerations for implementation

A multidimensional approach could provide a more complete picture of the scientific value of sharing by recognizing both the quality and usability of shared research products and evidence that they enable subsequent research, collaboration, and discovery. It could also provide greater recognition for scientific and technical contributions that may not be fully reflected in publication- and citation-based measures.

Care should be taken to avoid over-reliance on easily quantified measures such as numbers of datasets deposited, downloads, users, API calls, citations, or software activity. These measures may provide useful contextual information but do not necessarily indicate meaningful scientific use or impact. For example, a highly specialized resource may have relatively few users but enable an important scientific advance, while a widely accessed resource may not necessarily result in substantive reuse.

Measures should also account for differences across disciplines, research products, career stages, institutions, and the sensitivity or maturity of the resource. Appropriate controlled-access sharing should be recognized alongside open sharing, and investigators should not be disadvantaged when ethical, legal, privacy, security, consent, or governance considerations appropriately limit access.

Where possible, multiple indicators could be considered alongside qualitative information about the nature and significance of a contribution. A context-sensitive approach could help incentivize and recognize the quality, reuse, sustainability, and scientific impact of shared research products without reducing these contributions to a single metric or score or creating unnecessary administrative burden.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

171. University of New Mexico Health Sciences

I am responding: On behalf of an organization

Name: Stacy Catanach

Name of Organization: University of New Mexico Health Sciences

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 8/14/2026

1. Rigor and Reproducibility: UNM HSC Research Administration supports NIH's continued emphasis on rigor, reproducibility, transparency, and replication as essential components of high-quality biomedical research. Recognition of researchers who undertake replication studies, develop reproducible methodologies, transparently report results, and contribute to validation of scientific findings could strengthen the research enterprise and encourage activities that may not be adequately recognized through traditional publication and citation metrics.

Potential indicators could include completion and dissemination of replication or validation studies; use of transparent and well-documented research methods; preregistration when appropriate to the discipline; availability of protocols, analytic methods, and supporting documentation; and demonstrated adoption or reuse of reproducible methods by other researchers.

However, measures of rigor and reproducibility should remain primarily within the scientific evaluation process. Institutions should not be required to independently certify scientific rigor, validate reproducibility, or establish new administrative monitoring systems to verify scientific practices. NIH should also recognize that reproducibility may be demonstrated differently across basic, clinical, translational, behavioral, population health, and other research disciplines.

We believe implementation should build upon existing NIH requirements and scientific peer review rather than create duplicative proposal, progress-reporting, or institutional certification requirements. Where additional information is needed, NIH should use standardized investigator-reported information and existing federal systems whenever possible.

NIH should also consider providing dedicated funding for replication and reproducibility activities. Expectations for additional validation, replication, documentation, data management, or quality-control activities may require additional personnel, infrastructure, statistical support, and institutional resources. Expectations should therefore be accompanied by appropriate financial support rather than becoming unfunded institutional responsibilities.

2. Data, Software, and Model Sharing: UNM HSC Research Administration supports appropriate sharing of scientific data, software, models, protocols, and other research resources. Responsible sharing can increase the value of federal research investments, facilitate validation and secondary research, reduce

unnecessary duplication, and promote collaboration.

Potential measures of impact could include deposition of data or research resources in recognized repositories; assignment of persistent identifiers; documented reuse or citation of datasets, software, models, or protocols; development of broadly utilized research tools; and adherence to FAIR principles where appropriate.

NIH should recognize, however, that the ability to share and the cost of sharing vary significantly by research type. Human participant data, protected health information, proprietary information, Tribal data, sensitive datasets, intellectual property, controlled-access data, and large or technically complex datasets may require substantial institutional infrastructure and oversight.

Institutions may incur significant costs associated with data curation, storage, cybersecurity, repository fees, privacy review, access controls, software maintenance, licensing, and long-term preservation. If NIH expands expectations in this area, adequate funding should be available to support these activities.

Measures should recognize responsible sharing rather than simply the volume of information made public. A dataset that is appropriately restricted to protect participants should not be considered less valuable than an openly accessible dataset when restrictions are scientifically, ethically, legally, or culturally appropriate.

NIH should also rely upon existing repositories, persistent identifiers, and machine-readable information to measure sharing and reuse whenever possible rather than requiring institutions to manually track downstream use of research products.

3. Training and Mentorship: UNM HSC Research Administration strongly supports recognition of training and mentorship as important contributions to the biomedical research enterprise. Effective mentorship contributes to workforce development, research integrity, scientific independence, retention, and the development of future generations of researchers.

Potential indicators could include completion of structured mentoring activities; trainee progression and achievement of educational or career milestones; successful transition to independent research roles; development of mentoring plans; participation in mentor training; and demonstrated contributions to the professional development of students, trainees, junior investigators, and research staff.

We believe NIH should avoid relying solely on traditional outcomes such as the number of trainees who subsequently obtain NIH funding or tenure-track academic positions. Successful biomedical research careers include clinical research, industry, government, research administration, public health, technology transfer, data science, and other sectors that contribute to the research ecosystem.

Long-term career tracking also presents a substantial feasibility concern. Institutions may not have access to reliable information about individuals years after they complete training or leave the institution. NIH should use federal data sources, ORCID, researcher profiles, and voluntary trainee reporting where possible rather than placing responsibility on institutions to maintain indefinite longitudinal tracking systems.

Measures should also recognize the quality and depth of mentorship rather than simply the number of individuals mentored. Incentives that emphasize volume could unintentionally encourage superficial mentoring relationships.

4. Collaboration: UNM HSC Research Administration supports increased recognition of collaborative and team-based science. Modern biomedical research frequently requires multidisciplinary teams, shared infrastructure, specialized expertise, community partnerships, clinical collaborations, and contributions from individuals whose scientific impact may not be reflected through first- or senior-author publications.

Potential indicators could include documented investigator contribution statements; participation in multidisciplinary or multi-institutional research; development or leadership of collaborative research infrastructure; contributions to shared datasets, cores, methods, or research resources; and successful integration of expertise across disciplines.

NIH should recognize the full range of meaningful scientific contributions without requiring institutions to independently quantify or validate the relative contribution of every member of a research team. Standardized contribution frameworks, such as structured investigator contribution statements, may provide a more practical approach.

We believe NIH should also avoid measures that unintentionally favor large research teams or highly resourced institutions simply because they generate larger numbers of collaborations. Quality, scientific necessity, and meaningful contribution should be considered alongside quantity.

Recognition should extend beyond investigators to the research professionals and infrastructure that make collaborative science possible, including biostatisticians, data scientists, research coordinators, core facilities, research administrators, and other specialized contributors where appropriate.

5. Entrepreneurship and Translation: UNM HSC Research Administration supports recognizing activities that successfully translate scientific discoveries into products, technologies, diagnostics, therapeutics, clinical practices, and other applications that benefit patients and the public.

Potential indicators could include invention disclosures, patents, licensing activity, startup formation, industry partnerships, advancement of technologies toward clinical use, implementation of evidence-based interventions, regulatory milestones, and adoption of research findings into healthcare practice.

We believe NIH should exercise caution, however, in relying heavily on commercialization-based measures such as patents, licensing revenue, startup formation, or private investment. Such measures may favor institutions with substantial commercialization infrastructure or research portfolios concentrated in areas that readily produce patentable technologies.

Scientific translation also occurs without commercialization. Research may change clinical guidelines, improve healthcare delivery, inform public health practices, develop noncommercial tools, improve diagnostic approaches, or result in interventions adopted by communities or healthcare systems.

NIH should therefore define translation broadly enough to recognize multiple pathways from discovery to impact.

In addition, commercialization outcomes can occur many years after the underlying NIH-supported research. Institutions and investigators should not be expected to establish direct causal attribution between an individual NIH award and a downstream commercial product when numerous researchers, organizations, funding sources, and subsequent discoveries contributed to the outcome.

6. Foundational Scientific Exploration: UNM HSC Research Administration strongly supports NIH's recognition of foundational research and high-risk, high-reward scientific exploration. Fundamental discoveries frequently precede clinical or commercial applications by many years, and their ultimate impact may not be foreseeable during the period of an NIH award.

NIH should therefore avoid evaluating foundational research primarily through short-term measures such as commercialization, immediate clinical impact, citation counts, or measurable public-health outcomes.

Potential indicators could instead recognize generation of new hypotheses; development of novel methodologies; creation of foundational datasets or research resources; opening of new areas of scientific inquiry; replication or refutation of established assumptions; and contributions that subsequently enable other scientific discoveries.

NIH should also recognize that scientifically valuable research does not always produce the anticipated result. Well-designed studies that generate negative, null, or unexpected findings can contribute substantially to scientific knowledge and prevent unnecessary duplication of research.

Measures intended to reward impact should not create incentives for investigators or institutions to avoid scientifically important but uncertain research in favor of projects more likely to generate immediately measurable outcomes.

UNM HSC support's NIH preserving peer review and scientific judgment as central components of evaluating foundational research and should ensure that quantitative impact metrics supplement rather than replace expert assessment.

7. Public Impact: UNM HSC Research Administration supports NIH's goal of recognizing the broader public impact of biomedical research, including improvements in health outcomes, clinical practice, accessibility, public trust, economic activity, and societal benefit.

This category presents significant opportunities but also substantial measurement challenges.

Potential indicators could include incorporation of research findings into clinical guidelines or healthcare practice; implementation of evidence-based interventions; improvements in healthcare access or delivery; adoption of research findings by public agencies or community organizations; contributions to workforce development; public-facing research resources; community engagement; and demonstrated

improvements in health outcomes where attribution is reasonably possible.

We would hope that NIH recognize that public impact may look substantially different depending upon an institution's mission and the populations it serves. Measures should appropriately recognize research conducted in rural, underserved, and geographically dispersed communities, as well as research focused on improving healthcare delivery and access.

Public impact should not be measured primarily through economic return, commercialization, media attention, or other measures that may disproportionately favor certain research areas or highly resourced institutions.

Attribution also requires caution. Changes in health outcomes or clinical practice frequently result from decades of research involving numerous investigators, institutions, funding agencies, healthcare systems, and community partners. Requiring an investigator or institution to demonstrate that a specific NIH award directly produced a particular societal outcome could create substantial administrative burden while producing unreliable measures.

8. Other Areas: UNM HSC Research Administration supports NIH's effort to develop a broader and more meaningful understanding of scientific impact. Moving beyond publication counts, citation metrics, and grant funding has the potential to more appropriately recognize the many activities necessary to advance biomedical research.

However, implementation will be critical. UNM HSC suggests that NIH:

Use existing information whenever possible. NIH should leverage federal systems, repositories, persistent identifiers, ORCID, PubMed, and other machine-readable sources rather than requiring institutions and investigators to repeatedly enter the same information.

Avoid new institutional certifications. Measures intended to recognize scientific contributions should not evolve into institutional compliance certifications requiring central administrative offices to validate scientific impact.

Provide funding for new expectations. Expanded requirements for data sharing, reproducibility, long-term preservation, tracking, reporting, or research infrastructure should be accompanied by appropriate funding.

Maintain proportionality. Reporting expectations should be proportional to the size, complexity, duration, and risk of the funded activity.

Recognize institutional and disciplinary differences. Metrics should not unintentionally disadvantage institutions because of geography, institutional resources, research portfolio, patient population, commercialization infrastructure, or discipline.

Avoid excessive longitudinal tracking requirements. Institutions should not be expected to indefinitely track trainees, downstream data reuse, commercialization, clinical implementation, or societal outcomes

after an award has ended.

Separate scientific evaluation from administrative compliance. Indicators designed to recognize scientific contributions should primarily inform scientific evaluation rather than create additional administrative review and monitoring responsibilities.

Pilot new measures before broad implementation. NIH should test proposed indicators across different institution types, disciplines, career stages, and award mechanisms before establishing agency-wide requirements and assess both scientific value and administrative burden.

Avoid overreliance on quantitative metrics. Measures should supplement, rather than replace, expert scientific judgment. Metrics that become targets may unintentionally influence researcher behavior in ways that prioritize measurable activity over meaningful scientific contribution.

Provide clear definitions and standardized guidance. Any new measures should be consistently defined across NIH Institutes and Centers to minimize inconsistent interpretation and institutional implementation.

Overall, UNM HSC Research Administration supports NIH's objective of recognizing the broader range of activities that contribute to rigorous, collaborative, innovative, and publicly beneficial biomedical research. We encourage NIH to develop approaches that reward these contributions while ensuring that the mechanisms used to measure scientific impact do not themselves divert significant investigator and institutional resources away from conducting research.

172. Association of Biomolecular Resource Facilities

I am responding: On behalf of an organization

Name: Joe Dragavon

Name of Organization: Association of Biomolecular Resource Facilities

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/15/2026

1. Rigor and Reproducibility: The Association of Biomolecular Resource Facilities (ABRF) Committee for Core Rigor and Reproducibility (CCoRRe) commends the NIH for examining how scientific impact is measured and rewarded. CCoRRe is a standing committee within ABRF dedicated to advancing the rigor, reproducibility, and transparency of scientific research.

Core facilities, also referred to as Shared Research Resources (SRRs) and Shared Resource Laboratories (SRLs), are essential components of the biomedical research ecosystem. These entities provide not only advanced instrumentation, but also deep technical expertise, training, quality management, technology development, data stewardship, and community leadership. Together, these contributions directly support scientific excellence and reinforce public trust in research.

We are submitting these comments because core facilities generate a substantial proportion of research data at many institutions and serve as institutional anchors for rigorous, reproducible science. Strengthening standards within these environments has meaningful downstream impact on research quality and reliability; as such, strengthening standards within these environments has a broad and meaningful downstream impact on the quality and reliability of science overall.

Current assessment systems tend to emphasize publications, citation counts, journal metrics, and funding success, while underrecognizing critical contributions that enhance the rigor, reproducibility, transparency, accessibility, and long-term value of biomedical research.

NIH has a valuable opportunity to broaden its definition of scientific impact to more fully recognize and incentivize contributions that advance rigor, reproducibility, and transparency, and that enable high-quality, reliable science. We support NIH's efforts to align measures of scientific success with outcomes that strengthen scientific integrity and maximize public health impact.

Reproducibility and replicability are foundational to rigorous biomedical research. Core facilities and SRLs are uniquely positioned to enhance research quality through systemic implementation of technology expertise, quality assurance, protocol standardization, assay validation, experimental design consultation, and workforce training.

ABRF Recommendations

NIH should incentivize, measure, and recognize:

• Adoption and reporting rates of Research Resource Identifiers (RRIDs) across publications, datasets, and protocols
• Development and dissemination of validated protocols, standard operating procedures (SOPs), and reference materials (e.g., downloads, citations, reuse)
• Participation in benchmarking studies, proficiency testing programs, and interlaboratory comparisons, including performance relative to established standards
• Implementation of quality management systems, including metrics such as instrument uptime, calibration frequency, assay validation rates, and error reduction trends
• Publication and dissemination of replication studies, validation studies, and negative findings
• Documented contributions by core facilities and SRLs to study design, methodology selection, quality control, and data integrity (e.g., inclusion in grant applications, protocols, or contributor roles)

These indicators should be incorporated into grant review criteria, progress reports, and renewal evaluations to reward demonstrated commitment to rigor and reproducibility.
Scientific impact assessments should also include evidence that research outputs are reproducible, transparent, reusable, and supported by appropriate methodological documentation.

2. Data, Software, and Model Sharing: The value of scientific research extends beyond publications to include datasets, software, analytical workflows, protocols, and models that enable future discovery and validation.

ABRF CCoRRe Recommendations

NIH should recognize, measure, and incentivize:
• Adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, including compliance rates in data management plans and demonstrated accessibility of datasets within 12 months of study completion.
• Sharing of datasets, metadata, software tools, analytical workflows, protocols, and computational pipelines, measured by:
o Number of publicly available datasets/tools
o Dataset downloads and reuse frequency
o Citations of datasets and software
• Development and adoption of standards for data formatting, annotation, interoperability, and provenance tracking.
• Use of persistent identifiers (RRIDs, ORCID iDs, DOIs), measured by rates of inclusion in publications and datasets.
• Long-term stewardship metrics, including duration of data availability, dataset updates, and maintenance activities.
• Contributions by core facilities and SRLs to shared data infrastructure, including number of supported users, datasets managed, and pipelines maintained.

NIH could integrate these metrics into funding decisions, data sharing compliance evaluations, and

supplemental funding incentives for high-impact data resources.
Measures of impact should include data reuse, software adoption, protocol utilization, resource citations, and community uptake rather than relying solely on publication-based metrics.

3. Training and Mentorship: Core facilities and SRLs are major contributors to workforce development and provide training that is often not captured by traditional academic metrics.

ABRF CCoRRe Recommendations

NIH should recognize, measure, and reward:
• Training activities conducted by core facility scientists and staff, including:
o Number of trainees per year
o Training hours delivered
o Types of skills taught (technology, data analysis, rigor practices)
• Development and dissemination of educational programs, measured by adoption rates, downloads, and use across institutions.
• Mentorship outcomes, including trainee career progression, publications, and continued use of rigorous practices.
• Contributions to workforce development that improve scientific competency and research quality.
• Training efficacy metrics, including pre- and post-training assessments, trainee feedback, and demonstrated adoption of rigorous practices by trainees in their own research.

NIH could incorporate these metrics into grant progress reports and core facility funding mechanisms and provide supplemental funding or recognition for high-impact training programs. Potential indicators of impact include numbers of individuals trained, adoption of training materials, trainee outcomes, development of educational resources, and demonstrated improvements in research quality attributable to training activities.

4. Collaboration: Modern biomedical research increasingly relies on interdisciplinary teams that include investigators, clinicians, engineers, statisticians, data scientists, and core facility experts.
ABRF CCoRRe Recommendations

NIH should expand recognition of team science by:
• Developing mechanisms to document and quantify contributions using contributor role taxonomies (e.g., CRediT), including counts of technical, methodological, and advisory contributions.
• Encouraging use of ORCID identifiers to track contributions across projects and publications.
• Measuring participation in multidisciplinary, multi-investigator, and multi-institutional projects.
• Recognizing contributions that enable scientific success even when they are not reflected through traditional authorship positions (e.g., acknowledgments, documented consultation hours, technical impact statements).
These contributions should be considered in grant scoring, promotion of collaborative funding mechanisms, and evaluation of investigator impact.
Core facility scientists frequently contribute specialized expertise that directly influences study design, technology implementation, data interpretation, and research outcomes. These contributions should be recognized as meaningful indicators of scientific impact.

We recommend NIH develop guidance on appropriate roles for core facility scientists in authorship vs. acknowledgment to ensure transparent credit attribution while respecting disciplinary norms.

5. Entrepreneurship and Translation: Core facilities frequently serve as innovation hubs where new technologies are developed, optimized, validated, and disseminated.

ABRF CCoRRe Recommendations

NIH should recognize, measure, and incentivize:
• Development of novel methodologies, workflows, instrumentation, software tools, and analytical approaches, measured by adoption across projects and institutions.
• Technology validation and implementation metrics, including number of validated assays and deployment across user groups.
• Public-private partnerships and collaborations, including number of partnerships and resulting outputs.
• Technology transfer outcomes, including patents filed, licenses executed, and commercial products developed.
• Dissemination activities, measured by training workshops, protocol sharing, and cross-institutional adoption.
NIH could incorporate these indicators into translational funding mechanisms and reward successful implementation and dissemination through supplemental funding or programmatic recognition.
Measures of impact should include technology adoption, implementation across institutions, commercialization activities, and demonstrated improvements in research capabilities. Core facility-led innovation often addresses unmet methodological needs identified through direct engagement with end-user researchers, positioning these facilities as natural innovation hubs."

6. Foundational Scientific Exploration: Many transformative discoveries arise from exploratory research, technology development, and high-risk investigations whose value may not be immediately apparent.

ABRF CCoRRe Recommendations

NIH should encourage and measure:
• Investment in novel technologies and proof-of-concept studies, including number and outcomes of pilot projects.
• Utilization of shared research resources for exploratory work, including pilot usage rates and downstream funding success.
• Technical innovation outputs, such as newly developed methods or workflows and their later adoption.
• Long-term impact indicators, including downstream publications, grant funding, and technology diffusion over time.
NIH should align funding review criteria to value long-term influence and provide dedicated funding mechanisms that support early-stage, high-risk innovation.
Core facilities often provide environments where investigators can test innovative ideas, generate preliminary data, and evaluate new technologies.

7. Public Impact: Scientific impact should extend beyond academic metrics and include broader societal benefit.

ABRF CCoRRe Recommendations

NIH should recognize and measure:
• Improvements in reproducibility and research quality, including reductions in irreproducible findings and retraction rates.
• Contributions to clinical translation, public health outcomes, and evidence-based practices.
• Accessibility of scientific outputs, including availability of open datasets, protocols, and educational materials.
• Community engagement activities, including outreach events, training programs, and public-facing resources.
• Economic efficiencies, such as cost savings achieved through shared infrastructure and resource utilization.
• Return on investment metrics, including reuse of federally funded resources and downstream impact of funded work.

These measures could inform program evaluations and demonstrate public value, strengthening accountability for federal investments.
Investments in rigor, reproducibility, quality management, and scientific stewardship provide substantial public value by ensuring that scientific discoveries are trustworthy and reusable.

8. Other Areas: ABRF and CCoRRe recommend NIH consider several additional areas:

• Recognition of Technical Professionals – Career advancement metrics, retention rates, and documented contributions to research outputs.
• Shared Research Resource Contributions – Metrics such as number of users served, projects supported, and publications enabled.
• Quality Management and Accreditation – Adoption rates of quality frameworks, benchmarking participation, and performance improvement metrics.
• Research Resource Stewardship – Duration of resource availability, usage rates, and maintenance activities.
• Community Standards Development – Participation in standards initiatives and adoption rates of developed guidelines.
• Scientific Service and Leadership – Roles in professional societies, committees, and consensus-building efforts.

Cross-Cutting Recommendation: Recognition of Shared Research Resources and Core Facility Contributions

ABRF and the CCoRRe Committee encourage NIH to recognize SRRs, core facilities, and SRLs as critical contributors across all areas addressed in this RFI.

NIH should consider developing mechanisms to document, track, and reward these contributions through:

• Standardized acknowledgment and attribution of core facility contributions in publications and grants.

• Integration of ORCID identifiers and contributor role taxonomies into NIH reporting systems.
• Inclusion of core facility impact metrics (e.g., number of supported projects, training outputs, data generated) in grant applications and progress reports.
• Recognition of facilities that demonstrate excellence in rigor, reproducibility, quality management, and FAIR data practices.
• Incorporation of shared resource contributions into funding decisions, renewals, and institutional evaluations, including development of standardized reporting formats to reduce burden on core facility administrators."

Importantly, many activities essential to rigorous science—such as quality assurance, method validation, benchmarking, standards development, data stewardship, and training—are often performed by core facility personnel and are not adequately captured by traditional metrics.
Because shared research resources contribute directly to all NIH priority areas identified in this RFI, NIH should incorporate recognition of SRR and SRL contributions as a cross-cutting element within any future framework.

Conclusion

NIH has an opportunity to reshape scientific incentives by recognizing rigor, reproducibility, transparency, stewardship, training, collaboration, and shared research infrastructure as essential dimensions of scientific impact.

A modern assessment framework should reward not only scientific discoveries but also the measurable practices, contributions, and infrastructure that ensure research is rigorous, reproducible, transparent, reusable, and trustworthy.

Recognizing these contributions will strengthen the biomedical research enterprise, improve public trust, maximize the return on federal investments, and better align scientific incentives with practices that produce reliable and impactful science.
We are prepared to provide data, case studies, and further detail on implementing these recommendations, and welcome collaboration with NIH to pilot assessment frameworks.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/ABRF-CCoRRe-RFI-Response-081426.pdf

Description: Recommendations on behalf of the Association of Biomolecular Resource Facilities

Association of Biomolecular Resource Facilities
201 E. Main Street
Suite 810
Lexington, KY 40507
Tel: 859-514-0028 ♦ www.abrf.org Email: abrf@abrf.org

August 15, 2026

NIH Request for Information: Measuring and Rewarding Scientific Impact

Notice Number: NOT-OD-26-087

The Association of Biomolecular Resource Facilities (ABRF) Committee on Core Rigor and Reproducibility (CCoRRe) commends the NIH for examining how scientific impact is measured and rewarded. CCoRRe is a standing committee within ABRF dedicated to advancing the rigor, reproducibility, and transparency of scientific research.

Core facilities, also referred to as Shared Research Resources (SRRs) and Shared Resource Laboratories (SRLs), are essential components of the biomedical research ecosystem. These entities provide not only advanced instrumentation, but also deep technical expertise, training, quality management, technology development, data stewardship, and community leadership. Together, these contributions directly support scientific excellence and reinforce public trust in research.

We are submitting these comments because core facilities generate a substantial proportion of research data at many institutions and serve as institutional anchors for rigorous, reproducible science. Strengthening standards within these environments has meaningful downstream impact on research quality and reliability; as such, strengthening standards within these environments has a broad and meaningful downstream impact on the quality and reliability of science overall.

Current assessment systems tend to emphasize publications, citation counts, journal metrics, and funding success, while underrecognizing critical contributions that enhance the rigor, reproducibility, transparency, accessibility, and long-term value of biomedical research.

NIH has a valuable opportunity to broaden its definition of scientific impact to more fully recognize and incentivize contributions that advance rigor, reproducibility, and transparency, and that enable high-quality, reliable science. We support NIH's efforts to align measures of scientific success with outcomes that strengthen scientific integrity and maximize public health impact.

SECTION 1: RIGOR AND REPRODUCIBILITY

Reproducibility and replicability are foundational to rigorous biomedical research. Core facilities and SRLs are uniquely positioned to enhance research quality through systemic implementation of technology expertise, quality assurance, protocol standardization, assay validation, experimental design consultation, and workforce training.

ABRF CCoRRe Recommendations

NIH should incentivize, measure, and recognize:

- Adoption and reporting rates of Research Resource Identifiers (RRIDs) across publications, datasets, and protocols
- Development and dissemination of validated protocols, standard operating procedures (SOPs), and reference materials (e.g., downloads, citations, reuse)
- Participation in benchmarking studies, proficiency testing programs, and interlaboratory comparisons, including performance relative to established standards
- Implementation of quality management systems, including metrics such as instrument uptime, calibration frequency, assay validation rates, and error reduction trends
- Publication and dissemination of replication studies, validation studies, and negative findings
- Documented contributions by core facilities and SRLs to study design, methodology selection, quality control, and data integrity (e.g., inclusion in grant applications, protocols, or contributor roles)

These indicators should be incorporated into grant review criteria, progress reports, and renewal evaluations to reward demonstrated commitment to rigor and reproducibility.
Scientific impact assessments should also include evidence that research outputs are reproducible, transparent, reusable, and supported by appropriate methodological documentation.

SECTION 2: DATA, SOFTWARE, AND MODEL SHARING

The value of scientific research extends beyond publications to include datasets, software, analytical workflows, protocols, and models that enable future discovery and validation.

ABRF CCoRRe Recommendations

NIH should recognize, measure, and incentivize:

- Adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, including compliance rates in data management plans and demonstrated accessibility of datasets within 12 months of study completion.
- Sharing of datasets, metadata, software tools, analytical workflows, protocols, and computational pipelines, measured by:
    - Number of publicly available datasets/tools
    - Dataset downloads and reuse frequency
    - Citations of datasets and software

- Development and adoption of standards for data formatting, annotation, interoperability, and provenance tracking.
- Use of persistent identifiers (RRIDs, ORCID iDs, DOIs), measured by rates of inclusion in publications and datasets.
- Long-term stewardship metrics, including duration of data availability, dataset updates, and maintenance activities.
- Contributions by core facilities and SRLs to shared data infrastructure, including number of supported users, datasets managed, and pipelines maintained.

NIH could integrate these metrics into funding decisions, data sharing compliance evaluations, and supplemental funding incentives for high-impact data resources.
Measures of impact should include data reuse, software adoption, protocol utilization, resource citations, and community uptake rather than relying solely on publication-based metrics.

SECTION 3: TRAINING AND MENTORSHIP

Core facilities and SRLs are major contributors to workforce development and provide training that is often not captured by traditional academic metrics.

ABRF CCoRRe Recommendations

NIH should recognize, measure, and reward:

- Training activities conducted by core facility scientists and staff, including:
    - Number of trainees per year
    - Training hours delivered
    - Types of skills taught (technology, data analysis, rigor practices)
- Development and dissemination of educational programs, measured by adoption rates, downloads, and use across institutions.
- Mentorship outcomes, including trainee career progression, publications, and continued use of rigorous practices.
- Contributions to workforce development that improve scientific competency and research quality.
- **Training efficacy metrics**, including pre- and post-training assessments, trainee feedback, and demonstrated adoption of rigorous practices by trainees in their own research.

NIH could incorporate these metrics into grant progress reports and core facility funding mechanisms and provide supplemental funding or recognition for high-impact training programs.

Potential indicators of impact include numbers of individuals trained, adoption of training materials, trainee outcomes, development of educational resources, and demonstrated improvements in research quality attributable to training activities.

SECTION 4: COLLABORATION

Modern biomedical research increasingly relies on interdisciplinary teams that include investigators, clinicians, engineers, statisticians, data scientists, and core facility experts.

ABRF CCoRRe Recommendations

NIH should expand recognition of team science by:

- Developing mechanisms to document and quantify contributions using contributor role taxonomies (e.g., CRediT), including counts of technical, methodological, and advisory contributions.
- Encouraging use of ORCID identifiers to track contributions across projects and publications.
- Measuring participation in multidisciplinary, multi-investigator, and multi-institutional projects.
- Recognizing contributions that enable scientific success even when they are not reflected through traditional authorship positions (e.g., acknowledgments, documented consultation hours, technical impact statements).

These contributions should be considered in grant scoring, promotion of collaborative funding mechanisms, and evaluation of investigator impact.
Core facility scientists frequently contribute specialized expertise that directly influences study design, technology implementation, data interpretation, and research outcomes. These contributions should be recognized as meaningful indicators of scientific impact.
We recommend NIH develop guidance on appropriate roles for core facility scientists in authorship vs. acknowledgment to ensure transparent credit attribution while respecting disciplinary norms.

SECTION 5: ENTREPRENEURSHIP AND TRANSLATION

Core facilities frequently serve as innovation hubs where new technologies are developed, optimized, validated, and disseminated.

ABRF CCoRRe Recommendations

NIH should recognize, measure, and incentivize:

- Development of novel methodologies, workflows, instrumentation, software tools, and analytical approaches, measured by adoption across projects and institutions.

- Technology validation and implementation metrics, including number of validated assays and deployment across user groups.
- Public-private partnerships and collaborations, including number of partnerships and resulting outputs.
- Technology transfer outcomes, including patents filed, licenses executed, and commercial products developed.
- Dissemination activities, measured by training workshops, protocol sharing, and cross-institutional adoption.

NIH could incorporate these indicators into translational funding mechanisms and reward successful implementation and dissemination through supplemental funding or programmatic recognition.
Measures of impact should include technology adoption, implementation across institutions, commercialization activities, and demonstrated improvements in research capabilities. Core facility-led innovation often addresses unmet methodological needs identified through direct engagement with end-user researchers, positioning these facilities as natural innovation hubs."

SECTION 6: FOUNDATIONAL SCIENTIFIC EXPLORATION

Many transformative discoveries arise from exploratory research, technology development, and high-risk investigations whose value may not be immediately apparent.

ABRF CCoRRe Recommendations

NIH should encourage and measure:

- Investment in novel technologies and proof-of-concept studies, including number and outcomes of pilot projects.
- Utilization of shared research resources for exploratory work, including pilot usage rates and downstream funding success.
- Technical innovation outputs, such as newly developed methods or workflows and their later adoption.
- Long-term impact indicators, including downstream publications, grant funding, and technology diffusion over time.

NIH should align funding review criteria to value long-term influence and provide dedicated funding mechanisms that support early-stage, high-risk innovation.
Core facilities often provide environments where investigators can test innovative ideas, generate preliminary data, and evaluate new technologies.

SECTION 7: PUBLIC IMPACT

Scientific impact should extend beyond academic metrics and include broader societal benefit.

ABRF CCoRRe Recommendations

NIH should recognize and measure:

- Improvements in reproducibility and research quality, including reductions in irreproducible findings and retraction rates.
- Contributions to clinical translation, public health outcomes, and evidence-based practices.
- Accessibility of scientific outputs, including availability of open datasets, protocols, and educational materials.
- Community engagement activities, including outreach events, training programs, and public-facing resources.
- Economic efficiencies, such as cost savings achieved through shared infrastructure and resource utilization.
- Return on investment metrics, including reuse of federally funded resources and downstream impact of funded work.

These measures could inform program evaluations and demonstrate public value, strengthening accountability for federal investments.
Investments in rigor, reproducibility, quality management, and scientific stewardship provide substantial public value by ensuring that scientific discoveries are trustworthy and reusable.

SECTION 8: ADDITIONAL AREAS FOR CONSIDERATION

ABRF CCoRRe recommend NIH consider several additional areas:

- **Recognition of Technical Professionals** – Career advancement metrics, retention rates, and documented contributions to research outputs.
- **Shared Research Resource Contributions** – Metrics such as number of users served, projects supported, and publications enabled.
- **Quality Management and Accreditation** – Adoption rates of quality frameworks, benchmarking participation, and performance improvement metrics.
- **Research Resource Stewardship** – Duration of resource availability, usage rates, and maintenance activities.

- **Community Standards Development** – Participation in standards initiatives and adoption rates of developed guidelines.
- **Scientific Service and Leadership** – Roles in professional societies, committees, and consensus-building efforts.

Cross-Cutting Recommendation: Recognition of Shared Research Resources and Core Facility Contributions

ABRF and the CCoRRe Committee encourage NIH to recognize SRRs, core facilities, and SRLs as critical contributors across all areas addressed in this RFI.

NIH should consider developing mechanisms to document, track, and reward these contributions through:

- Standardized acknowledgment and attribution of core facility contributions in publications and grants.
- Integration of ORCID identifiers and contributor role taxonomies into NIH reporting systems.
- Inclusion of core facility impact metrics (e.g., number of supported projects, training outputs, data generated) in grant applications and progress reports.
- Recognition of facilities that demonstrate excellence in rigor, reproducibility, quality management, and FAIR data practices.
- Incorporation of shared resource contributions into funding decisions, renewals, and institutional evaluations, including development of standardized reporting formats to reduce burden on core facility administrators."

Importantly, many activities essential to rigorous science—such as quality assurance, method validation, benchmarking, standards development, data stewardship, and training—are often performed by core facility personnel and are not adequately captured by traditional metrics. Because shared research resources contribute directly to all NIH priority areas identified in this RFI, NIH should incorporate recognition of SRR and SRL contributions as a cross-cutting element within any future framework.

Conclusion

NIH has an opportunity to reshape scientific incentives by recognizing rigor, reproducibility, transparency, stewardship, training, collaboration, and shared research infrastructure as essential dimensions of scientific impact.

A modern assessment framework should reward not only scientific discoveries but also the measurable practices, contributions, and infrastructure that ensure research is rigorous, reproducible, transparent, reusable, and trustworthy.

Recognizing these contributions will strengthen the biomedical research enterprise, improve public trust, maximize the return on federal investments, and better align scientific incentives with practices that produce reliable and impactful science.

We are prepared to provide data, case studies, and further detail on implementing these recommendations, and welcome collaboration with NIH to pilot assessment frameworks.

Committee on Core Rigor and Reproducibility
Association of Biomolecular Resource Facilities

173. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Organizational Official

Date Submitted: 8/15/2026

1. Rigor and Reproducibility: I would commend two areas for enhancing rigor and reproducibility:
1)Rigor: encourage registered reports as a form of publishing. Much science is wasted or slowed through poorly thought out controls and designs. Negative results that are useful to the field but not ground breaking also rarely find their way into publications. Registered Reports, with peer-reviewed methods and provisional acceptance regardless of the results, are a way to address these. Perhaps they could be encouraged with some kind of "priority review voucher" for future grant submissions, much the same as the highly successful FDA program.

2) Reproducibility: encourage each new PhD or masters trainee to have a replication project as part of their work portfolio. This allows them to learn the field, learn techniques, and publish on a shorter timescale than their main thesis. Dedicated repository or journals for this would probably help as well.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: As a leader at a small biotech that has applied to and successfully completed SBIRs, I feel strongly that this program is crucial to the science ecosystem. It's impact can be amplified by accelerating the timelines on which awards are made. The lag time of ~9 months simply is too long for the pace that many innovative biotechs want to work at. Can companies be vetted a single time and then have expedited reviews in the future? Can administrative fluff be removed to condense that time between step?

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

174. Licklider, Inc.

I am responding: On behalf of an organization

Name: Tasuku Kobayashi

Name of Organization: Licklider, Inc.

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Organizational Official

Date Submitted: 8/16/2026

1. Rigor and Reproducibility: DISCLOSURE. The respondent is a company developing verification software for statistical analyses in the life sciences. This response does not ask NIH to adopt, evaluate, or endorse any specific product. A current implementation by the respondent is cited only where explicitly identified as feasibility evidence; the policy proposal does not depend on that implementation.

SUMMARY. NIH already requires investigators to describe plans for rigor and for data sharing, and operates extensive disclosure and reporting instruments across the research life cycle. What remains underdeveloped is a standardized, NIH-wide mechanism for independently machine-checking reported quantitative results against their declared evidence, declared methods, and specified computation; NIH's own implementation plan describes composite measures in this area as still under development. We propose one primary candidate measurable indicator — Machine-Verifiable Coverage, reported as a vector rather than a single ratio — and one secondary indicator suitable only for piloting, an Independent Verification Pass Rate, together with the open infrastructure and measurement procedure needed to compute them. Both are indicators of transparency and checkability, not of scientific quality or validity, and we propose the measurement construct itself for pilot validation: we do not ask NIH to treat it as validated. We state the guarantee boundary explicitly and repeatedly, because misreading verification status as evidence of validity is itself the most foreseeable unintended consequence of any indicator in this area. Since this RFI was issued, OSTP has separately called for building verification infrastructure at scale — investment in verification systems and open standards, standards enabling machine-auditable replication packages, and rewards for those who replicate or disprove influential results (Science: A New Golden Age, July 2026). This response addresses a narrower measurement question within that direction: which verification states should be observable, and how they should be aggregated without implying scientific validity.

DEFINITIONS AND SCOPE. We follow the definitions in the OSTP Gold Standard Science guidance (June 2025) and the NIH implementation plan (August 2025), under which replicability denotes performing the same experiment or study using the same methods and conditions to achieve the same result, and reproducibility denotes independent researchers testing a hypothesis through multiple methods. We note explicitly that these assignments are reversed relative to the National Academies consensus report (NASEM 2019), where reproducibility denotes obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis. The property our indicators measure corresponds to NASEM's computational reproducibility, a property that has no dedicated name in the

2025 scheme. To avoid ambiguity, this response uses the term machine-verifiable execution consistency: the ability of software, following an open specification, to confirm that reported quantitative results are consistent with the declared data, the declared methods, and the specified computation. Such checking establishes consistency with declarations; it does not establish that the declared analysis was the one historically performed. NASEM (2019) is explicit that "exact reproducibility does not guarantee the correctness of the computation." Nothing proposed below claims otherwise, and the indicator design treats that limit as a design input rather than a caveat.

THE MEASUREMENT GAP. NIH already operates extensive policy instruments across the research life cycle — the rigor and transparency review criteria in effect since FY2016, authentication plans, the Data Management and Sharing Policy in effect since 2023, the Genomic Data Sharing Policy, the 2024 Public Access Policy, and clinical trial registration and results reporting — including post-execution disclosure and reporting. What remains underdeveloped is a standardized, NIH-wide mechanism for independently machine-checking reported quantitative computations against their declared inputs and methods. A decade of empirical work shows why that layer matters and why statements are weak proxies for practice. In a study of 3,556 articles across 333 BMC journals, 123 of 1,792 manuscripts (6.8 percent) whose authors had stated in a data availability statement that they would share data actually provided it when asked (Gabelica et al. 2022). In a large-scale study of more than 2,000 publicly deposited research code-and-data packages, 74 percent of the R files they contained failed to complete on initial execution (Trisovic et al. 2022). Among 25 articles carrying open-data badges in a leading psychology journal, only 36 percent could be fully analytically reproduced from the original data without author assistance — computational reproducibility in the NASEM sense — and the dominant failure cause was analytical procedures too ambiguously reported to re-execute (Hardwicke et al. 2021). At the same time, policy stringency and enforcement are associated with substantial changes in reporting, data reusability, and re-execution outcomes: a mandatory open-data policy at one journal was followed by an increase in data availability statements from 25 to 78 percent of articles, with data judged reusable in principle rising from 22 to 62 percent (Hardwicke et al. 2018), and journal policy strictness correlates with re-execution success (Trisovic et al. 2022). Finally, the most mature machine checks in existence are automated consistency checks of reported statistics: for supported, machine-detected APA-style null-hypothesis test results, statcheck validation studies reported classification accuracy between 96.2 and 99.9 percent — which does not imply equivalent detection coverage of all statistical results in an article — and such checks have found at least one internally inconsistent statistical result in roughly half of checkable psychology articles (Nuijten et al. 2016; Nuijten, van Assen, Hartgerink, Epskamp & Wicherts 2017), with granularity checks finding a similar rate in a separate sample (Brown & Heathers 2017). These findings support testing execution-side indicators — provided their guarantee boundary is designed in from the start.

PROPOSED INDICATOR 1 — MACHINE-VERIFIABLE COVERAGE (MVC): A CANDIDATE INDICATOR PROPOSED FOR PILOT VALIDATION. MVC measures how much of a research output's quantitative claims can be, and has been, independently machine-checked. It is deliberately reported as a vector, not a single ratio, because collapsing it to one number invites both gaming and misreading. For a given article or project, MVC reports:

- N_major: the number of major quantitative claims, drawn from a claim inventory declared by the

authors and confirmed during review;
- N_eligible: the subset of those claims that fall under a published eligibility rule (i.e., a public verification profile exists for the analysis class);
- N_recorded: the subset represented in machine-verifiable structured records — including, where applicable, claims inside a declared closed verification surface, defined below;
- N_verified: the subset whose records have passed independent verification under the applicable public verification profile (record integrity, declared provenance relations, specified computation, numerical consistency);
- N_bound: the subset of verified claims whose reported values in the final research artifact have additionally been machine-matched, within declared numerical tolerances, to the corresponding structured record;
- N_unsupported: claims whose analysis class has no verification profile yet, reported as a distinct state and never counted as a failure.

These states are nested: N_bound ≤ N_verified ≤ N_recorded ≤ N_eligible ≤ N_major, with N_unsupported = N_major − N_eligible. Two coverage ratios are derived and always reported together: coverage over all major claims (N_bound / N_major) and coverage over eligible claims (N_bound / N_eligible); when N_eligible = 0, the eligible-claims ratio is reported as NOT APPLICABLE, never as zero. Reporting both ratios is an anti-gaming measure: if only the eligible-claims ratio were used, there would be an incentive to classify inconvenient claims as ineligible; if only the all-claims ratio were used, fields with many not-yet-supported analysis classes would be structurally penalized. The distinct unsupported state prevents unsupported methods from being mislabeled as verification failures; it does not by itself remove disincentives to methodological novelty, since all-claims coverage still falls when many claims are unsupported. For this reason, all-claims coverage must not be used to rank investigators, fields, or methods.

MVC is itself an unvalidated measurement construct, and this response treats that as a first-order fact rather than a footnote. Its measurement properties are not yet established: whether independent raters construct the same claim inventory (inter-rater agreement on N_major), whether eligibility classification is consistent across raters, whether independently developed verifiers agree (implementation concordance), and how sensitive the vector is to choices of claim granularity, are all open questions. We therefore propose MVC as a candidate indicator whose reliability and validity should be a primary object of the pilot described below — not as an indicator ready for adoption.

A declared closed verification surface, referenced above, is a delimited set of research outputs — for example, the specified quantitative elements of a figure, its legend, and the corresponding Methods and Results statements — that are generated from, or explicitly bound to, a single structured analysis record, such that every declared in-scope value or structured relation covered by the applicable verification profile can be machine-checked against that record. Within such a surface, checking is binary and complete relative to the declared surface and its verification contract; the coverage vector then records how much of the article's claim inventory lies inside declared surfaces. The two mechanisms are complementary: surfaces make the numerator crisp, and the inventory-based denominators make selective declaration visible.

The final-artifact value binding component (N_bound) is essential rather than optional. One practical advantage of text-based checks such as statcheck is that they test values as actually published; a structured record, by itself, is a self-declaration. Binding record values to the values reported in the final artifact preserves the property that verification confronts the published claim, not a parallel account of it. Consistency of content outside a declared surface — including any subsequent edits in word processors or publisher typesetting — is established only by this explicit value binding and is never presumed.

WHAT MVC ESTABLISHES AND WHAT IT DOES NOT. For claims counted in N_bound, the measurement establishes only that the covered quantitative elements are represented in structured records, that the reported computations have been independently re-derived from the declared inputs under an open specification, and that the published values match those computations within declared tolerances. It does not establish that the experimental design was appropriate, that the data are authentic, that the recorded analysis was the one historically performed, that the statistical model was well specified, that causal interpretations are warranted, or that the scientific conclusions are true. An internally consistent fabrication would pass; a retrospectively constructed record that is internally consistent would pass; a correct analysis expressed in a not-yet-supported method would register as unsupported, not as failed. Machine verification of this kind is a lower bound on internal consistency — in NASEM's framing, an indicator of transparency serving indirect assessment — and any deployment should prohibit representing verification status as evidence of scientific validity. We recommend that this prohibition be an explicit design requirement of any NIH adoption, including in how results are displayed: verification output should take the form of a vector of named checks with explicit not-evaluated dimensions (for example: artifact integrity, recorded-input binding, specified computation, numerical consistency, final-artifact value binding — each PASS/FAIL/UNSUPPORTED — alongside experimental design, data authenticity, causal validity, independent replication — each NOT EVALUATED), rather than a single "verified" label.

PROPOSED INDICATOR 2 — INDEPENDENT VERIFICATION PASS RATE (IVPR), PILOT ONLY. Among structured records submitted for a defined scope, IVPR is the share that pass independent verification of record integrity, declared provenance relations, specified computation, and numerical consistency. We propose IVPR only for piloting and program evaluation, not for investigator-level assessment, because it lacks predictive-validity evidence and is the component most exposed to Goodhart dynamics if elevated into a reward. Its pilot value is diagnostic: it measures where and why verification fails (environment capture, ambiguous specification, tooling gaps), which is exactly the information needed to improve infrastructure.

INFRASTRUCTURE REQUIRED. Computing these indicators requires: an open, versioned specification for structured analysis records, such that — given the required record and evidence — any authorized third party can execute verification under the public specification without proprietary issuer logic or dependence on any single service (issuer-independent verification, a property of open specifications and third-party execution rather than of any count of implementations); public conformance fixtures so that implementations can be checked against one another; ideally, more than one independently developed verifier implementation with distinct numerical backends, to reduce shared-implementation error — with cross-implementation concordance itself measured in the pilot rather than assumed; a registry with

trusted timestamps so that selective disclosure among registered records can be detected — unregistered analyses remain outside detection; and reuse of existing standards vocabularies rather than invention — W3C PROV for provenance relations, RO-Crate and CodeMeta for packaging and software metadata, FAIR4RS for research software, and the NISO RP-31-2021 reproducibility badging taxonomy for any user-facing status vocabulary, with terminology explicitly mapped to the definitions used in this response.

Institutional precedents demonstrate that third-party verification is operable in practice: the ACM TOMS Replicated Computational Results initiative (since 2015), the CODECHECK certificate register, the cascad certification agency (including third-party checking of analyses on restricted data inside a secure computing environment), and publisher-side computational verification workflows. This proposal does not claim novelty in third-party computational verification itself; those programs demonstrate that such verification events are operationally feasible. The additional element proposed here is to test whether those events can be represented as a standardized, machine-readable measurement state — combining a public eligibility profile, an independent recomputation outcome, final-artifact value binding, explicit not-evaluated dimensions, and portfolio-level coverage denominators — designed for funder-level aggregation and integration with existing processes. Whether that additional measurement layer is useful to NIH is a question for the proposed pilot, not an assumption of this response.

As evidence that a substrate of this kind is implementable (offered as feasibility evidence, not as evidence of policy merit): an open record schema, an independently runnable open-source verifier, and public conformance fixtures for a deliberately narrow, explicitly experimental verification contract covering one analysis class (Welch two-group comparisons) are publicly available under an Apache-2.0 license, released by the respondent and fixed by the immutable version identifier given in the attached references, so that this evidence does not change after submission. The current demonstration implements only the structured-record and scoped statistical-recomputation portion of this proposal. It does not implement MVC claim inventories, closed verification surfaces, final-artifact value binding, registry-based coverage measurement, multiple independent verifier implementations, or secure-enclave execution; those are proposed pilot components.

IMPLEMENTATION APPROACHES. Integration should reuse existing process surfaces across the award life cycle rather than create new bureaucracy, and should respect the timing of what exists when. Pre-award, applications can state whether, and for which analysis classes, the project intends to produce verification-eligible artifacts — alongside the rigor and transparency sections and Data Management and Sharing plans that applications already contain; the claim inventory itself concerns completed outputs and cannot exist at application time. Post-award, claim inventories, records, and coverage reporting can attach to annual progress reporting, where NIH already tracks data-sharing implementation, and to publication workflows. Evaluation of any pilot can use the implementation-science frameworks NIH has already designated for Gold Standard Science assessment (CFIR and RE-AIM). NIH could pilot two pathways for the same class of eligible computational analyses. In a legacy pathway — which would require no new authoring environment — a machine-readable record is constructed retrospectively from the analysis artifacts and documentation investigators already produce; this measures the practical ceiling of post-hoc reconstruction, because information that was not captured at execution time cannot reliably be reconstructed later and frequently requires author assistance, as the reproducibility literature

documents. In a native pathway, which would separately test participating tools capable of execution-time record capture, tools emit the record at execution time and may generate a declared closed verification surface from the same structured record. Within that declared surface, deterministic generation can structurally preclude drift along the canonical generation path; consistency with content outside the surface is still established only by explicit value binding and is not presumed. Across the two pathways the pilot should compare verification coverage, missing-information rates, author and reviewer burden, final-artifact binding success, observed gaming behavior, and failure modes. Neither pathway should be treated as producing evidence of scientific validity.

PILOT DESIGN. We recommend a voluntary, registration-based pilot in one or two Institutes or Centers with computation-heavy portfolios, with a concurrent comparison group, prospectively sized with prespecified primary outcomes; any inferential comparison should have a justified sample size. The primary question of the pilot should be measurement itself: can MVC be measured reproducibly, at tolerable burden, without unacceptable gaming or field bias? Concretely, the pilot should measure inter-rater agreement on claim inventories and on eligibility classification; concordance between independently developed verifier implementations; false-pass and false-fail rates against a manual expert-audit benchmark; the sensitivity of the coverage vector to claim granularity; burden by career stage and institution type; observed gaming; and discipline differences. The legacy-versus-native comparison described above is the secondary, implementation-level question. The rationale for building evaluation in from the start is empirical: nominally similar incentive interventions have produced divergent results in different contexts — open-data badges were associated with a rise from under 3 percent to roughly 39 percent data sharing in one journal's observational data, while a randomized trial at another journal did not detect an increase, with a highly imprecise estimate that could not rule out meaningful effects in either direction — so context-dependence should be assumed and measured rather than argued.

POTENTIAL BENEFITS. Execution-side indicators of this kind could add a standardized measurement layer to NIH's developing metric framework where composite measures are still under development; could allow journal reviewers, reproducibility auditors, and programmatic reviewers of completed outputs to spend less attention on arithmetic consistency and more on design and interpretation; could reduce the cost of computational re-execution and of evidence review — whether they affect the cost of experimental replication should be tested rather than assumed; may support long-term preservation of analysis evidence through registry and archival integration; and could give existing mandates (data sharing, rigor criteria) an execution-side outcome measure they currently lack.

UNINTENDED CONSEQUENCES AND ANTI-GAMING DESIGN. We enumerate the principal failure modes and their status honestly, because several are structural rather than fixable. Mitigable by design: surface or claim cherry-picking (mitigated, not eliminated, by the dual denominators and the claim inventory); denominator manipulation via narrow "major claim" definitions (claim inventories confirmed at review); eligibility laundering, i.e., classifying inconvenient claims as ineligible (published eligibility rules plus the all-claims ratio); selective disclosure of passing records (registry with timestamps; unregistered analyses remain undetectable); verifier monoculture and self-verification (third-party execution required; cross-implementation concordance measured); template farming and claim salami-slicing (claim inventories subject to independent review and audit; raw claim counts never themselves used as rewards;

judgments of claim granularity and importance remain a human review function); post-publication mutation of reported values (detectable on re-verification of the published artifact, conditional on re-verification actually being run — not structurally prevented). Structurally unavoidable, and therefore boundary rather than bug: semantic overinterpretation in prose outside declared surfaces; analysis-path selection among internally consistent alternatives; analyses that are wrong in design or model specification yet perfectly executed; and internally consistent fabrication, including retrospective record construction. For these, the honest countermeasure is the display and reporting prohibition described above — verification status must never be presented as validity — plus the separate mechanisms science already uses (audit, raw-data requirements, replication). Two further consequences deserve explicit monitoring: discipline and institutional inequity, since computational fields and well-resourced institutions can generate records more cheaply (partially addressed by scope limitation, hosted verification options, and the unsupported state); and checkbox compliance, the documented fate of data availability statements, which argues for pairing any indicator with spot enforcement rather than self-report alone. Consistent with DORA, the Leiden Manifesto, the Metric Tide, and CoARA, these indicators should support qualitative assessment and never substitute for it; the vector reporting format is chosen specifically so that no single number exists to be ranked on.

FEASIBILITY ACROSS CAREER STAGES, DISCIPLINES, AND INSTITUTION TYPES. Scope should begin with narrow computational analysis classes for which a public verification profile can be specified, benchmarked, and independently tested — the setting in which validated automated checks already exist. Restricted and controlled-access data could participate in a pilot through secure-enclave execution: cascad/CASD demonstrates the feasibility of third-party computational checking without public data release, and adapting the machine-verifiable architecture proposed here to such environments would require separate implementation and validation; issuer independence can in principle be preserved even for hosted execution so long as the specification is open and third parties can execute verification. Smaller institutions without local compute could use hosted verification under the same open contract. Petabyte-scale computations and analyses locked to proprietary software environments should be explicitly out of scope for an initial indicator rather than papered over. The pilot should stratify burden and adoption by career stage rather than assume an advantage or disadvantage in advance.

A PDF version of this response, including full references, is attached.

2. Data, Software, and Model Sharing: This response extends the framework proposed under Topic 1 (Rigor and Reproducibility) to sharing incentives; the definitions, guarantee boundary, anti-gaming design, and pilot structure described there apply here and are not repeated.

FROM EXISTENCE METRICS TO VERIFIABILITY METRICS. Many scalable automated indicators of sharing measure reporting or existence — a data-availability statement, a repository deposit, a code link — rather than successful reuse or re-execution. Existence can be measured at scale: automated extraction across 2.75 million open-access articles estimated mentions of data-sharing information in 14.5 percent of articles and of code-sharing information in 2.5 percent (Serghiou et al. 2021). But the evidence shows that existence alone can be a weak proxy for the outcomes sharing policy actually seeks: 6.8 percent actual provision among authors who had stated willingness to share (Gabelica et al. 2022); 74 percent initial execution failure among the R files in publicly deposited code-and-data packages (Trisovic et al. 2022); and full unassisted analytic reproduction from original data in only about a third of badge-carrying

articles (Hardwicke et al. 2021). Rewarding existence therefore invites the documented checkbox outcome. We propose that NIH pilot whether verified-use and binding states — the full Machine-Verifiable Coverage vector restricted to shared artifacts, interpreted qualitatively — can complement availability and deposit evidence in future sharing incentives. Consistent with Topic 1, this proposes the construct for validation before any connection to rewards; raw counts of bound claims or of deposits, both trivially inflatable, should never themselves be reward-relevant.

SOFTWARE SHARING. NIH-wide research-software sharing is addressed primarily through best-practice guidance rather than a uniform mandatory software-sharing policy. Rather than proposing new mandates, we propose a measurable standard that any future incentive could attach to: software shared with sufficient metadata and environment capture to permit independent re-execution — reusing CodeMeta and FAIR4RS for metadata and interoperability, while a separate verification profile specifies the dependency, environment, and execution requirements that metadata alone does not guarantee — with re-execution success measured rather than assumed. This would convert software sharing from a statement into an event that can be counted, and could provide an execution-side outcome measure for future NIH software-sharing incentives or reporting, without assuming that the Data Management and Sharing Policy itself applies to research software. This initial proposal does not attempt to define a separate indicator for computational-model sharing.

EQUITY ACROSS DATA TYPES. Verification-based sharing measures must not be limited to data that can be posted publicly. Controlled-access and restricted data could participate through verification executed inside secure computing environments, for which an operating precedent exists in third-party checking without public data release (cascad/CASD), with only the verification outcome — not the data — leaving the environment; adapting the proposed machine-verifiable architecture to such environments would require separate implementation and validation. This matters for feasibility across fields: human-subjects research would otherwise be structurally excluded from any verification-linked measure, which would be both inequitable and a perverse incentive against sensitive-data research.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/NOT-OD-26-087_Licklider_response.pdf

Description: PDF version of the responses submitted in the Topic 1 (Rigor and Reproducibility) and Topic 2 (Data, Software, and Model Sharing) fields, with the full reference list. Content is identical to the form fields; the references (33 sources, including the immutable version identifier for the feasibility evidence cited in Topic 1) appear only in this PDF.

Response to NIH Request for Information NOT-OD-26-087

Measuring and Rewarding Scientific Impact

Respondent: Licklider, Inc. (Kabushiki Kaisha Licklider), Tokyo, Japan — responding on behalf of an organization (private company; developer of open verification infrastructure for statistical analyses in the life sciences). Submitted August 2026 via the NIH Office of Science Policy comment form. This document is the PDF version of the form-field responses to Topics 1 and 2, with full references.

Topic 1: Rigor and Reproducibility

DISCLOSURE. The respondent is a company developing verification software for statistical analyses in the life sciences. This response does not ask NIH to adopt, evaluate, or endorse any specific product. A current implementation by the respondent is cited only where explicitly identified as feasibility evidence; the policy proposal does not depend on that implementation.

SUMMARY. NIH already requires investigators to describe plans for rigor and for data sharing, and operates extensive disclosure and reporting instruments across the research life cycle. What remains underdeveloped is a standardized, NIH-wide mechanism for independently machine-checking reported quantitative results against their declared evidence, declared methods, and specified computation; NIH's own implementation plan describes composite measures in this area as still under development. We propose one primary candidate measurable indicator — Machine-Verifiable Coverage, reported as a vector rather than a single ratio — and one secondary indicator suitable only for piloting, an Independent Verification Pass Rate, together with the open infrastructure and measurement procedure needed to compute them. Both are indicators of transparency and checkability, not of scientific quality or validity, and we propose the measurement construct itself for pilot validation: we do not ask NIH to treat it as validated. We state the guarantee boundary explicitly and repeatedly, because misreading verification status as evidence of validity is itself the most foreseeable unintended consequence of any indicator in this area. Since this RFI was issued, OSTP has separately called for building verification infrastructure at scale — investment in verification systems and open standards, standards enabling machine-auditable replication packages, and rewards for those who replicate or disprove influential results (Science: A New Golden Age, July 2026). This response addresses a narrower measurement question within that direction: which verification states should be observable, and how they should be aggregated without implying scientific validity.

DEFINITIONS AND SCOPE. We follow the definitions in the OSTP Gold Standard Science guidance (June 2025) and the NIH implementation plan (August 2025), under which replicability denotes performing the same experiment or study using the same methods and conditions to achieve the same result, and reproducibility denotes independent researchers testing a hypothesis through multiple methods. We note explicitly that these assignments are reversed relative to the National Academies consensus report (NASEM 2019), where reproducibility denotes obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis. The property our indicators measure corresponds to NASEM's computational reproducibility, a property that has no dedicated name in the 2025 scheme. To avoid ambiguity, this response uses the term machine-verifiable execution consistency: the ability of software, following an open specification, to confirm that reported quantitative results are consistent with the declared data, the declared methods, and the specified computation. Such checking establishes consistency with declarations; it does not establish that the

declared analysis was the one historically performed. NASEM (2019) is explicit that "exact reproducibility does not guarantee the correctness of the computation." Nothing proposed below claims otherwise, and the indicator design treats that limit as a design input rather than a caveat.

THE MEASUREMENT GAP. NIH already operates extensive policy instruments across the research life cycle — the rigor and transparency review criteria in effect since FY2016, authentication plans, the Data Management and Sharing Policy in effect since 2023, the Genomic Data Sharing Policy, the 2024 Public Access Policy, and clinical trial registration and results reporting — including post-execution disclosure and reporting. What remains underdeveloped is a standardized, NIH-wide mechanism for independently machine-checking reported quantitative computations against their declared inputs and methods. A decade of empirical work shows why that layer matters and why statements are weak proxies for practice. In a study of 3,556 articles across 333 BMC journals, 123 of 1,792 manuscripts (6.8 percent) whose authors had stated in a data availability statement that they would share data actually provided it when asked (Gabelica et al. 2022). In a large-scale study of more than 2,000 publicly deposited research code-and-data packages, 74 percent of the R files they contained failed to complete on initial execution (Trisovic et al. 2022). Among 25 articles carrying open-data badges in a leading psychology journal, only 36 percent could be fully analytically reproduced from the original data without author assistance — computational reproducibility in the NASEM sense — and the dominant failure cause was analytical procedures too ambiguously reported to re-execute (Hardwicke et al. 2021). At the same time, policy stringency and enforcement are associated with substantial changes in reporting, data reusability, and re-execution outcomes: a mandatory open-data policy at one journal was followed by an increase in data availability statements from 25 to 78 percent of articles, with data judged reusable in principle rising from 22 to 62 percent (Hardwicke et al. 2018), and journal policy strictness correlates with re-execution success (Trisovic et al. 2022). Finally, the most mature machine checks in existence are automated consistency checks of reported statistics: for supported, machine-detected APA-style null-hypothesis test results, statcheck validation studies reported classification accuracy between 96.2 and 99.9 percent — which does not imply equivalent detection coverage of all statistical results in an article — and such checks have found at least one internally inconsistent statistical result in roughly half of checkable psychology articles (Nuijten et al. 2016; Nuijten, van Assen, Hartgerink, Epskamp & Wicherts 2017), with granularity checks finding a similar rate in a separate sample (Brown & Heathers 2017). These findings support testing execution-side indicators — provided their guarantee boundary is designed in from the start.

PROPOSED INDICATOR 1 — MACHINE-VERIFIABLE COVERAGE (MVC): A CANDIDATE INDICATOR PROPOSED FOR PILOT VALIDATION. MVC measures how much of a research output's quantitative claims can be, and has been, independently machine-checked. It is deliberately reported as a vector, not a single ratio, because collapsing it to one number invites both gaming and misreading. For a given article or project, MVC reports:

- N_major: the number of major quantitative claims, drawn from a claim inventory declared by the authors and confirmed during review;
- N_eligible: the subset of those claims that fall under a published eligibility rule (i.e., a public verification profile exists for the analysis class);
- N_recorded: the subset represented in machine-verifiable structured records — including, where applicable, claims inside a declared closed verification surface, defined below;
- N_verified: the subset whose records have passed independent verification under the applicable public verification profile (record integrity, declared provenance relations, specified computation, numerical

consistency);

- N_bound: the subset of verified claims whose reported values in the final research artifact have additionally been machine-matched, within declared numerical tolerances, to the corresponding structured record;
- N_unsupported: claims whose analysis class has no verification profile yet, reported as a distinct state and never counted as a failure.

These states are nested: N_bound ≤ N_verified ≤ N_recorded ≤ N_eligible ≤ N_major, with N_unsupported = N_major – N_eligible. Two coverage ratios are derived and always reported together: coverage over all major claims (N_bound / N_major) and coverage over eligible claims (N_bound / N_eligible); when N_eligible = 0, the eligible-claims ratio is reported as NOT APPLICABLE, never as zero. Reporting both ratios is an anti-gaming measure: if only the eligible-claims ratio were used, there would be an incentive to classify inconvenient claims as ineligible; if only the all-claims ratio were used, fields with many not-yet-supported analysis classes would be structurally penalized. The distinct unsupported state prevents unsupported methods from being mislabeled as verification failures; it does not by itself remove disincentives to methodological novelty, since all-claims coverage still falls when many claims are unsupported. For this reason, all-claims coverage must not be used to rank investigators, fields, or methods.

MVC is itself an unvalidated measurement construct, and this response treats that as a first-order fact rather than a footnote. Its measurement properties are not yet established: whether independent raters construct the same claim inventory (inter-rater agreement on N_major), whether eligibility classification is consistent across raters, whether independently developed verifiers agree (implementation concordance), and how sensitive the vector is to choices of claim granularity, are all open questions. We therefore propose MVC as a candidate indicator whose reliability and validity should be a primary object of the pilot described below — not as an indicator ready for adoption.

A declared closed verification surface, referenced above, is a delimited set of research outputs — for example, the specified quantitative elements of a figure, its legend, and the corresponding Methods and Results statements — that are generated from, or explicitly bound to, a single structured analysis record, such that every declared in-scope value or structured relation covered by the applicable verification profile can be machine-checked against that record. Within such a surface, checking is binary and complete relative to the declared surface and its verification contract; the coverage vector then records how much of the article's claim inventory lies inside declared surfaces. The two mechanisms are complementary: surfaces make the numerator crisp, and the inventory-based denominators make selective declaration visible.

The final-artifact value binding component (N_bound) is essential rather than optional. One practical advantage of text-based checks such as statcheck is that they test values as actually published; a structured record, by itself, is a self-declaration. Binding record values to the values reported in the final artifact preserves the property that verification confronts the published claim, not a parallel account of it. Consistency of content outside a declared surface — including any subsequent edits in word processors or publisher typesetting — is established only by this explicit value binding and is never presumed.

WHAT MVC ESTABLISHES AND WHAT IT DOES NOT. For claims counted in N_bound, the measurement establishes only that the covered quantitative elements are represented in structured records, that the reported computations have been independently re-derived from the declared inputs under an open specification, and that the published values match those computations within declared tolerances. It does not establish that the experimental design was appropriate, that the data are authentic, that the recorded analysis was

the one historically performed, that the statistical model was well specified, that causal interpretations are warranted, or that the scientific conclusions are true. An internally consistent fabrication would pass; a retrospectively constructed record that is internally consistent would pass; a correct analysis expressed in a not-yet-supported method would register as unsupported, not as failed. Machine verification of this kind is a lower bound on internal consistency — in NASEM's framing, an indicator of transparency serving indirect assessment — and any deployment should prohibit representing verification status as evidence of scientific validity. We recommend that this prohibition be an explicit design requirement of any NIH adoption, including in how results are displayed: verification output should take the form of a vector of named checks with explicit not-evaluated dimensions (for example: artifact integrity, recorded-input binding, specified computation, numerical consistency, final-artifact value binding — each PASS/FAIL/UNSUPPORTED — alongside experimental design, data authenticity, causal validity, independent replication — each NOT EVALUATED), rather than a single "verified" label.

PROPOSED INDICATOR 2 — INDEPENDENT VERIFICATION PASS RATE (IVPR), PILOT ONLY. Among structured records submitted for a defined scope, IVPR is the share that pass independent verification of record integrity, declared provenance relations, specified computation, and numerical consistency. We propose IVPR only for piloting and program evaluation, not for investigator-level assessment, because it lacks predictive-validity evidence and is the component most exposed to Goodhart dynamics if elevated into a reward. Its pilot value is diagnostic: it measures where and why verification fails (environment capture, ambiguous specification, tooling gaps), which is exactly the information needed to improve infrastructure.

INFRASTRUCTURE REQUIRED. Computing these indicators requires: an open, versioned specification for structured analysis records, such that — given the required record and evidence — any authorized third party can execute verification under the public specification without proprietary issuer logic or dependence on any single service (issuer-independent verification, a property of open specifications and third-party execution rather than of any count of implementations); public conformance fixtures so that implementations can be checked against one another; ideally, more than one independently developed verifier implementation with distinct numerical backends, to reduce shared-implementation error — with cross-implementation concordance itself measured in the pilot rather than assumed; a registry with trusted timestamps so that selective disclosure among registered records can be detected — unregistered analyses remain outside detection; and reuse of existing standards vocabularies rather than invention — W3C PROV for provenance relations, RO-Crate and CodeMeta for packaging and software metadata, FAIR4RS for research software, and the NISO RP-31-2021 reproducibility badging taxonomy for any user-facing status vocabulary, with terminology explicitly mapped to the definitions used in this response.

Institutional precedents demonstrate that third-party verification is operable in practice: the ACM TOMS Replicated Computational Results initiative (since 2015), the CODECHECK certificate register, the cascad certification agency (including third-party checking of analyses on restricted data inside a secure computing environment), and publisher-side computational verification workflows. This proposal does not claim novelty in third-party computational verification itself; those programs demonstrate that such verification events are operationally feasible. The additional element proposed here is to test whether those events can be represented as a standardized, machine-readable measurement state — combining a public eligibility profile, an independent recomputation outcome, final-artifact value binding, explicit not-evaluated dimensions, and portfolio-level coverage denominators — designed for funder-level aggregation and integration with existing processes. Whether that additional measurement layer is useful to NIH is a question for the proposed pilot, not an assumption of this response.

As evidence that a substrate of this kind is implementable (offered as feasibility evidence, not as evidence of policy merit): an open record schema, an independently runnable open-source verifier, and public conformance fixtures for a deliberately narrow, explicitly experimental verification contract covering one analysis class (Welch two-group comparisons) are publicly available under an Apache-2.0 license, released by the respondent and fixed by the immutable version identifier given in the attached references, so that this evidence does not change after submission. The current demonstration implements only the structured-record and scoped statistical-recomputation portion of this proposal. It does not implement MVC claim inventories, closed verification surfaces, final-artifact value binding, registry-based coverage measurement, multiple independent verifier implementations, or secure-enclave execution; those are proposed pilot components.

IMPLEMENTATION APPROACHES. Integration should reuse existing process surfaces across the award life cycle rather than create new bureaucracy, and should respect the timing of what exists when. Pre-award, applications can state whether, and for which analysis classes, the project intends to produce verification-eligible artifacts — alongside the rigor and transparency sections and Data Management and Sharing plans that applications already contain; the claim inventory itself concerns completed outputs and cannot exist at application time. Post-award, claim inventories, records, and coverage reporting can attach to annual progress reporting, where NIH already tracks data-sharing implementation, and to publication workflows. Evaluation of any pilot can use the implementation-science frameworks NIH has already designated for Gold Standard Science assessment (CFIR and RE-AIM). NIH could pilot two pathways for the same class of eligible computational analyses. In a legacy pathway — which would require no new authoring environment — a machine-readable record is constructed retrospectively from the analysis artifacts and documentation investigators already produce; this measures the practical ceiling of post-hoc reconstruction, because information that was not captured at execution time cannot reliably be reconstructed later and frequently requires author assistance, as the reproducibility literature documents. In a native pathway, which would separately test participating tools capable of execution-time record capture, tools emit the record at execution time and may generate a declared closed verification surface from the same structured record. Within that declared surface, deterministic generation can structurally preclude drift along the canonical generation path; consistency with content outside the surface is still established only by explicit value binding and is not presumed. Across the two pathways the pilot should compare verification coverage, missing-information rates, author and reviewer burden, final-artifact binding success, observed gaming behavior, and failure modes. Neither pathway should be treated as producing evidence of scientific validity.

PILOT DESIGN. We recommend a voluntary, registration-based pilot in one or two Institutes or Centers with computation-heavy portfolios, with a concurrent comparison group, prospectively sized with prespecified primary outcomes; any inferential comparison should have a justified sample size. The primary question of the pilot should be measurement itself: can MVC be measured reproducibly, at tolerable burden, without unacceptable gaming or field bias? Concretely, the pilot should measure inter-rater agreement on claim inventories and on eligibility classification; concordance between independently developed verifier implementations; false-pass and false-fail rates against a manual expert-audit benchmark; the sensitivity of the coverage vector to claim granularity; burden by career stage and institution type; observed gaming; and discipline differences. The legacy-versus-native comparison described above is the secondary, implementation-level question. The rationale for building evaluation in from the start is empirical: nominally similar incentive interventions have produced divergent results in different contexts — open-data badges were associated with a rise from under 3 percent to roughly 39 percent data sharing in one journal's observational data, while a randomized trial at another journal did not detect an increase, with a highly imprecise estimate that

could not rule out meaningful effects in either direction — so context-dependence should be assumed and measured rather than argued.

POTENTIAL BENEFITS. Execution-side indicators of this kind could add a standardized measurement layer to NIH's developing metric framework where composite measures are still under development; could allow journal reviewers, reproducibility auditors, and programmatic reviewers of completed outputs to spend less attention on arithmetic consistency and more on design and interpretation; could reduce the cost of computational re-execution and of evidence review — whether they affect the cost of experimental replication should be tested rather than assumed; may support long-term preservation of analysis evidence through registry and archival integration; and could give existing mandates (data sharing, rigor criteria) an execution-side outcome measure they currently lack.

UNINTENDED CONSEQUENCES AND ANTI-GAMING DESIGN. We enumerate the principal failure modes and their status honestly, because several are structural rather than fixable. Mitigable by design: surface or claim cherry-picking (mitigated, not eliminated, by the dual denominators and the claim inventory); denominator manipulation via narrow "major claim" definitions (claim inventories confirmed at review); eligibility laundering, i.e., classifying inconvenient claims as ineligible (published eligibility rules plus the all-claims ratio); selective disclosure of passing records (registry with timestamps; unregistered analyses remain undetectable); verifier monoculture and self-verification (third-party execution required; cross-implementation concordance measured); template farming and claim salami-slicing (claim inventories subject to independent review and audit; raw claim counts never themselves used as rewards; judgments of claim granularity and importance remain a human review function); post-publication mutation of reported values (detectable on re-verification of the published artifact, conditional on re-verification actually being run — not structurally prevented). Structurally unavoidable, and therefore boundary rather than bug: semantic overinterpretation in prose outside declared surfaces; analysis-path selection among internally consistent alternatives; analyses that are wrong in design or model specification yet perfectly executed; and internally consistent fabrication, including retrospective record construction. For these, the honest countermeasure is the display and reporting prohibition described above — verification status must never be presented as validity — plus the separate mechanisms science already uses (audit, raw-data requirements, replication). Two further consequences deserve explicit monitoring: discipline and institutional inequity, since computational fields and well-resourced institutions can generate records more cheaply (partially addressed by scope limitation, hosted verification options, and the unsupported state); and checkbox compliance, the documented fate of data availability statements, which argues for pairing any indicator with spot enforcement rather than self-report alone. Consistent with DORA, the Leiden Manifesto, the Metric Tide, and CoARA, these indicators should support qualitative assessment and never substitute for it; the vector reporting format is chosen specifically so that no single number exists to be ranked on.

FEASIBILITY ACROSS CAREER STAGES, DISCIPLINES, AND INSTITUTION TYPES. Scope should begin with narrow computational analysis classes for which a public verification profile can be specified, benchmarked, and independently tested — the setting in which validated automated checks already exist. Restricted and controlled-access data could participate in a pilot through secure-enclave execution: cascad/CASD demonstrates the feasibility of third-party computational checking without public data release, and adapting the machine-verifiable architecture proposed here to such environments would require separate implementation and validation; issuer independence can in principle be preserved even for hosted execution so long as the specification is open and third parties can execute verification. Smaller institutions without local compute could use hosted verification under the same open contract. Petabyte-scale computations and analyses

locked to proprietary software environments should be explicitly out of scope for an initial indicator rather than papered over. The pilot should stratify burden and adoption by career stage rather than assume an advantage or disadvantage in advance.

A PDF version of this response, including full references, is attached.

Topic 2: Data, Software, and Model Sharing

This response extends the framework proposed under Topic 1 (Rigor and Reproducibility) to sharing incentives; the definitions, guarantee boundary, anti-gaming design, and pilot structure described there apply here and are not repeated.

FROM EXISTENCE METRICS TO VERIFIABILITY METRICS. Many scalable automated indicators of sharing measure reporting or existence — a data-availability statement, a repository deposit, a code link — rather than successful reuse or re-execution. Existence can be measured at scale: automated extraction across 2.75 million open-access articles estimated mentions of data-sharing information in 14.5 percent of articles and of code-sharing information in 2.5 percent (Serghiou et al. 2021). But the evidence shows that existence alone can be a weak proxy for the outcomes sharing policy actually seeks: 6.8 percent actual provision among authors who had stated willingness to share (Gabelica et al. 2022); 74 percent initial execution failure among the R files in publicly deposited code-and-data packages (Trisovic et al. 2022); and full unassisted analytic reproduction from original data in only about a third of badge-carrying articles (Hardwicke et al. 2021). Rewarding existence therefore invites the documented checkbox outcome. We propose that NIH pilot whether verified-use and binding states — the full Machine-Verifiable Coverage vector restricted to shared artifacts, interpreted qualitatively — can complement availability and deposit evidence in future sharing incentives. Consistent with Topic 1, this proposes the construct for validation before any connection to rewards; raw counts of bound claims or of deposits, both trivially inflatable, should never themselves be reward-relevant.

SOFTWARE SHARING. NIH-wide research-software sharing is addressed primarily through best-practice guidance rather than a uniform mandatory software-sharing policy. Rather than proposing new mandates, we propose a measurable standard that any future incentive could attach to: software shared with sufficient metadata and environment capture to permit independent re-execution — reusing CodeMeta and FAIR4RS for metadata and interoperability, while a separate verification profile specifies the dependency, environment, and execution requirements that metadata alone does not guarantee — with re-execution success measured rather than assumed. This would convert software sharing from a statement into an event that can be counted, and could provide an execution-side outcome measure for future NIH software-sharing incentives or reporting, without assuming that the Data Management and Sharing Policy itself applies to research software. This initial proposal does not attempt to define a separate indicator for computational-model sharing.

EQUITY ACROSS DATA TYPES. Verification-based sharing measures must not be limited to data that can be posted publicly. Controlled-access and restricted data could participate through verification executed inside secure computing environments, for which an operating precedent exists in third-party checking without public data release (cascad/CASD), with only the verification outcome — not the data — leaving the environment; adapting the proposed machine-verifiable architecture to such environments would require separate implementation and validation. This matters for feasibility across fields: human-subjects research would otherwise be structurally excluded from any verification-linked measure, which would be both inequitable and a perverse incentive against sensitive-data research.

References

1. NIH. NOT-OD-26-087: Request for Information on Measuring and Rewarding Scientific Impact. Released June 18, 2026.

2. Executive Order 14303, Restoring Gold Standard Science (May 23, 2025). Federal Register.

3. OSTP. Guidance for Implementing Gold Standard Science (June 23, 2025).

4. NIH. Leading in Gold Standard Science: Implementation Plan (August 22, 2025).

5. OSTP. Science: A New Golden Age (July 21, 2026).

6. NASEM. Reproducibility and Replicability in Science. National Academies Press, 2019.

7. NASEM. Fostering Integrity in Research. National Academies Press, 2017.

8. NRC. Assessing the Reliability of Complex Models. National Academies Press, 2012.

9. Gabelica M, Boj■i■ R, Puljak L. J Clin Epidemiol 2022. doi:10.1016/j.jclinepi.2022.05.019

10. Trisovic A, Lau MK, Pasquier T, Crosas M. Sci Data 2022. doi:10.1038/s41597-022-01143-6

11. Hardwicke TE, et al. R Soc Open Sci 2021. doi:10.1098/rsos.201494

12. Hardwicke TE, et al. R Soc Open Sci 2018. doi:10.1098/rsos.180448

13. Stodden V, Seiler J, Ma Z. PNAS 2018. doi:10.1073/pnas.1708290115

14. Serghiou S, et al. PLoS Biol 2021. doi:10.1371/journal.pbio.3001107

15. Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM. Behav Res Methods 2016. doi:10.3758/s13428-015-0664-2

16. Nuijten MB, van Assen MALM, Hartgerink CHJ, Epskamp S, Wicherts JM. The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. 2017 (preprint).

17. Brown NJL, Heathers JAJ. Soc Psychol Personal Sci 2017. doi:10.1177/1948550616673876

18. Kidwell MC, et al. PLoS Biol 2016. doi:10.1371/journal.pbio.1002456

19. Rowhani-Farid A, Aldcroft A, Barnett AG. R Soc Open Sci 2020. doi:10.1098/rsos.191818

20. DORA. San Francisco Declaration on Research Assessment, 2013.

21. Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. The Leiden Manifesto. Nature 2015;520:429. doi:10.1038/520429a

22. Wilsdon J, et al. The Metric Tide. UKRI, 2015.

23. CoARA. Agreement on Reforming Research Assessment, 2022.

24. NISO RP-31-2021. Reproducibility Badging and Definitions.

25. Heroux MA. ACM TOMS Replicated Computational Results initiative, 2015.

26. Nüst D, Eglen SJ. CODECHECK. F1000Research 2021.

27. cascad — Certification Agency for Scientific Code and Data (with CASD secure computing environment).

28. W3C. PROV-Overview, 2013.

29. Soiland-Reyes S, et al. RO-Crate. Data Science 2022. doi:10.3233/DS-210053

30. Barker M, et al. FAIR Principles for Research Software (FAIR4RS). Sci Data 2022. doi:10.1038/s41597-022-01710-x

31. Wilkinson MD, et al. FAIR Guiding Principles. Sci Data 2016. doi:10.1038/sdata.2016.18

32. Respondent's open verification demonstration (record schema, open-source verifier, conformance fixtures; Apache-2.0). Repository: https://github.com/licklider-ai/nomue-verifier — tag august-experimental-demo-2026-08-16, commit 59197f8e2ef692c6c8434cdb057be7bf876bc1db; distributable SHA-256 c23ec9d41ba5cf5d79f9eb435671ee563ea438dae04d09b7a86abbd78345e664. As of August 16, 2026.

33. NIH. NOT-OD-21-013: Final NIH Policy for Data Management and Sharing (effective January 25, 2023).

34. NIH. NOT-OD-24-175: Reporting Data Management and Sharing (DMS) Plan Activities in Research Performance Progress Reports (RPPRs) Submitted on or After October 1, 2024.

35. NIH Office of Data Science Strategy. Best Practices for Sharing Research Software (FAQ).

---

175. Endocrine Society

I am responding: On behalf of an organization

Name: Joseph Laakso

Name of Organization: Endocrine Society

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Director, Science Policy

Date Submitted: 8/16/2026

1. Rigor and Reproducibility: Rigor and reproducibility are essential for advancing research and transforming science into discoveries that benefit public health and build trust in science. The National Institutes of Health (NIH) should revise the NIH Data Management and Sharing Policy to create standard operating procedures for the management and sharing of data, guide the review of data and data analysis, and strengthen data integrity practices.

All NIH intramural and extramural researchers supported by training grants or career development awards are required to take Responsible Conduct of Research (RCR) courses. NIH should consider expanding RCR training to all researchers (graduate students, postdocs, staff, PIs) to develop and enhance their foundational knowledge of experimental design, replication, and reproducibility.

A cultural shift is needed in the academic system to incentivize replication studies. The current tenure system rewards scientific advancements. Novel discoveries and related publications are what help protect a Principal Investigator's (PI's) job security and lead to promotions. The "publish or perish" mentality discourages replication studies as the publication of "negative" or non-innovative data is not seen as something that will help advance their research or their career. Incentives and rewards must be built into the grant application system to consider replication in the context of rigor and reproducibility and shift the culture of the tenure system.

Implementing plans to validate previously published scientific data and to generate novel data are two important but separate goals that require different incentives and rewards. To support ongoing and future research efforts, NIH should consider providing PIs with additional financial support for additional materials and other costs associated with replicating their own scientific findings when needed. While the onus is on the PI for ensuring all lab personnel are adhering to best practices in the lab to maximize rigor and reproducibility efforts and avoid scientific misconduct, the NIH should consider providing workshops and resources to help PIs and their institutions with training to maintain consistency across the U.S. Evaluation of plans for addressing rigor, reproducibility, and transparency should be underscored through the grant review process.

To address the reproducibility of previously published fundings, NIH should create a dedicated funding

stream (such as a new R award) to support the materials and personnel needed for the replication of high-impact findings and should coordinate which research findings are prioritized for replication to avoid duplicative efforts. To incentivize researchers to replicate and publish replication studies of previously published studies, NIH should partner with high-impact journals to provide a pathway for publications of replication efforts that help researchers build prestige. Building upon NIH's desire to support and expand data sharing among researchers, NIH could require that all raw data be uploaded for sharing, except when disclosure of such data would potentially identify human subjects. Supplemental data repositories can be cost limiting for journals. NIH should find a way to partner with journals to support these extra costs. It is important to note that replication of clinical trials may be challenging, particularly for studies involving rare diseases or diseases that affect specific populations. There may not be enough patients in the U.S. or worldwide for a new cohort to yield statistical power for a replication study.

2. Data, Software, and Model Sharing: NIH must develop clearly quantifiable metrics to assess the utility and reuse of shared data, software, and models. Study section review is essential for the evaluation of these shared resources to ensure fairness amongst researchers. NIH could consider creating a "citation index" as a metric to quantify the use and reuse of shared scientific data and tools. A combination of the number of downloads, number of article citations, and number of derivative products that were developed from these resources could be used to develop the citation index.

To incentivize the development and validation of new software and models for the advancement of biomedical research, NIH should consider developing a new grant category for methodology development and in silico modeling development. Through this mechanism, projects that have clear plans for producing new technologies, methods, and tools could be prioritized as a way to incentivize these important activities.

3. Training and Mentorship: Training and mentoring graduate students and postdoctoral researchers (postdocs) is essential for the professional development of the next generation of scientists. Recognizing and rewarding effective training and mentorship will foster healthier and more productive research environments that advance medical research.

While mentorship is an essential skill for lab leaders to have, it is not routinely taught. Instead, mentorship skills are often derived from learned experiences, leaving trainee experiences to vary greatly across PIs and institutions. For mentoring to be successful, formal training in mentorship for PIs and trainees is needed to create a baseline mentoring experience to support the PI and all trainees and staff in the lab; the creation and distribution of supportive resources would further such efforts. NIH can use existing mentoring programs as a model for program development, such as the Howard Hughes Medical Institute (HHMI) Gilliam Fellows Program, which incorporates the professional development of the mentor-trainee relationship into its overall training experience. The mentor is required to go through a year-long mentorship development course to help the mentor build skills to create a healthy, supportive lab environment for their trainees' success.

Mentorship should be viewed as a measurable competency. Evidence of effective mentorship could be collected by NIH regularly at intervals throughout the trainee's time with particular attention paid to the experiences of trainees on training grants and fellowships. To develop a measurable metric for

evaluation, the mentor and mentee can create a shared training plan and conduct progress checks to track the progress and execution of the training plan. To protect trainees from potential retaliation for unfavorable evaluations, the information collected should remain confidential or should be aggregated (if there is enough aggregate information to protect anonymity). Should ineffective mentoring be identified during or after the training period, care should be taken to ensure that any consequences do not indirectly harm the trainee (e.g., negative impacts on current or future funding—whether through research project grants, training grants, fellowships, etc.— could materially harm the trainee's progress and prospects). Alternatively, a final exit survey could be conducted to gather feedback about both the mentor and mentee experience.

Quality and effective mentorship is not sufficiently recognized or rewarded through the traditional academic system. PIs do not receive adequate financial support or recognition for the time and effort required to provide excellent mentorship of graduate students and postdocs. To help shift the current academic culture, NIH should consider providing more tangible support for mentors on the K series of awards and training grants, namely providing funding and protected time, as good mentorship requires substantial time and effort. Additionally, non-monetary recognition of mentorship that can be placed on curriculum vitae (CVs) and biosketches could be modeled from NIH intramural awards from ICs such as NIA and NIMH to further showcase a PI's effectiveness in mentoring.

4. Collaboration: Collaboration is important for advancing research in the U.S. and around the world and is traditionally acknowledged through authorship on publications. The ordering of authors on a manuscript often fails to highlight and recognize all critical scientific contributions. NIH should develop a subsequent RFI to identify and utilize a new recognition system to better attribute collaborative efforts and credits to manuscript authors. An existing example of such a recognition system is called Contributor Roles Taxonomy (CRediT). CRediT recognizes 14 contributor roles that are designed to represent the different types of contributions that investigators make for scholarly publications. When developing metrics to assess collaborative contributions, NIH must gather additional input from the scientific community to ensure that the myriad of ways to contribute scientifically are appropriately captured.

5. Entrepreneurship and Translation: If entrepreneurship is to be used as a metric for measuring scientific impact, NIH should note that most scientists are not educated or trained on the topic of entrepreneurship. Educational opportunities and training will be needed to develop foundational knowledge in business development, strategy, fundraising, and technology transfer. In parallel to training, NIH could consider developing a mentoring system to partner investigators who are new to entrepreneurship with those who are experienced in entrepreneurship. NIH could incentivize entrepreneurship through the incorporation of entrepreneurial training for trainees and their mentors in training grant program curriculum.

6. Foundational Scientific Exploration: We identify three considerations that should be addressed to incentivize and reward foundational scientific exploration through high-risk, high-reward initiatives. First, funding for exploratory, high-risk, high-reward projects should not detract from funding for traditional research project grants. Second, training graduate students and postdocs requires additional oversight, support, and protection when working on high-risk, high-reward projects. In this regard, NIH should advise PIs to provide their trainees with a diversified project portfolio. If a high-risk, high-reward project does not work out, trainees should have other projects with less risk to support their continued scientific training. Last, when conducting high-risk, high-reward research, there is a risk that the project will not

produce the anticipated results (i.e. yield negative data). Value and recognition for these worthwhile studies, regardless of the results, should be attributed to the researcher through publication or other means.

It is important to note that identifying and measuring the impact of high-risk, high-reward initiatives can be subjective. Study section review should be maintained as part of the scientific review process to evaluate and prioritize which high-risk, high-reward initiatives should be supported and to avoid penalizing investigators who took on high-risk, high-reward projects that did not yield groundbreaking results.

7. Public Impact: As responsible stewards of U.S. taxpayer dollars, scientists have the responsibility to share their research progress and outcomes with community members and build trust in science through public engagement. The public engagement of science is a learned skill that needs practice and refinement to maximize impact and success. Not all researchers have the opportunity to develop this skillset and put their skills to practice by speaking with non-scientists about their research.

NIH should provide workshops or other methods of science communication and public engagement training for all NIH-supported researchers. Training opportunities will make assessing and rewarding public engagement and impact more feasible and fairer. Otherwise, without such training, scientists could be unfairly penalized through the grant review process for failing to create effective public engagement plans.

The NIH should consider adopting models and practices used by agencies and organizations such as the National Science Foundation (NSF) and Patient-Centered Outcomes Research Institute (PCORI). In addition to publishing a post-study report that includes a summary for non-scientists to read, PCORI requires awardees to describe plans for the dissemination of their work beyond publications in the scientific/medical literature. NSF grant applicants are required to describe the broader impact of their work and explain how their research benefits society. NSF uses five criteria to evaluate broader impacts which includes the potential benefit to society, ability to explore transformative concepts, and feasibility. NSF does not prescribe what activities meet their broader-impacts goals, but some of their examples of desired outcomes such as increasing public scientific literacy, education, development of the STEM workforce, and increasing the economic competitiveness of the U.S. As part of the implementation process to measure public engagement and impact, the NIH should consider providing examples of meaningful engagement that will be recognized while maintaining flexibility to support novel and meaningful activities. Ideas of public engagement opportunities could include partnerships with community organizations, advocacy organizations, or science communicators, activities that specifically engage with patients or community members, or opportunities to shape policy development.

To capture, assess, and recognize researchers' efforts on the public impact of research, the NIH may need to modify the grant application to allow researchers to report their public engagement activities and outcomes. This could be done by modifying the biosketch to include a section on public engagement or by adding an additional section to the grant application itself.

Measuring and assessing the effectiveness and return on investment (ROI) as a means of quantifying scientific impact is subjective. Public impact requires both short-term and long-term monitoring and

evaluation of the target audience. Additional funding and resources are needed to support short-term and long-term monitoring of proposed efforts to more accurately measure and evaluate ROI.

8. Other Areas: N/A

176. Cambridge University Press and University of Cambridge Libraries and Archives

I am responding: On behalf of an organization

Name: Monica Westin

Name of Organization: Cambridge University Press and University of Cambridge Libraries and Archives

Type of Organization: Not Applicable

Role: Other

Date Submitted: 8/17/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Cambridge-Library-Press-Response-to-NIH-RFI.docx

Response to NOT-OD-26-087: Request for Information on Measuring and Rewarding Scientific Impact

For reference: NOT-OD-26-087: Request for Information on Measuring and Rewarding Scientific Impact

Cambridge University Press and University of Cambridge Libraries and Archives strongly support the NIH's effort to realign research incentives with the values of rigor, openness, collaboration, and societal benefit.

Meaningful reform of research assessment requires coordinated action across the research ecosystem, including publishers, institutions, funders, and researchers themselves.

Many of the commendable principles reflected in this Request for Information are already being implemented via initiatives such as the San Francisco Declaration on Research Assessment (DORA) and the Coalition for Advancing Research Assessment (CoARA). Both Cambridge University Press and the University of Cambridge are signatories of DORA. The Press is also a sustaining supporting member of DORA and serves on its Steering Committee, with University of Cambridge Libraries and Archives serving as an alternate member. Across the University of Cambridge, responsible research assessment has already been embedded into institutional practice. The University has also successfully implemented the DORA principles and is progressing changes towards responsible research assessment and open research practices. As a result, the university's recruitment, promotion and reward processes now place substantially less emphasis on publication counts and journal-based metrics.

Cambridge University Press' recent *Publishing Futures* report, a sector-wide review of more than 3,000 stakeholders in the scholarly research ecosystem, highlights the urgent need to reform research-assessment incentives and develop stronger ways to measure research impact, changes that will deliver meaningful benefits for both the scientific community and society at large.

For example, in line with NIH's note that current metrics of scientific success rely heavily on individual-level indicators of productivity, such as publication counts and citation metrics, *Publishing Futures* found that existing academic reward systems worldwide still commonly tie career advancement to the number of published journal articles. Incentivising quantity over quality has resulted in ballooning publication volumes, some of questionable quality to the research ecosystem, which in turn increases the costs of the entire research dissemination system. We know the NIH also shares our concern about these costs.

Fixing these system-wide problems will be challenging but achievable and will depend on rethinking how we measure and reward research so that incentives align with meaningful, high-quality impact.

We therefore strongly support NIH's focus on developing measurable indicators and practical implementation approaches that better recognise rigorous, transparent, and collaborative research.

Our recommendations include:

1. Implement Responsible Research Assessment Principles

DORA provides a well-established framework for improving research assessment across funders, institutions, and publishers. NIH should encourage DORA-aligned assessment practices that:

- evaluate research on its intrinsic quality and contribution rather than the prestige of the publication venue or Journal Impact Factor
- recognise the full range of research outputs, including datasets, software, protocols, and other products that enable scientific progress
- establish transparent and explicit criteria for funding and career advancement
- use quantitative indicators responsibly to support, rather than replace, expert qualitative assessment, while avoiding inappropriate reliance on institutional rankings or journal metrics

These principles are directly aligned with NIH's objective of rewarding scientific rigor, reproducibility, and long-term scientific contribution.

DORA has recently published guidelines for funders that may be particularly useful to NIH colleagues: A Practical Guide to Implementing Responsible Research Assessment at Research Funding Organizations.

2. Reward FAIR Data, software, and model sharing

NIH should explicitly recognise and reward FAIR (Findable, Accessible, Interoperable, and Reusable) research outputs as indicators of scientific impact.

Funding applications and research assessment should consider not only publication plans but also commitments to depositing well-documented datasets, software, computational models, and other research objects in FAIR-compliant repositories.

Both Cambridge University Press and the University of Cambridge actively support FAIR principles. The university’s Research Data Management Service provides institutional support for FAIR data stewardship, while the Press’ open research policies promote these practices that improve the discoverability, accessibility, interoperability, and long-term reuse of research outputs.

Rewarding FAIR-compliant sharing would create stronger incentives for reproducibility, secondary analysis, and cumulative scientific progress.

3. Recognise collaboration and scholarly service as academic labour

Research assessment must better recognise collaborative contributions that are essential to maintaining a healthy research ecosystem but are often invisible within current reward systems.

The *Publishing Futures* report highlighted widespread concern that activities such as peer review, editorial service, mentoring, community leadership, and stewardship of shared research infrastructure receive insufficient recognition despite their importance to research quality.

We encourage NIH to recognise these activities within assessment and reporting frameworks and to encourage institutions to treat peer review, editorial work, and other forms of scholarly service as valued academic research contributions, consistent with emerging practices across the sector.

Conclusion

The reforms proposed in this Request for Information represent an important opportunity to shift research assessment away from quantity-based proxies and toward demonstrable scientific contributions. By adopting responsible assessment frameworks, rewarding FAIR data and software sharing, and valuing collaboration and scholarly service, NIH can accelerate a culture change that is already underway across the global research community.

Such reforms would strengthen scientific rigor, improve reproducibility, encourage openness, and better align incentives with the activities that produce robust, trustworthy, and societally valuable research.

177. Daniel M Merfeld

I am responding: On behalf of myself

Name: Daniel M Merfeld

Name of Organization: Ohio State University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/17/2026

1. Rigor and Reproducibility: Rigor and reproducibility should be recognized as scientific contributions in their own right. Replication studies, independent validation, robustness analyses, and efforts to identify or correct limitations in existing findings may be essential to scientific progress even when they produce fewer publications or citations than novel findings.

A contribution-provenance approach could help make these contributions visible. For example, a replication or validation study could be linked to the research it evaluates and, in turn, to subsequent research that relies on the validated finding. Such relationships would provide evidence of a contribution to the reliability of the scientific record rather than simply counting the resulting publication.

NIH should therefore consider indicators such as independent replication, validation, reproducibility of methods and results, identification and correction of errors, and documented downstream use of validated findings. These measures should supplement rather than replace traditional publication and citation measures.

Importantly, the system should avoid creating incentives to pursue only easily replicable or low-risk research. The objective should be to recognize rigor and reproducibility as additional dimensions of scientific contribution while preserving appropriate incentives for exploratory, foundational, and high-risk research.

2. Data, Software, and Model Sharing: I suggest that NIH consider moving beyond counting whether data, software, and models are shared toward measuring and recognizing how those research products are subsequently used.

A scientist may spend years creating a dataset that enables dozens or hundreds of subsequent studies. Another may develop software that becomes an essential tool for an entire research community. Traditional publication and citation metrics can substantially under represent these contributions.

NIH should therefore recognize not only the creation and sharing of research products, but also their documented reuse. Potential indicators could include the number of independent research groups using a dataset or software tool, resulting publications, use in replication or validation studies, incorporation into subsequent datasets or models, and downstream contributions to clinical or translational research.

There is already substantial infrastructure on which such an approach could build. DataCite provides persistent identifiers and metadata relationships for research outputs and explicitly recognizes data citation as a mechanism for providing credit to researchers. OpenAIRE's research graph connects publications, datasets, software, researchers, projects, and other research products and records relationships among them, including provenance information. Crossref is also developing infrastructure for tracking data citations and research outputs that are reused, cited, and built upon.

This suggests that NIH need not create an entirely new system. Instead, NIH could explore whether existing scholarly infrastructure, together with NIH's own iCite system, could support a scientific contribution-provenance approach.

For example:

Dataset A → Paper B → Paper C → Review D → Clinical application E

The objective would not be to arbitrarily transfer a fraction of every downstream citation back to the creator of Dataset A. Rather, the system could make the downstream pathway visible as evidence in the researcher's assessment portfolio:

"This dataset was directly reused in 37 publications by 24 independent research groups, including studies that subsequently contributed to additional research outputs and translational activities."

This distinction is important. I would favor documenting downstream provenance rather than creating another single numerical impact score.

A portfolio-based system could retain traditional measures such as publications and citations while adding evidence of the creation, reuse, and downstream contribution of datasets, software, models, methods, and protocols.

NIH could initially pilot such an approach in several fields and career stages, compare it with conventional assessment, and evaluate both its usefulness and potential unintended consequences. The goal would be to determine whether documenting the pathway from research contribution to subsequent scientific use provides evaluators with meaningful information that is currently missing.

The broader principle is:

Scientific assessment should ask not only what a researcher produced, but how those contributions were subsequently used and what scientific activity they helped enable.

3. Training and Mentorship: N/A

4. Collaboration: Current research assessment often assigns credit primarily through publications, authorship position, citations, and grant leadership. These measures can under represent substantive contributions made by members of large or interdisciplinary teams.

A contribution-provenance approach could provide an additional way to recognize team science.

Rather than treating a publication or research project as a single unit of credit, research outputs could be connected to specific contributions using persistent identifiers and structured contributor information. Those contributions could include data generation and curation, software development, methodology, models, protocols, analysis, infrastructure, and other substantive roles.

The important next step would be to preserve evidence of what happens to those contributions downstream.

For example, if one member of a team creates a dataset, another develops software to analyze it, and a third performs the analysis that becomes the basis for subsequent studies, the research record could preserve those relationships. The resulting portfolio would show not only authorship but the contribution pathway.

This should not become a formula for mechanically dividing citations among coauthors. Rather, it should provide evaluators with additional evidence about the nature, extent, and subsequent use of an individual's contribution.

Existing infrastructures such as CRediT, persistent identifiers, DataCite, Crossref, and OpenAIRE provide components of such a system. NIH could explore integrating or extending these approaches and piloting contribution-provenance records in team-based research.

This would be particularly valuable for researchers whose major contributions are difficult to capture through first-author or senior-author counts, including data scientists, software developers, methodologists, infrastructure builders, and other members of interdisciplinary teams.

The objective should be to recognize substantive contribution rather than simply position within an authorship hierarchy.

5. Entrepreneurship and Translation: Patents and other forms of innovation should be included in scientific assessment portfolios. A patent, however, should not be treated as equivalent to a successfully translated technology. Patent filing, patent issuance, licensing, commercialization, clinical development, and demonstrated health impact represent different stages of a contribution pathway.

NIH's own technology-transfer activities illustrate the value of following such pathways. NIH has documented how licensed technologies can generate subsequent inventions, research and development, clinical trials, and ultimately health and economic benefits.

The same principle could be applied more broadly to scientific research: assessment should preserve evidence of the pathway from an original contribution through subsequent research and, where applicable, translation.

I would also recommend retaining established measures such as publications, citations, and grants rather

than attempting to replace them with a new metric. These measures have limitations, but they also have intrinsic value and provide useful evidence of scientific activity and influence. Completely discarding them would be particularly problematic for investigators whose career development was guided by the existing system and its rules.

The better approach is a portfolio of evidence.

Traditional measures could be supplemented by evidence concerning data and software reuse, replication, mentorship, collaboration, patents, translation, and downstream scientific utilization.

This is particularly important for foundational and high-risk research. Scientific value may not become apparent for years or decades, and an assessment system that rewards only immediate citations, applications, or public attention could systematically disadvantage research whose eventual contribution is substantial but delayed.

I therefore recommend against a single composite "scientific impact score." Different forms of scientific contribution are not interchangeable, and a new metric would inevitably become a new target for optimization.

Instead, NIH could develop a multidimensional portfolio in which traditional metrics remain visible alongside evidence of additional contributions and their subsequent use.

The guiding question should be:
What did this scientist contribute, how was that contribution used, and what subsequent scientific activity did it help enable?

That approach would broaden recognition without pretending that the diverse forms of scientific value can be reduced to one number.

6. Foundational Scientific Exploration: Scientific assessment should preserve and recognize foundational and exploratory research even when its eventual impact is not immediately apparent. Scientific value may emerge years or decades after an initial discovery, and foundational work may enable subsequent research in ways that are difficult to anticipate when the original work is evaluated.

This is another reason to avoid replacing existing measures with a single "scientific impact score." A metric based primarily on immediate citations, applications, commercialization, or public attention could systematically disadvantage high-risk, exploratory, or foundational research whose importance becomes apparent only over time.

Instead, NIH should consider a multidimensional portfolio in which traditional measures such as publications, citations, and grants remain visible alongside evidence of other forms of contribution, including subsequent reuse of datasets, methods, software, models, and findings.

The emerging infrastructure for linking research outputs also creates an opportunity to document these longer-term contribution pathways. A foundational discovery could be linked to subsequent publications,

methods, datasets, technologies, or clinical applications that build upon it. Such provenance would provide evidence of downstream scientific contribution without requiring NIH to predict in advance which research will ultimately prove important.

The goal should therefore not be to determine a single numerical value for every scientific contribution, but to provide evaluators with a richer body of evidence about what was contributed, how it was subsequently used, and what scientific activity it helped enable.

This approach would preserve incentives for both foundational exploration and practical application while recognizing that the time horizon and pathway to impact can differ dramatically across fields and individual research programs.

7. Public Impact: N/A

8. Other Areas: I encourage NIH to consider the overall structure of scientific assessment as an additional area for consideration. In my view, the scientific community is unlikely to reach agreement on a single metric of scientific impact, and attempting to create one would likely be counterproductive.

A better approach would be a portfolio of complementary measures. Established indicators such as publications, citations, and grants should be retained because they provide useful evidence of scientific activity and influence. They should be supplemented by measures of contributions that are currently less visible, including datasets, software, methods, replication, collaboration, mentorship, patents, translation, and downstream reuse.

The objective should be to provide evaluators with a richer body of evidence rather than to generate a new composite score. Different investigators, disciplines, career stages, and types of research will appropriately have different patterns of contribution.

The emerging ability to link research products through persistent identifiers and scholarly knowledge graphs creates an opportunity to make these contributions—and their subsequent use—more visible. NIH could help develop and test such a portfolio-based approach while studying its feasibility and unintended consequences.

In short, I would encourage NIH to move from asking "What is this scientist's impact score?" toward asking "What did this scientist contribute, how was it used, and what subsequent scientific activity did it help enable?"

178. NCI Office of Data Sharing and Trans-NCI DMS Working Group National Cancer Institute

I am responding: On behalf of an organization

Name: Mousumi Ghosh, Ph.D.

Name of Organization: NCI Office of Data Sharing and Trans-NCI DMS Working Group National Cancer Institute

Type of Organization: Government Agency

Role: Government Official

Date Submitted: 8/17/2026

1. Rigor and Reproducibility: Reproducibility and replicability are foundational elements of rigorous biomedical research. As biomedical technologies become increasingly sophisticated and data-intensive, ensuring that experimental findings can be reproduced within and across institutions requires standardized methodologies, comprehensive documentation, transparent reporting, robust data stewardship including provenance tracking for shared data, and effective incentive mechanisms. All aspects of scientific data, in this context, are the key to unlocking the ground truths of many studies. In that regard, shared resource and core facilities are uniquely positioned to advance these goals because they operate at the intersection of technology, experimental design, data generation and management, and investigator training. Unlike individual laboratories, core facilities routinely implement standardized operating procedures, instrument calibration and maintenance programs, quality assurance measures, and metadata capture workflows that reduce technical variability and support consistent data generation. Core facilities maintain records of instrument settings, analytical pipelines, reagent lots, quality control metrics, and sample processing procedures, providing the detailed provenance information necessary to reproduce experiments. Core facilities frequently generate data for PIs; tagging it with accurate metadata and formatting it to community standards ensures FAIR compliance when sharing. This process also saves PIs time and effort for publications or DMS policy requirements.

Establishing institutional mechanisms and infrastructure that allow datasets, workflows, and analytical pipelines generated in core facilities to be systematically archived and reused would create opportunities for independent validation studies and cross-institutional replication efforts. Additionally, core facilities can be foundational for inter-institutional proficiency testing, developing reference standards, and replication studies. Funding to expand this inter-institutional activity would strengthen confidence in experimental findings, accelerate adoption of robust methodologies and streamline the data management process.

Developing Incentives to Reward Replication Science. Incentives for rigor and reproducibility must start with data generation. Institutions and funding agencies should recognize and reward activities such as data curation, standards development, protocol validation, quality assurance, and replication studies. Creation of widely accepted metrics that acknowledge the generation of reusable datasets, adherence to

community standards, successful independent validation of findings, and reuse/ impact of shared data would encourage broader participation in reproducibility efforts. In addition, researchers who use and reuse data to replicate/validate previous findings as well as reveal novel scientific discovery should be given credits and incentives to do so where scientific data is one of the key research products along with publications, software, patents, or clinical trials. Recently, the NEI-led S-Index Challenge called on the broad community to develop data sharing metrics to incentivize researchers for rigor and reproducibility studies. Other mechanisms of incentives are cash prize awards, such as NINDS's Replication Prize, and community-driven data use events to democratize scientific data for validation and research findings such as NCI's Data Jamboree and other hackathons would be very helpful.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

179. Federation of American Societies for Experimental Biology (FASEB)

I am responding: On behalf of an organization

Name: Crytal J. Neely

Name of Organization: Federation of American Societies for Experimental Biology (FASEB)

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/17/2026

1. Rigor and Reproducibility: The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to respond to the National Institutes of Health (NIH) Request for Information (NOT-OD-26-087), Measuring and Rewarding Scientific Impact. Representing 21 American scientific societies and more than 100,000 researchers across the biological and biomedical sciences, FASEB commends NIH's efforts to develop approaches that more fully recognize the full range of contributions that advance biological and biomedical research while maintaining the highest standards of scientific rigor and public trust.

FASEB's member societies steward critical components of the biological and biomedical research enterprise by publishing peer-reviewed journals, convening scientific meetings, and providing mentoring and career development programs that help ensure federally funded discoveries are rigorously evaluated, disseminated, and built upon by the broader research community, including industry. FASEB approaches this RFI both as a representative of individual investigators and as a steward of the infrastructure that translates NIH-funded research into scientific and public value, consistent with the principles of Executive Order 14303, Restoring Gold Standard Science (May 23, 2025). FASEB believes that rigorous, expert peer review remains the cornerstone of evaluating scientific quality, maintaining the credibility of science, and ensuring responsible stewardship of federal research investments.

NIH's reporting and evaluation frameworks represent only one component of the broader research assessment ecosystem. Many meaningful indicators of scientific quality, rigor, and impact already exist but are not consistently recognized or incorporated into evaluation processes. These indicators should complement—not replace—expert scientific peer review, which remains the most effective mechanism for evaluating rigor, originality, significance, and scientific quality in both funding decisions and scholarly publication. Peer-reviewed publications remain an essential indicator of scientific productivity; however, publication counts alone do not capture the full value of a scientist's contributions. Notably, even journals evaluate their own performance using a portfolio of complementary metrics rather than publication volume alone.

FASEB encourages NIH to partner with scientific societies, research institutions, and other stakeholders to modernize research assessment in ways that are evidence-based, practical, and appropriate across disciplines, career stages, institution types, and resource levels. FASEB stands ready to support NIH by identifying meaningful scientific impact indicators and assessment approaches, evaluating

implementation strategies, and helping minimize unnecessary administrative burden, particularly for smaller and less-resourced institutions.

Topic 1: Rigor and Reproducibility
FASEB recommends NIH:
Establish clear definitions of reproducibility and causal research before incorporating them into scientific impact assessment. NIH should clearly define these concepts and communicate consistent expectations before incorporating them into evaluation or reporting. Scientific societies can help ensure these definitions are interpreted consistently across disciplines and implemented equitably.

Complement traditional productivity measures with evidence of rigorous scientific practice. Where measurable indicators are appropriate, NIH should emphasize established practices that support rigor, including authentication of key biological and chemical resources, adherence to community reporting standards, appropriate preregistration of confirmatory studies, and successful independent validation. These indicators provide an additional layer of value and understanding of the incentive structure at institutions.

Integrate transparency for generative AI into existing reproducibility policies. As AI becomes increasingly integrated into biological and biomedical research, NIH should incorporate documentation of AI tools, such as the model, provider, and version where appropriate, within existing authentication and reporting requirements rather than creating separate reporting systems. Building on NIH's existing Enhancing Reproducibility through Rigor and Transparency guidance (NIH, 2025), this approach will promote transparency, support computational reproducibility, and minimize additional administrative burden while allowing expectations to evolve alongside advances in AI.

Invest in infrastructure that enables rigorous science. Reproducibility depends on sustained access to shared research resources, core facilities, modern instrumentation, and skilled technical staff. NIH should ensure that investments in reproducibility are matched by investments in the infrastructure needed to support them; otherwise, expectations for rigor may outpace institutional capacity.

Recognize replication as a valuable scientific contribution. Independent replication strengthens confidence in scientific findings but remains underrecognized within traditional systems of academic reward. NIH should recognize replication studies as distinct scholarly contributions within existing reporting mechanisms and continue developing funding opportunities that support high-priority replication efforts. To foster this commitment to replication, today, many societies have at least one journal in their portfolio that allows for recognition of and publication of replication studies, and several societies have supported the community further by dedicating special issues to replication studies.

Support scientific societies as critical to the research ecosystem as they seek to advance reproducibility. Biological and biomedical scientific societies are uniquely positioned to develop discipline-specific standards, identify high-priority opportunities for replication, disseminate best practices, and provide training in rigorous and reproducible research methods. NIH should continue to support society-led initiatives, including the development of training resources and the integration of rigor and reproducibility modules into programs serving researchers across career stages, including predoctoral

trainees, postdoctoral researchers, and early-career investigators.

Scientific impact cannot be meaningfully assessed without confidence in the quality and reliability of the underlying science. FASEB encourages NIH to build its scientific impact framework on the foundation of rigorous, expert peer review, which remains the cornerstone of research assessment and the credibility of science. Responsible indicators of rigor should supplement, not substitute for, expert scientific judgment and should be implemented in partnership with the scientific community, consistent with the principles of Executive Order 14303, Restoring Gold Standard Science.

2. Data, Software, and Model Sharing: Peer-reviewed publications remain the primary scholarly output of biological and biomedical research and an essential means of communicating scientific discoveries. The value of biological and biomedical research, however, also extends to data, software, computational models, and other enduring research outputs that enable validation, reuse, and new lines of investigation. NIH has made considerable progress through its Data Management and Sharing Policy (NIH, 2023). Building on this progress, NIH should recognize and reward the creation, stewardship, and sharing of high-quality research resources to maximize the return on the nation's investment in biological and biomedical research.

FASEB recommends NIH:
Recognize data, software, and computational models as additional scholarly products alongside publications. These outputs should be evaluated in addition to publications in funding decisions and progress reporting. NIH should continue incorporating these contributions into existing investigator reporting mechanisms, including the Biographical Sketch Common Form and Contributions to Science, so that researchers receive appropriate credit for developing resources that enable discovery across the biological and biomedical enterprise.

Reward quality, reuse, and interoperability. The value of shared research resources lies in their accessibility, reliability, interoperability, and scientific utility, not simply whether they are deposited. Assessment systems that prioritize quantity over quality, usability, or scientific value risk encouraging compliance-driven behavior without meaningfully advancing research. Where measurable indicators are appropriate, NIH should emphasize adherence to the FAIR Principles (Wilkinson et al., 2016), documented reuse, appropriate citation, interoperability, and downstream scientific impact. Consistent with FASEB's longstanding position on data management and access, incentives should encourage the sharing of high-value research resources while avoiding requirements that create unnecessary administrative burden with limited scientific benefit.

Recognize stewardship as a valuable contribution. Long-term stewardship of datasets, software, repositories, and other shared research resources requires sustained scientific expertise. NIH should recognize activities such as curation, quality control, documentation, and long-term maintenance as distinct, valuable contributions within investigator assessment. Recognizing stewardship will encourage sustained investment in high-quality research resources that remain reliable, reusable, and valuable over time.

Invest in the infrastructure that enables responsible sharing and long-term stewardship of research resources. Expectations for sharing data, software, biospecimens, model organisms, reagents, and other

valuable research resources should be matched by sustained investments in repositories, biobanks, research resource centers, core infrastructure, cloud resources, metadata standards, and the skilled technical workforce required to curate and maintain them. Without this support, valuable scientific resources may become inaccessible, limiting reproducibility, reducing reuse, and diminishing the long-term return on NIH's investment. Long-term stewardship of these resources should not depend solely on individual research grants.

Promote transparency and reproducibility for computational models. As artificial intelligence and computational modeling become increasingly integrated into biological and biomedical research, NIH should encourage transparent documentation of model provenance, versioning, computational workflows, and other information needed to support reproducibility, as appropriate. Building these expectations into existing reporting practices will improve transparency, strengthen confidence in AI-enabled research, and allow guidance to evolve alongside advances in computational science.

Providing funding for, and rewarding, the creation and stewardship of the comprehensive knowledge ecosystem, including peer-reviewed publications, data, software, and computational models, is essential to maximizing the return on the nation's investment in research.

3. Training and Mentorship: NIH's mission depends on a strong, sustainable biological and biomedical research workforce, and mentorship is among the most important contributions scientists make to that workforce. Yet mentorship remains largely invisible in traditional measures of scientific productivity and impact. As a federation of scientific societies, FASEB has extensive experience advancing research skill development, mentorship, and workforce development through NIH-funded programs, scientific society initiatives, and evidence-based policy recommendations. Drawing on this experience, FASEB encourages NIH to recognize mentorship as a valuable contribution, reward high-quality training environments, and adopt assessment approaches that reflect the various career pathways essential to the biological and biomedical research enterprise.

FASEB recommends NIH:
Recognize mentorship as a valuable contribution. Effective mentorship should be recognized alongside scientific achievement in NIH assessment and reporting. Building on recommendations from the NIH Advisory Committee to the Director Working Group on Re-envisioning NIH-Supported Postdoctoral Training (NIH ACD, 2023), NIH should incorporate structured documentation of mentoring activities, mentor development, and scholar outcomes into relevant existing reporting mechanisms. Assessment should emphasize mentoring quality and scholar success rather than simply the number of scholars supervised, while recognizing participation in evidence-based mentor training as an indicator of scientific leadership.

Leverage existing NIH training programs rather than creating new assessment systems. Institutional training grants, particularly T32 programs, represent NIH's most established framework for evaluating research training because they already incorporate structured training plans, mentor oversight, career development activities, and longitudinal scholar outcome reporting. NIH should build on these established mechanisms to identify effective mentoring practices, evaluate training environments, and disseminate evidence-based approaches while minimizing additional administrative burden.

Evaluate training through meaningful scholar outcomes that recognize various career pathways. NIH should recognize successful career outcomes across academia, industry, government, nonprofit organizations, education, and other sectors. Building on existing trainee outcome reporting and standardized career taxonomies such as Unified Career Outcomes Taxonomy (UCOT; Stayart et al., 2020), indicators including degree completion, career transitions, K99-to-R00 transition, K-to-R conversion, and time to research independence can provide meaningful measures of program effectiveness when interpreted within program context.

Recognize competency-based development as an indicator of quality. Effective research training develops competencies in scientific rigor, leadership, communication, professionalism, responsible conduct of research, and career development. NIH should encourage competency-based assessment that complements traditional research outputs while remaining flexible across career stages and career pathways.

Support mentoring networks that reflect today's biological and biomedical workforce. Scientists increasingly benefit from mentoring networks that include research, career, peer, and cross-sector mentors. NIH should recognize and encourage distributed mentoring models that strengthen career development, broaden access to expertise, and better reflect the collaborative nature of today's biological and biomedical research enterprise.

Recognize the complementary role of scientific societies in workforce development. Scientific societies provide mentoring, leadership development, networking, career exploration, and professional communities that complement institutional training. These activities expand access to expertise, strengthen professional networks, and complement institutional training across career stages. NIH should continue partnering with scientific societies to extend the reach and impact of its workforce development investments.

Mentorship is among the most consequential—and least recognized—contributions scientists make to the biological and biomedical research enterprise. Assessment approaches should support long-term program improvement rather than short-term funding decisions and account for institutional context, discipline, career stage, and the time required for meaningful career outcomes to emerge. By recognizing mentoring excellence and high-quality training environments, NIH can strengthen the biological and biomedical workforce and, in turn, the nation's innovation economy.

4. Collaboration: Collaboration has become a defining feature of modern biological and biomedical research, with discoveries increasingly emerging from multidisciplinary teams that span institutions, sectors, and scientific disciplines. Through its member societies, scientific meetings, and collaborative initiatives, FASEB has long supported team science and interdisciplinary collaboration. As NIH develops approaches for recognizing and assessing scientific impact, assessment should recognize the range of scientific contributions that enable collaborative research while remaining appropriate across disciplines, career stages, and research settings.

FASEB recommends NIH:

Recognize the full range of scientific contributions that enable team science. Scientific impact should not be inferred solely from a singular metric; a matrix approach is recommended. Rigorous peer review should continue evaluating the quality of collaborative science, while structured contribution taxonomies such as the Contributor Roles Taxonomy (CRediT) developed by the National Information Standards Organization (NISO) provide additional context regarding individual contributions. NIH should incorporate these taxonomies into existing reporting mechanisms to provide additional context regarding individual contributions.

Evaluate collaboration based on scientific contributions rather than team size. Assessment should recognize interdisciplinary integration, shared leadership, effective coordination among collaborators, and the scientific advances enabled by collaboration rather than relying on coauthor counts or other quantitative measures.

Recognize cross-sector collaboration as an indicator of valuable impact. Biological and biomedical discovery increasingly depends on partnerships among academia, industry, government, healthcare systems, nonprofit organizations, patient communities, and scientific societies. NIH should recognize collaborations that meaningfully advance scientific discovery, translation, workforce development, or public health.

Recognize collaborative leadership within NIH assessment and reporting. NIH should continue recognizing multiple principal investigators, co-investigators, and other collaborative leadership roles within its funding and research assessment processes, reinforcing institutional recognition of team-based science.

Ensure collaborative science supports early-career investigators. Assessment approaches should enable students, postdoctoral scholars, and early-stage investigators to clearly document their individual scientific contributions within collaborative projects. Participation in team science should provide opportunities to demonstrate scientific independence rather than obscure individual contributions.

Avoid metrics that create unintended incentives. Measures based solely on coauthor counts, collaboration team size, or other quantitative proxies are poor indicators of meaningful collaboration and may encourage honorary authorship or unnecessary expansion of research teams. Collaboration should instead be evaluated based on documented scientific contributions, the quality and impact of the collaborative work, and disciplinary norms.

Scientific progress increasingly depends on effective collaboration across disciplines, institutions, and sectors. By recognizing meaningful contributions to team science, NIH can strengthen collaborative innovation while ensuring assessment remains equitable, transparent, and reflective of modern biological and biomedical research.

5. Entrepreneurship and Translation: Biological and biomedical research creates value by generating new knowledge and translating discoveries into improved health. Successful translation depends on sustained investment in investigator-initiated foundational research together with effective cross-sector partnerships. Through their peer-reviewed journals, scientific meetings, and professional networks, FASEB's member societies help advance discoveries from the laboratory to practical application. As NIH

develops approaches for recognizing scientific impact, entrepreneurial and translational activities should be recognized as important scientific contributions without diminishing the vital importance of foundational discovery research.

FASEB recommends NIH:
Recognize translational research as a valuable scientific contribution. NIH should recognize contributions that advance discoveries toward clinical practice, public health, diagnostics, therapeutics, medical technologies, and other real-world applications. Assessment should recognize progress across the translational continuum, including advancement through the research and development pipeline and the maturation of early-stage discoveries, rather than focusing solely on commercial endpoints.

Recognize entrepreneurial and innovation activities that responsibly advance biological and biomedical discoveries. Activities such as technology development, invention disclosure, patenting, licensing, startup formation, and regulatory advancement can represent meaningful scientific contributions when they advance the translation of NIH-supported discoveries into products, services, and technologies that improve health. Recognizing these activities within existing reporting mechanisms will encourage investigators to pursue translational research while preserving support for foundational discovery. Because commercialization infrastructure, including technology transfer offices, industry partnerships, and access to capital, is unevenly distributed, NIH should avoid indicators that disadvantage investigators at smaller or less-resourced institutions or reward patenting activity without evidence of meaningful translational or public benefit.

Recognize the role of the Bayh-Dole framework in moving federally funded discoveries toward public benefit. For more than four decades, the Bayh-Dole Act has enabled discoveries made with federal support to advance toward application through invention disclosure, patenting, and licensing, catalyzing the development of new drugs, diagnostics, and medical technologies. NIH's assessment framework should reinforce the successful incentives established under the Bayh-Dole Act by recognizing technology transfer, patenting, licensing, and related translational activities as legitimate scientific contributions while continuing to value the foundational research on which these activities depend.

Support cross-sector partnerships that accelerate bench-to-bedside translation. Successful translation depends on collaboration among academic institutions, industry, government agencies, healthcare systems, patient organizations, philanthropies, and scientific societies. NIH should recognize meaningful partnerships that move discoveries toward application and improve public health, while ensuring that individual investigator contributions within these partnerships remain visible in assessment.

Evaluate translation using meaningful evidence of scientific advancement. Assessment should emphasize evidence that discoveries are progressing toward application, including adoption of new methods or technologies, advancement through the development pipeline, and incorporation into clinical practice or public health. Quantitative measures should be interpreted within the context of discipline, career stage, and the long timelines, often a decade or more for new therapeutics, that translation frequently requires, so that early-career researchers are not disadvantaged when translational outcomes fall outside their typical evaluation windows.

Translation is one of the principal ways biological and biomedical research delivers benefits to patients and society, but it depends on sustained investments in rigorous, peer-reviewed basic research. By recognizing entrepreneurship and translation as complementary dimensions of scientific impact and continuing to support the Bayh-Dole framework, NIH can strengthen the continuum from discovery to application while preserving incentives for foundational science.

6. Foundational Scientific Exploration: Foundational scientific exploration is the wellspring of biological and biomedical progress. The discoveries that ultimately transform human health frequently originate in curiosity-driven investigation whose applications cannot be foreseen at the outset and may take years or decades to emerge. Sustained progress in translational and clinical research depends on a continuous supply of the new knowledge, mechanisms, and targets that fundamental research generates. As NIH develops approaches to measuring scientific impact, FASEB encourages the agency to also recognize, protect, and reward foundational biological and biomedical exploration, recognizing that assessment systems built solely around near-term or easily quantified outcomes risk undervaluing the basic science on which future advances depend.

FASEB recommends NIH:
Protect foundational research from assessment systems that favor near-term outcomes. Because the impact of basic biological and biomedical science often emerges years or decades later through unpredictable pathways, NIH should adopt approaches to evaluation that account for the long and uncertain timelines of discovery research.

Recognize the advancement of fundamental knowledge as a core indicator of scientific impact. Expanding our understanding of biological systems, including fundamental cell, molecular, and organismal biology and the discovery of new biological concepts and therapeutic targets, is a foundational contribution to the biological and biomedical research enterprise, independent of near-term application. NIH should recognize contributions that advance fundamental knowledge, generate new concepts and tools, or open new lines of inquiry as meaningful indicators of impact within existing evaluation and reporting mechanisms.

Support high-risk, high-reward research through appropriate assessment. Transformative discovery often depends on investigations that carry significant uncertainty and may not yield the anticipated result. NIH should ensure that assessment approaches do not penalize scientifically rigorous research that does not produce the anticipated findings and should recognize well-executed negative or null results as meaningful contributions to scientific knowledge.

Account for long and unpredictable timelines in evaluating foundational research. Measures tied to short reporting windows cannot capture the trajectory of basic science. NIH should favor evaluation approaches that accommodate extended timeframes, avoid drawing conclusions about impact prematurely, and allow the significance of foundational biological and biomedical discoveries to emerge over time. Because foundational discovery is inherently unpredictable, impact assessment frameworks should continue to support a diverse portfolio of investigator-initiated, curiosity-driven research rather than directing resources primarily toward predefined priorities or readily measurable near-term outcomes.

Foundational biological and biomedical science exploration is the source of discoveries that ultimately transform health, yet its value is the hardest to capture through conventional measures of impact. FASEB urges NIH to design assessment approaches that recognize, protect, and reward curiosity-driven science, thereby safeguarding the discovery research from which tomorrow's advances will arise.

7. Public Impact: Biological and biomedical research generates value through publications, datasets, software, computational models, and other enduring scientific outputs. It also delivers value in improving health, strengthening the economy, informing clinical practice, building public trust, and benefiting society in ways that often emerge over many years. As NIH develops approaches to assess public impact, FASEB encourages the agency to recognize the many pathways through which research benefits the public. Public impact is multidimensional and should be assessed through a matrix of complementary indicators that reflect NIH's mission.

FASEB recommends NIH:
Recognize improvements in health and clinical practice as core indicators of public impact. NIH should also recognize contributions that improve health outcomes, inform clinical care, support implementation of evidence-based interventions, and broaden access to effective care, including the responsible return of research results where appropriate.

Incorporate validated measures of economic and societal benefit. NIH-supported research drives innovation, economic growth ($2.57 for each $1.00 invested), workforce development, and national competitiveness. Where appropriate, NIH should also incorporate validated measures of economic impact alongside other metrics to better communicate the broad public return on federal investment in biological and biomedical research.

Recognize public engagement and trustworthy science communication as scientific contributions. Building and maintaining public trust requires sustained engagement with patients, communities, educators, policymakers, and the public. NIH should also recognize meaningful contributions to science communication, patient and community engagement, transparency, and other activities that strengthen confidence in biological and biomedical research. Scientific societies are important partners in developing and disseminating evidence-based communication practices and fostering public trust.

Avoid simplistic measures of public impact. No single metric can adequately capture the many ways biological and biomedical research benefits society, and a matrix of indicators is warranted. Likewise, economic measures should complement, not replace, assessment of health outcomes and broader public benefit. NIH should instead adopt a balanced framework that recognizes health outcomes, clinical practice, economic growth, accessibility, public trust, and societal benefit as complementary dimensions of impact.

Recognize public impact within existing application, reporting, and research assessment processes. Investigators should be able to document meaningful public impacts, including contributions to clinical guidelines, public policy, accessibility, community engagement, the responsible return of research results, and public understanding of science. Narrative descriptions should complement quantitative indicators by capturing important additional impacts.

Partner with scientific societies to strengthen public engagement and scientific literacy. Scientific societies play an essential role in communicating biological and biomedical scientific advances, convening experts, engaging communities across the country, and developing trusted educational resources. NIH should continue partnering with scientific societies to strengthen public understanding of biological and biomedical research and ensure that measures of public impact reflect the range of disciplines and communities they serve.

Public impact is the ultimate expression of NIH's mission. FASEB encourages NIH to adopt assessment approaches that recognize the various ways biological and biomedical research improves lives by valuing publications alongside datasets, software, computational models, and other scientific contributions, while remaining flexible, evidence-based, and appropriate across scientific disciplines and stages of research.

In conclusion, FASEB supports NIH's efforts to develop a more comprehensive approach to measuring and rewarding scientific impact that recognizes the full range of contributions advancing biological and biomedical research. As NIH develops this framework, we encourage the agency to preserve rigorous, expert peer review as the foundation for evaluating scientific quality while adopting responsible, evidence-based indicators that complement—not replace—scientific judgment.

FASEB urges NIH to adopt the recommendations outlined above to ensure that scientific impact is assessed in ways that strengthen rigor, encourage innovation, recognize the range of contributions across the biological and biomedical research enterprise, and minimize unnecessary administrative burden. FASEB and its member societies welcome the opportunity to provide additional information and to work with NIH to develop and implement an assessment framework that advances scientific excellence while maximizing public benefit.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/FASEB_Response-to-NIH-RFI-on-Measuring-Scientific-Impact.pdf

Description: FASEB's Response Letter to NIH's RFI on Measuring and Rewarding Scientific Impact

FASEB
FEDERATION OF AMERICAN SOCIETIES FOR EXPERIMENTAL BIOLOGY

REPRESENTING OVER 100,000 RESEARCHERS
6120 Executive Boulevard | Suite 230 | Rockville, MD 20852
faseb.org

August 17, 2026

Submitted electronically via the [Comment Form: Measuring and Rewarding Scientific Impact - Office of Science Policy]

The Honorable Jayanta Bhattacharya, M.D., Ph.D.
Director, National Institutes of Health
9000 Rockville Pike
Bethesda, MD 20892

Re: Comments of the Federation of American Societies for Experimental Biology (FASEB) on the National Institutes of Health Request for Information (NOT-OD-26-087), *Measuring and Rewarding Scientific Impact*

Dear Director Bhattacharya:

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to respond to the National Institutes of Health (NIH) Request for Information (NOT-OD-26-087), M*easuring and Rewarding Scientific Impact*. Representing 21 American scientific societies and more than 100,000 researchers across the biological and biomedical sciences, FASEB commends NIH's efforts to develop approaches that more fully recognize the full range of contributions that advance biological and biomedical research while maintaining the highest standards of scientific rigor and public trust.

FASEB's member societies steward critical components of the biological and biomedical research enterprise by publishing peer-reviewed journals, convening scientific meetings, and providing mentoring and career development programs that help ensure federally funded discoveries are rigorously evaluated, disseminated, and built upon by the broader research community, including industry. FASEB approaches this RFI both as a representative of individual investigators and as a steward of the infrastructure that translates NIH-funded research into scientific and public value, consistent with the principles of Executive Order 14303, *Restoring Gold Standard Science* (May 23, 2025). FASEB believes that rigorous, expert peer review remains the cornerstone of evaluating scientific quality, maintaining the credibility of science, and ensuring responsible stewardship of federal research investments.

NIH's reporting and evaluation frameworks represent only one component of the broader research assessment ecosystem. Many meaningful indicators of scientific quality, rigor, and impact already exist but are not consistently recognized or incorporated into evaluation processes. These indicators should complement—not replace—expert scientific peer review, which remains the most effective mechanism for evaluating rigor, originality, significance, and

scientific quality in both funding decisions and scholarly publication. Peer-reviewed publications remain an essential indicator of scientific productivity; however, publication counts alone do not capture the full value of a scientist's contributions. Notably, even journals evaluate their own performance using a portfolio of complementary metrics rather than publication volume alone.

FASEB encourages NIH to partner with scientific societies, research institutions, and other stakeholders to modernize research assessment in ways that are evidence-based, practical, and appropriate across disciplines, career stages, institution types, and resource levels. FASEB stands ready to support NIH by identifying meaningful scientific impact indicators and assessment approaches, evaluating implementation strategies, and helping minimize unnecessary administrative burden, particularly for smaller and less-resourced institutions.

Topic 1: Rigor and Reproducibility

FASEB recommends NIH:

Establish clear definitions of reproducibility and causal research before incorporating them into scientific impact assessment. NIH should clearly define these concepts and communicate consistent expectations before incorporating them into evaluation or reporting. Scientific societies can help ensure these definitions are interpreted consistently across disciplines and implemented equitably.

Complement traditional productivity measures with evidence of rigorous scientific practice. Where measurable indicators are appropriate, NIH should emphasize established practices that support rigor, including authentication of key biological and chemical resources, adherence to community reporting standards, appropriate preregistration of confirmatory studies, and successful independent validation. These indicators provide an additional layer of value and understanding of the incentive structure at institutions.

Integrate transparency for generative AI into existing reproducibility policies. As AI becomes increasingly integrated into biological and biomedical research, NIH should incorporate documentation of AI tools, such as the model, provider, and version where appropriate, within existing authentication and reporting requirements rather than creating separate reporting systems. Building on NIH's existing *Enhancing Reproducibility through Rigor and Transparency* guidance (NIH, 2025), this approach will promote transparency, support computational reproducibility, and minimize additional administrative burden while allowing expectations to evolve alongside advances in AI.

Invest in infrastructure that enables rigorous science. Reproducibility depends on sustained access to shared research resources, core facilities, modern instrumentation, and skilled technical staff. NIH should ensure that investments in reproducibility are matched by

investments in the infrastructure needed to support them; otherwise, expectations for rigor may outpace institutional capacity.

Recognize replication as a valuable scientific contribution. Independent replication strengthens confidence in scientific findings but remains underrecognized within traditional systems of academic reward. NIH should recognize replication studies as distinct scholarly contributions within existing reporting mechanisms and continue developing funding opportunities that support high-priority replication efforts. To foster this commitment to replication, today, many societies have at least one journal in their portfolio that allows for recognition of and publication of replication studies, and several societies have supported the community further by dedicating special issues to replication studies.

Support scientific societies as critical to the research ecosystem as they seek to advance reproducibility. Biological and biomedical scientific societies are uniquely positioned to develop discipline-specific standards, identify high-priority opportunities for replication, disseminate best practices, and provide training in rigorous and reproducible research methods. NIH should continue to support society-led initiatives, including the development of training resources and the integration of rigor and reproducibility modules into programs serving researchers across career stages, including predoctoral trainees, postdoctoral researchers, and early-career investigators.

Scientific impact cannot be meaningfully assessed without confidence in the quality and reliability of the underlying science. FASEB encourages NIH to build its scientific impact framework on the foundation of rigorous, expert peer review, which remains the cornerstone of research assessment and the credibility of science. Responsible indicators of rigor should supplement, not substitute for, expert scientific judgment and should be implemented in partnership with the scientific community, consistent with the principles of Executive Order 14303, *Restoring Gold Standard Science*.

Topic 2: Data, Software, and Model Sharing

Peer-reviewed publications remain the primary scholarly output of biological and biomedical research and an essential means of communicating scientific discoveries. The value of biological and biomedical research, however, also extends to data, software, computational models, and other enduring research outputs that enable validation, reuse, and new lines of investigation. NIH has made considerable progress through its *Data Management and Sharing Policy* (NIH, 2023). Building on this progress, NIH should recognize and reward the creation, stewardship, and sharing of high-quality research resources to maximize the return on the nation's investment in biological and biomedical research.

FASEB recommends NIH:

Recognize data, software, and computational models as additional scholarly products alongside publications. These outputs should be evaluated in addition to publications in funding decisions and progress reporting. NIH should continue incorporating these contributions into existing investigator reporting mechanisms, including the Biographical Sketch Common Form and Contributions to Science, so that researchers receive appropriate credit for developing resources that enable discovery across the biological and biomedical enterprise.

Reward quality, reuse, and interoperability. The value of shared research resources lies in their accessibility, reliability, interoperability, and scientific utility, not simply whether they are deposited. Assessment systems that prioritize quantity over quality, usability, or scientific value risk encouraging compliance-driven behavior without meaningfully advancing research. Where measurable indicators are appropriate, NIH should emphasize adherence to the FAIR Principles (Wilkinson et al., 2016), documented reuse, appropriate citation, interoperability, and downstream scientific impact. Consistent with FASEB's longstanding position on data management and access, incentives should encourage the sharing of high-value research resources while avoiding requirements that create unnecessary administrative burden with limited scientific benefit.

Recognize stewardship as a valuable contribution. Long-term stewardship of datasets, software, repositories, and other shared research resources requires sustained scientific expertise. NIH should recognize activities such as curation, quality control, documentation, and long-term maintenance as distinct, valuable contributions within investigator assessment. Recognizing stewardship will encourage sustained investment in high-quality research resources that remain reliable, reusable, and valuable over time.

Invest in the infrastructure that enables responsible sharing and long-term stewardship of research resources. Expectations for sharing data, software, biospecimens, model organisms, reagents, and other valuable research resources should be matched by sustained investments in repositories, biobanks, research resource centers, core infrastructure, cloud resources, metadata standards, and the skilled technical workforce required to curate and maintain them. Without this support, valuable scientific resources may become inaccessible, limiting reproducibility, reducing reuse, and diminishing the long-term return on NIH's investment. Long-term stewardship of these resources should not depend solely on individual research grants.

Promote transparency and reproducibility for computational models. As artificial intelligence and computational modeling become increasingly integrated into biological and biomedical research, NIH should encourage transparent documentation of model provenance, versioning, computational workflows, and other information needed to support reproducibility, as appropriate. Building these expectations into existing reporting practices will improve

transparency, strengthen confidence in AI-enabled research, and allow guidance to evolve alongside advances in computational science.

Providing funding for, and rewarding, the creation and stewardship of the comprehensive knowledge ecosystem, including peer-reviewed publications, data, software, and computational models, is essential to maximizing the return on the nation's investment in research.

Topic 3: Training and Mentoring

NIH's mission depends on a strong, sustainable biological and biomedical research workforce, and mentorship is among the most important contributions scientists make to that workforce. Yet mentorship remains largely invisible in traditional measures of scientific productivity and impact. As a federation of scientific societies, FASEB has extensive experience advancing research skill development, mentorship, and workforce development through NIH-funded programs, scientific society initiatives, and evidence-based policy recommendations. Drawing on this experience, FASEB encourages NIH to recognize mentorship as a valuable contribution, reward high-quality training environments, and adopt assessment approaches that reflect the various career pathways essential to the biological and biomedical research enterprise.

FASEB recommends NIH:

Recognize mentorship as a valuable contribution. Effective mentorship should be recognized alongside scientific achievement in NIH assessment and reporting. Building on recommendations from the NIH Advisory Committee to the Director Working Group on Re-envisioning NIH-Supported Postdoctoral Training (NIH ACD, 2023), NIH should incorporate structured documentation of mentoring activities, mentor development, and scholar outcomes into relevant existing reporting mechanisms. Assessment should emphasize mentoring quality and scholar success rather than simply the number of scholars supervised, while recognizing participation in evidence-based mentor training as an indicator of scientific leadership.

Leverage existing NIH training programs rather than creating new assessment systems. Institutional training grants, particularly T32 programs, represent NIH's most established framework for evaluating research training because they already incorporate structured training plans, mentor oversight, career development activities, and longitudinal scholar outcome reporting. NIH should build on these established mechanisms to identify effective mentoring practices, evaluate training environments, and disseminate evidence-based approaches while minimizing additional administrative burden.

Evaluate training through meaningful scholar outcomes that recognize various career pathways. NIH should recognize successful career outcomes across academia, industry, government, nonprofit organizations, education, and other sectors. Building on existing trainee outcome reporting and standardized career taxonomies such as Unified Career Outcomes

Taxonomy (UCOT; Stayart et al., 2020), indicators including degree completion, career transitions, K99-to-R00 transition, K-to-R conversion, and time to research independence can provide meaningful measures of program effectiveness when interpreted within program context.

Recognize competency-based development as an indicator of quality. Effective research training develops competencies in scientific rigor, leadership, communication, professionalism, responsible conduct of research, and career development. NIH should encourage competency-based assessment that complements traditional research outputs while remaining flexible across career stages and career pathways.

Support mentoring networks that reflect today's biological and biomedical workforce. Scientists increasingly benefit from mentoring networks that include research, career, peer, and cross-sector mentors. NIH should recognize and encourage distributed mentoring models that strengthen career development, broaden access to expertise, and better reflect the collaborative nature of today's biological and biomedical research enterprise.

Recognize the complementary role of scientific societies in workforce development. Scientific societies provide mentoring, leadership development, networking, career exploration, and professional communities that complement institutional training. These activities expand access to expertise, strengthen professional networks, and complement institutional training across career stages. NIH should continue partnering with scientific societies to extend the reach and impact of its workforce development investments.

Mentorship is among the most consequential—and least recognized—contributions scientists make to the biological and biomedical research enterprise. Assessment approaches should support long-term program improvement rather than short-term funding decisions and account for institutional context, discipline, career stage, and the time required for meaningful career outcomes to emerge. By recognizing mentoring excellence and high-quality training environments, NIH can strengthen the biological and biomedical workforce and, in turn, the nation's innovation economy.

Topic 4: Collaboration

Collaboration has become a defining feature of modern biological and biomedical research, with discoveries increasingly emerging from multidisciplinary teams that span institutions, sectors, and scientific disciplines. Through its member societies, scientific meetings, and collaborative initiatives, FASEB has long supported team science and interdisciplinary collaboration. As NIH develops approaches for recognizing and assessing scientific impact, assessment should recognize the range of scientific contributions that enable collaborative research while remaining appropriate across disciplines, career stages, and research settings.

FASEB recommends NIH:

Recognize the full range of scientific contributions that enable team science. Scientific impact should not be inferred solely from a singular metric; a matrix approach is recommended. Rigorous peer review should continue evaluating the quality of collaborative science, while structured contribution taxonomies such as the Contributor Roles Taxonomy (CRediT) developed by the National Information Standards Organization (NISO) provide additional context regarding individual contributions. NIH should incorporate these taxonomies into existing reporting mechanisms to provide additional context regarding individual contributions.

Evaluate collaboration based on scientific contributions rather than team size. Assessment should recognize interdisciplinary integration, shared leadership, effective coordination among collaborators, and the scientific advances enabled by collaboration rather than relying on coauthor counts or other quantitative measures.

Recognize cross-sector collaboration as an indicator of valuable impact. Biological and biomedical discovery increasingly depends on partnerships among academia, industry, government, healthcare systems, nonprofit organizations, patient communities, and scientific societies. NIH should recognize collaborations that meaningfully advance scientific discovery, translation, workforce development, or public health.

Recognize collaborative leadership within NIH assessment and reporting. NIH should continue recognizing multiple principal investigators, co-investigators, and other collaborative leadership roles within its funding and research assessment processes, reinforcing institutional recognition of team-based science.

Ensure collaborative science supports early-career investigators. Assessment approaches should enable students, postdoctoral scholars, and early-stage investigators to clearly document their individual scientific contributions within collaborative projects. Participation in team science should provide opportunities to demonstrate scientific independence rather than obscure individual contributions.

Avoid metrics that create unintended incentives. Measures based solely on coauthor counts, collaboration team size, or other quantitative proxies are poor indicators of meaningful collaboration and may encourage honorary authorship or unnecessary expansion of research teams. Collaboration should instead be evaluated based on documented scientific contributions, the quality and impact of the collaborative work, and disciplinary norms.

Scientific progress increasingly depends on effective collaboration across disciplines, institutions, and sectors. By recognizing meaningful contributions to team science, NIH can strengthen collaborative innovation while ensuring assessment remains equitable, transparent, and reflective of modern biological and biomedical research.

Topic 5: Entrepreneurship and Translation

Biological and biomedical research creates value by generating new knowledge and translating discoveries into improved health. Successful translation depends on sustained investment in investigator-initiated foundational research together with effective cross-sector partnerships. Through their peer-reviewed journals, scientific meetings, and professional networks, FASEB's member societies help advance discoveries from the laboratory to practical application. As NIH develops approaches for recognizing scientific impact, entrepreneurial and translational activities should be recognized as important scientific contributions without diminishing the vital importance of foundational discovery research.

FASEB recommends NIH:

Recognize translational research as a valuable scientific contribution. NIH should recognize contributions that advance discoveries toward clinical practice, public health, diagnostics, therapeutics, medical technologies, and other real-world applications. Assessment should recognize progress across the translational continuum, including advancement through the research and development pipeline and the maturation of early-stage discoveries, rather than focusing solely on commercial endpoints.

Recognize entrepreneurial and innovation activities that responsibly advance biological and biomedical discoveries. Activities such as technology development, invention disclosure, patenting, licensing, startup formation, and regulatory advancement can represent meaningful scientific contributions when they advance the translation of NIH-supported discoveries into products, services, and technologies that improve health. Recognizing these activities within existing reporting mechanisms will encourage investigators to pursue translational research while preserving support for foundational discovery. Because commercialization infrastructure, including technology transfer offices, industry partnerships, and access to capital, is unevenly distributed, NIH should avoid indicators that disadvantage investigators at smaller or less-resourced institutions or reward patenting activity without evidence of meaningful translational or public benefit.

Recognize the role of the Bayh-Dole framework in moving federally funded discoveries toward public benefit. For more than four decades, the Bayh-Dole Act has enabled discoveries made with federal support to advance toward application through invention disclosure, patenting, and licensing, catalyzing the development of new drugs, diagnostics, and medical technologies. NIH's assessment framework should reinforce the successful incentives established under the

Bayh-Dole Act by recognizing technology transfer, patenting, licensing, and related translational activities as legitimate scientific contributions while continuing to value the foundational research on which these activities depend.

Support cross-sector partnerships that accelerate bench-to-bedside translation. Successful translation depends on collaboration among academic institutions, industry, government agencies, healthcare systems, patient organizations, philanthropies, and scientific societies. NIH should recognize meaningful partnerships that move discoveries toward application and improve public health, while ensuring that individual investigator contributions within these partnerships remain visible in assessment.

Evaluate translation using meaningful evidence of scientific advancement. Assessment should emphasize evidence that discoveries are progressing toward application, including adoption of new methods or technologies, advancement through the development pipeline, and incorporation into clinical practice or public health. Quantitative measures should be interpreted within the context of discipline, career stage, and the long timelines, often a decade or more for new therapeutics, that translation frequently requires, so that early-career researchers are not disadvantaged when translational outcomes fall outside their typical evaluation windows.

Translation is one of the principal ways biological and biomedical research delivers benefits to patients and society, but it depends on sustained investments in rigorous, peer-reviewed basic research. By recognizing entrepreneurship and translation as complementary dimensions of scientific impact and continuing to support the Bayh-Dole framework, NIH can strengthen the continuum from discovery to application while preserving incentives for foundational science.

Topic 6: Foundational Scientific Exploration

Foundational scientific exploration is the wellspring of biological and biomedical progress. The discoveries that ultimately transform human health frequently originate in curiosity-driven investigation whose applications cannot be foreseen at the outset and may take years or decades to emerge. Sustained progress in translational and clinical research depends on a continuous supply of the new knowledge, mechanisms, and targets that fundamental research generates. As NIH develops approaches to measuring scientific impact, FASEB encourages the agency to also recognize, protect, and reward foundational biological and biomedical exploration, recognizing that assessment systems built solely around near-term or easily quantified outcomes risk undervaluing the basic science on which future advances depend.

FASEB recommends NIH:

Protect foundational research from assessment systems that favor near-term outcomes. Because the impact of basic biological and biomedical science often emerges years or decades

later through unpredictable pathways, NIH should adopt approaches to evaluation that account for the long and uncertain timelines of discovery research.

Recognize the advancement of fundamental knowledge as a core indicator of scientific impact. Expanding our understanding of biological systems, including fundamental cell, molecular, and organismal biology and the discovery of new biological concepts and therapeutic targets, is a foundational contribution to the biological and biomedical research enterprise, independent of near-term application. NIH should recognize contributions that advance fundamental knowledge, generate new concepts and tools, or open new lines of inquiry as meaningful indicators of impact within existing evaluation and reporting mechanisms.

Support high-risk, high-reward research through appropriate assessment. Transformative discovery often depends on investigations that carry significant uncertainty and may not yield the anticipated result. NIH should ensure that assessment approaches do not penalize scientifically rigorous research that does not produce the anticipated findings and should recognize well-executed negative or null results as meaningful contributions to scientific knowledge.

Account for long and unpredictable timelines in evaluating foundational research. Measures tied to short reporting windows cannot capture the trajectory of basic science. NIH should favor evaluation approaches that accommodate extended timeframes, avoid drawing conclusions about impact prematurely, and allow the significance of foundational biological and biomedical discoveries to emerge over time. Because foundational discovery is inherently unpredictable, impact assessment frameworks should continue to support a diverse portfolio of investigator-initiated, curiosity-driven research rather than directing resources primarily toward predefined priorities or readily measurable near-term outcomes.

Foundational biological and biomedical science exploration is the source of discoveries that ultimately transform health, yet its value is the hardest to capture through conventional measures of impact. FASEB urges NIH to design assessment approaches that recognize, protect, and reward curiosity-driven science, thereby safeguarding the discovery research from which tomorrow's advances will arise.

Topic 7: Public Impact

Biological and biomedical research generates value through publications, datasets, software, computational models, and other enduring scientific outputs. It also delivers value in improving health, strengthening the economy, informing clinical practice, building public trust, and benefiting society in ways that often emerge over many years. As NIH develops approaches to assess public impact, FASEB encourages the agency to recognize the many pathways through

which research benefits the public. Public impact is multidimensional and should be assessed through a matrix of complementary indicators that reflect NIH's mission.

FASEB recommends NIH:

Recognize improvements in health and clinical practice as core indicators of public impact. NIH should also recognize contributions that improve health outcomes, inform clinical care, support implementation of evidence-based interventions, and broaden access to effective care, including the responsible return of research results where appropriate.

Incorporate validated measures of economic and societal benefit. NIH-supported research drives innovation, economic growth ($2.57 for each $1.00 invested), workforce development, and national competitiveness. Where appropriate, NIH should also incorporate validated measures of economic impact alongside other metrics to better communicate the broad public return on federal investment in biological and biomedical research.

Recognize public engagement and trustworthy science communication as scientific contributions. Building and maintaining public trust requires sustained engagement with patients, communities, educators, policymakers, and the public. NIH should also recognize meaningful contributions to science communication, patient and community engagement, transparency, and other activities that strengthen confidence in biological and biomedical research. Scientific societies are important partners in developing and disseminating evidence-based communication practices and fostering public trust.

Avoid simplistic measures of public impact. No single metric can adequately capture the many ways biological and biomedical research benefits society, and a matrix of indicators is warranted. Likewise, economic measures should complement, not replace, assessment of health outcomes and broader public benefit. NIH should instead adopt a balanced framework that recognizes health outcomes, clinical practice, economic growth, accessibility, public trust, and societal benefit as complementary dimensions of impact.

Recognize public impact within existing application, reporting, and research assessment processes. Investigators should be able to document meaningful public impacts, including contributions to clinical guidelines, public policy, accessibility, community engagement, the responsible return of research results, and public understanding of science. Narrative descriptions should complement quantitative indicators by capturing important additional impacts.

Partner with scientific societies to strengthen public engagement and scientific literacy. Scientific societies play an essential role in communicating biological and biomedical scientific advances, convening experts, engaging communities across the country, and developing trusted

educational resources. NIH should continue partnering with scientific societies to strengthen public understanding of biological and biomedical research and ensure that measures of public impact reflect the range of disciplines and communities they serve.

Public impact is the ultimate expression of NIH's mission. FASEB encourages NIH to adopt assessment approaches that recognize the various ways biological and biomedical research improves lives by valuing publications alongside datasets, software, computational models, and other scientific contributions, while remaining flexible, evidence-based, and appropriate across scientific disciplines and stages of research.

In conclusion, FASEB supports NIH's efforts to develop a more comprehensive approach to measuring and rewarding scientific impact that recognizes the full range of contributions advancing biological and biomedical research. ***As NIH develops this framework, we encourage the agency to preserve rigorous, expert peer review as the foundation for evaluating scientific quality while adopting responsible, evidence-based indicators that complement—not replace—scientific judg*ment**.

FASEB urges NIH to adopt the recommendations outlined above to ensure that scientific impact is assessed in ways that strengthen rigor, encourage innovation, recognize the range of contributions across the biological and biomedical research enterprise, and minimize unnecessary administrative burden. FASEB and its member societies welcome the opportunity to provide additional information and to work with NIH to develop and implement an assessment framework that advances scientific excellence while maximizing public benefit.

Respectfully submitted,

[redacted]

Michael N. Lehman, PhD
President
Federation of American Societies for Experimental Biology

cc: Darla P. Henderson, PhD
Executive Director & Chief Executive Officer

Contact: dhenderson@faseb.org | 301-634-7526

180. UNC Center for Health Promotion and Disease Prevention

I am responding: On behalf of an organization

Name: Caroline Dilworth

Name of Organization: UNC Center for Health Promotion and Disease Prevention

Type of Organization: Not Applicable

Role: Other

Role - Other: Deputy Director

Date Submitted: 8/17/2026

1. Rigor and Reproducibility: o Significance of the health problem addressed - # people impacted, severity of the issue, benefit of prevention
o Potential for scale up and dissemination
o Feasibility of broad implementation of evidence-informed interventions
o Evidence of community-engagement and input on research priorities
o Cost-effectiveness
o Evidence of clinical significance of change/improvement in proposed mechanisms and outcomes (rather than statistical significance only)

2. Data, Software, and Model Sharing: o University financial support and personnel effort for data management, storage, and sharing across disciplines

3. Training and Mentorship: o Joint publications with community partners
o Evidence of broad reach/impact
o Evidence of prioritizing trainees – for example, as first author on publications
o Interdisciplinary collaborations

4. Collaboration: o Collaboration across disciplines
o Rewarding more than first author publications, expecting multiple

5. Entrepreneurship and Translation: o Balancing encouragement for innovation with restrictions related to COI
o Recognition of social entrepreneurship and the role in translation to the community

6. Foundational Scientific Exploration: o Apply this approach to translational as well as technological research

7. Public Impact: o Implementation research including cost-effectiveness
o Importance of community engaged research related to economic development
o Importance of science communication and training
o Potential utility of “Ripple Effects Mapping”:
https://communityengagement.uncg.edu/research_collabs/ripple-effects-mapping/

8. Other Areas: N/A

181. Andrew Kant

I am responding: On behalf of myself

Name: Andrew Kant

Name of Organization: UNC-Chapel Hill

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/17/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: The gap in current reporting

NIH states that more can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts, and separately identifies economic growth among the dimensions of public impact worth recognizing.

The barrier is not willingness. Investigators pursue this work, and institutions fund offices to support it. What is missing is a defined category for these outputs in the instruments through which scientific contribution gets recorded.

The structure of progress reporting shows this plainly. Research Performance Progress Reports provide a dedicated field for websites and internet sites, a field for technologies or techniques, and a field for other products that resolves in practice to publications and shared research resources. Publications are captured as structured identifiers and carry compliance enforcement. A licensed technology, a formed company, a regulatory clearance, or an adopted device design has no comparable field.

The same asymmetry appears at program level. NCATS supporting instructions for hub progress reports in the Clinical and Translational Science Awards Program require recipients to report on funded pilot projects, and provide a recommended format of roughly twenty structured fields per project, publications among them as indexed identifiers. Return on investment is collected as free text, limited to 250 characters. NCATS states that the recommended format exists to support aggregation of pilot data and reporting on program impact. Publication output can therefore be compared across the consortium. Translational progression cannot.

The categories through which hubs report key outcomes reflect the same gap. They span disease areas,

populations, workforce, and health equity. None covers a product, a technology transferred, an industry partnership, or economic activity. A program cannot report an outcome for which its reporting vocabulary has no term.

I direct a medical device development program within a CTSA hub, and the practical effect is familiar. A multi-year device development effort and a one-hour consultation are recorded as the same unit of service activity. Work that reaches a regulatory or design milestone has no field to occupy, so it goes into narrative and drops out of anything that can be counted.

The consequence extends past individual credit. NIH cannot presently answer, at the portfolio level, how often its investments reach the point of delivering something a patient can receive.

Recommendation 1: Define a category for translational outputs where contribution is recorded

NIH should add a named, structured category for translational outputs to the biosketch and to progress reporting, populated by verifiable identifiers rather than narrative assertion. Anchors already exist: patent numbers, FDA submission and clearance or approval identifiers, SBIR and STTR award numbers, and invention and licensing records already reported to the government under existing obligations.

This asks for no new data collection. Much of the information is already reported. It is reported as an institutional compliance obligation to the government rather than as a record of individual scientific contribution, and it never returns to the investigator or reaches reviewers in any form. Route what exists into the instruments where merit is assessed.

Feasibility across institution types is high, because the category records events that either occurred or did not. An investigator at a small institution with one licensed technology can report it as readily as an investigator at a large one with many. No institutional infrastructure is required to participate.

Recommendation 2: Measure translational progression, not market outcomes

What goes into the category matters as much as its existence, and an outcome-anchored indicator would do harm.

Commercial outcomes materialize years after any award period ends. They depend on capital availability, corporate strategy, market size, and reimbursement, none of which an investigator controls. An indicator anchored to market outcomes would reward timing and market conditions rather than scientific contribution. It would penalize work in rare disease, pediatrics, maternal health, and other areas where clinical need is high and commercial return is limited. It would also create pressure to sustain projects that should be stopped.

The more defensible unit of measure is the completed de-risking transition. Candidate indicators include:

- Regulatory pathway determined, with predicate or comparator identified where applicable
- Pre-submission or equivalent agency meeting requested and held

- Investigational application filed (IND, IDE, or equivalent)
- Design specification frozen, or verification and validation testing completed
- Invention disclosed, and patent application filed or claim allowed
- Technology optioned or licensed to a development partner
- Sponsored research or co-development agreement executed with a commercial partner
- Translational funding secured downstream of the award (SBIR, STTR, or equivalent)
- Documented decision not to proceed, with the technical or clinical basis recorded

Each occurs inside an award period, falls within the investigator's control, and is documentable when it happens rather than years later.

That last item carries more weight than its placement suggests. A well-reasoned decision not to proceed should count as a completed transition, not as an absence of output. Recognizing negative determinations removes the incentive to keep unviable work alive for reporting purposes, and it treats efficient failure as what it is: a contribution to translational efficiency.

There is a further reason to measure progression rather than outcomes. Translational projects frequently receive federal support several development cycles before any commercial event, and reporting obligations end when the supporting award ends. An indicator that waits for a market outcome will capture nothing, whatever the underlying work achieved.

Verification

Every indicator above is anchored to a record generated by a third party: a patent office, a regulatory agency, a funding agency, or an executed agreement between institutions. None depends on self-assessment. Each can be audited against an external source. This is what separates the proposed category from narrative impact statements, which cannot be verified, compared, or aggregated.

Feasibility across career stages

The current gap falls hardest on early-career investigators. Translational development routinely runs five to ten years, overlapping the period during which an investigator assembles a promotion and tenure record. An assistant professor who commits effort to advancing a technology toward clinical use spends that time against a record that counts publications and grants, and receives nothing in the formal accounting for the technology itself. The rational response is to defer the work until after promotion, or to decline it.

Institutions largely follow NIH in what they count. A defined category at the federal level would give promotion committees a recognized reference point, at no cost to NIH beyond the definition itself. That is the mechanism by which a reporting change becomes a change in behavior.

Thank You

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

Description: A defined reporting category for translational and entrepreneurial outputs.

182. Beth Israel Lahey Health

I am responding: On behalf of an organization

Name: Gyongyi Szabo

Name of Organization: Beth Israel Lahey Health

Type of Organization: Healthcare Institution

Role: Organizational Official

Date Submitted: 8/17/2026

1. Rigor and Reproducibility: In the grant review process, there should be consideration for the value of Rigor and Reproducibility, particularly for grants that are deemed as addressing topics that are not novel. For example, a new grant application that is on a topic that already has some existence of prior art, could be still highly valuable from the standpoint of Reproducibility and broaden understanding of previously reported findings.

2. Data, Software, and Model Sharing: Scientific impact should include creation of high-quality, reusable data, software, and research resources. However, there is no one-size-fits-all approach to data sharing across biomedical disciplines. Current data-sharing requirements often represent substantial, unfunded effort for investigators, while long-term maintenance and curation exceed the capacity of individual laboratories. Institutions are increasingly responsible for ensuring regulatory compliance but require sustained federal investment in infrastructure, personnel, and data stewardship. NIH should recognize institutional support for data sharing and provide dedicated resources rather than relying on unfunded mandates.

3. Training and Mentorship: Impact should extend beyond publications to developing the next generation of scientists. Indicators could include mentoring quality, trainee career outcomes across multiple career paths, and contributions to research integrity and independence. Expanding mentorship will require recognizing the significant time commitment of senior investigators. NIH should consider in-kind or other mechanisms that compensate mentoring effort, creating incentives for experienced researchers to mentor while strengthening career development and independence of early-stage investigators. In addition to the often-uncompensated time and effort of mentors, host laboratories and institutions provide substantial in kind contribution for trainees including laboratory equipment, access to highly specialized core facilities, access to and maintenance of animal colonies and laboratory space. None of these critical elements in the success of emerging scientists/investigators are compensated for in the current funding opportunities

4. Collaboration: Modern biomedical research depends on multidisciplinary teams. NIH should recognize meaningful contributions to collaborative science through leadership, shared infrastructure, cross-disciplinary partnerships, and coordination of multi-site studies. Evaluation systems could better recognize substantive contributions from all members of successful research teams. The new requirements and rules related to international collaborations are particularly limiting. With globalization in all areas of science, it is unavoidable that lead laboratories and scientists will have multiple

international collaborations and publications. Even for review articles, most leading journals require multiple authorship from different countries/continents.

5. Entrepreneurship and Translation: Appropriate indicators should include clinical implementation, regulatory milestones, licensing, industry partnerships, incorporation into clinical guidelines, and demonstrated public benefit. Evaluation should recognize that meaningful translation often requires long timelines and collaborative partnerships.

6. Foundational Scientific Exploration: Basic and high-risk research should be rewarded even when outcomes are uncertain. Assessment should recognize development of transformative methods, foundational datasets, and new scientific concepts that enable future advances. Evaluation should emphasize scientific quality and long-term influence.

7. Public Impact: Public engagement should be recognized as a core component of scientific impact. Researchers are uniquely positioned to improve public understanding of science, yet outreach and science communication are rarely supported or rewarded. This represents a missed opportunity that contributes to misunderstanding and mistrust of science. NIH should provide dedicated funding for public engagement activities by individual investigators or through coordinated institutional programs, ensuring that communication with the public is recognized as an important and supported component of the research enterprise.

8. Other Areas: N/A

183. Weill Cornell Medicine

I am responding: On behalf of an organization

Name: Dr. Hugh Hemmings

Name of Organization: Weill Cornell Medicine

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 8/17/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/WCM-Response_RFI-Metrics-for-Scientific-Success_August-2026.pdf

Weill Cornell Medicine
Office of the Research Dean

Hugh C. Hemmings, Jr, MD, PhD, FRCA
Senior Associate Dean for Research
Joseph F. Artusio, Jr Professor and Chair
Department of Anesthesiology
Professor of Pharmacology
Weill Cornell Medicine
Anesthesiologist-in-Chief
NewYork-Presbyterian Hospital

August 19, 2026

NIH Office of Science Policy
6705 Rockledge Drive
Suite 630
Bethesda, MD 20892

Re: NOT-OD-26-087 — Request for Information on Measuring and Rewarding Scientific Impact

Dear NIH Office of Science Policy,

Weill Cornell Medicine is committed to excellence in patient care, scientific discovery and the education of future physicians in New York City and around the world. The doctors and scientists of Weill Cornell Medicine — faculty from Weill Cornell Medical College, Weill Cornell Graduate School of Medical Sciences, and Weill Cornell Physician Organization — are engaged in world-class clinical care and cutting-edge research that connect patients to the latest treatment innovations and prevention strategies. Located in the heart of the Upper East Side's scientific corridor, Weill Cornell Medicine's powerful network of collaborators extends to its parent university Cornell University and through affiliations and collaborations across the globe. Weill Cornell Medicine faculty provide exemplary patient care at NewYork-Presbyterian/Weill Cornell Medical Center, NewYork-Presbyterian Westchester Behavioral Health Center, NewYork-Presbyterian Lower Manhattan Hospital, NewYork-Presbyterian Queens and NewYork-Presbyterian Brooklyn Methodist Hospital. Weill Cornell Medicine is also affiliated with Houston Methodist.

On behalf of Weill Cornell Medicine, thank you for the opportunity to provide comments in response to the NIH Request for Information (RFI) on Measuring and Rewarding Scientific Impact. We commend the NIH for continuing to seek external input as it considers policy changes with broad implications for the research community. Meaningful stakeholder engagement is essential to ensuring that new policies are informed by the perspectives of those who will be directly affected by them.

We preface our response to the individual topics with three limitations about NIH's effort to identify measurable indicators and incentives of scientific success beyond publication counts and citation metrics.

First, scientists' incentives are largely intrinsic. Curiosity and professional purpose are the motivators and progress and discovery are the incentives. Scientists require adequate resources and institutional support. When these resources are more abundant, scientists may be able to achieve more, and when resources are withdrawn from one line of research and made available for another, scientists may change course accordingly. Based on our anecdotal observations, if resources are held equal, scientists do not become more creative or more productive in proportion to external incentives.

Second, scientific success is defined more broadly than the rigor, reproducibility, data sharing, mentorship, collaboration, and translational impact criteria NIH proposes to measure. Success can mean answering a question or formulating a better one. It can mean discovering something with no obvious utility today that may, decades later, prove foundational to drug

discovery - or it can mean solving the last, technically modest but functionally essential step in a multi-step process built on many other scientists' work over preceding decades. Importantly, we would like to emphasize that foundational knowledge is required before any applied use can be undertaken. As an example, without a firm understanding of the basics of DNA base-pairing, we could not develop gene editing nor advance its applications for patients.

Third, scientists, through their advanced content expertise, remain the best available judges of scientific success. The indicators of an individual scientist's success extend beyond publication counts and citation metrics but are rooted largely in the impact of their published work, whether a sustained, often-cited body of work or a single paper too recent to have accrued citations. These indicators include, but are not limited to, winning competitive fellowships and grants; being sought out by students and postdoctoral fellows who wish to join their lab; being invited to speak at, chair, or organize national and international scientific conferences; being asked to serve on or lead national guidelines panels; being asked to review manuscripts for leading journals; being invited to join journal editorial boards or scientific advisory boards for foundations, NGOs, and companies; election to honorific scientific societies; and prizes for exceptional achievement. These indicators vary by career stage and field. They are qualitative, not quantitative, and in that sense, evaluable but not measurable.

Considering the noted limitations of quantitative metrics, we propose initial ideas for developing frameworks with measurable indicators.

1. Rigor and Reproducibility

Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings.

Response:

- Reproducibility is foundational to scientific progress. Rigor should be assessed at the level of study design and methodology in addition to after-the-fact judgments about whether results can be replicated. Institutions already support this by building rigor training into graduate and postdoctoral curricula and by ensuring institutional review processes (IRB, IACUC, biostatistics cores) are adequately resourced to support investigators in study design.

- Infrastructure and incentives for replication should be separated from the incentives that drive original discovery. A dedicated funding mechanism for replication studies would allow institutions and investigators to pursue reproducibility work without diverting resources from novel research and would give replication studies a clear, fundable, and publishable home.

- Reproducibility efforts are most productive when targeted rather than universal. NIH should prioritize replication investment in areas where a reproducibility failure carries the greatest downstream cost, e.g., findings poised to inform clinical trials or policy changes, rather than mandating uniform replication requirements across all funded research.

2. Data, Software, and Model Sharing

Sharing scientific data and tools accelerates biomedical research discovery, in part, by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies. NIH seeks to incentivize and reward these activities to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.

Response:

- Academic medical centers face difficulty reconciling multiple, sometimes conflicting regulatory frameworks governing the protection and sharing of data. A clearer NIH statement on standards for sharing de-identified data without risk of penalty paired with resources to help institutions meet those standards and support for the considerable costs of sharing, would substantially advance high-quality, widespread data sharing.

- Existing NIH initiatives, including the Data Management and Sharing Policy (DMSP) and the Data Sharing Index (S-Index), provide useful guidance for investigators and administrators. Clarifying in DMSP guidance that investigators need not pay publisher fees to comply with PubMed Central posting requirements would reduce friction and increase compliance. In addition, DMSP and S-Index measures should explicitly account for sensitive data (e.g., protected health information) and direct investigators and administrators to institutional secure enclaves and legal data-sharing agreements that enable cross-organizational data sharing.

- Models and software that leverage statistical techniques, such as machine learning and artificial intelligence, may contain sensitive data (e.g., protected health information) that, while challenging to reidentify, present a challenge for sharing. Coordination among NIH and partners regarding expectations and requirements for sharing of AI/ML-based models and software can potentially reduce friction for investigators and administrators going forward.

3. Training and Mentorship

NIH recognizes that advancing its mission requires promoting the growth and stability of the biomedical research workforce. While NIH supports training and protects time through existing policies and programs, additional indicators are needed to assess and quantify the impact of these efforts.

Response:

- A best practice could be for training programs to track outcomes beyond publications, including attainment of independent funding; career advancement and retention in research and academic medicine; leadership roles among alumni and former trainees; and contributions to science communication. Mentorship impact can be assessed through trainee engagement, career progression, and the development of mentoring networks that sustain the next generation of investigators.

- At the medical student level, useful indicators could include: the percentage of students participating in research; the percentage receiving competitive funding during medical school; the percentage taking a dedicated research year; and the percentage with presentations or publications during medical school.

- Current NIH evaluation criteria commonly capture trainee outcomes, mentorship quality, workforce development, and training activities, but give less weight to science communication with public audiences, STEM outreach and community education, leadership development and service contributions and the extent to which trainees “pay it forward” by mentoring others. These activities strengthen the biomedical workforce, build public trust in science, expand pathways into research careers, and improve retention, yet are not consistently captured in traditional measures of training success.

4. Collaboration

As biomedical research becomes increasingly interdisciplinary and convergent, collaboration is growing in both scale and complexity. NIH seeks new approaches for assessing and recognizing contributions to team-based science.

Response:

- Team-based science is now essential to progress. In fields such as cancer research, where interventions targeting oncogenic signaling can have favorable or unfavorable effects, assembling multidisciplinary teams of experts across these domains is often necessary for meaningful advances.

- NIH should recognize contributions to team-based science by expanding programs such as the Collaborative Program Grant for Multidisciplinary Teams (RM1) beyond NIGMS to other institutes. While other institutes have their own team-science mechanisms, the RM1's key distinguishing feature, a single application with one unified project, rather than a collection of sub-projects as in P01/P50/U54 structures, led by three to six principal investigators sharing management responsibility, merits broader adoption.

5. Entrepreneurship and Translation

To advance the translation of discoveries into products for people, NIH fosters the maturation of discoveries from bench to bedside and promotes innovation across sectors. More can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission.

Response:
Any single measure would be imperfect, but the following offer a starting point:

Short-term indicators
- Number of technology disclosures made to the institution's technology transfer office
- Number of filed U.S. patent applications
- Number of active licenses or options transferring faculty IP to commercial partners
- Number of corporate-sponsored research agreements (and associated funding)

Longer-term indicators
- Products on the market based on faculty IP
- Licensing revenue received from licensed faculty IP, as a proxy for economic impact
- Number of start-up companies founded on faculty IP that have received at least $2–5M in institutional or professional investment

6. Foundational Scientific Exploration

Basic research and high-risk, high-reward initiatives represent foundational scientific exploration with an inherent risk of failure but the potential for paradigm-shifting breakthroughs. Further embedding these approaches within the academic enterprise is essential to ensuring continued support for bold and innovative research directions.

Response:
- To insulate foundational research from risk-averse review, NIH should evaluate proposals on the significance of the scientific question and the originality of the approach using metrics tailored to protect early-stage investigators rather than on the extent of preliminary data or perceived certainty of success. Success should be assessed using long-horizon indicators of scientific value, including adoption of new methods and tools, reuse of FAIR datasets, independent replication, and dissemination of rigorous null or negative findings. For clinical conditions where progress has been limited (i.e. Alzheimer's disease given the historical focus on amyloid and tau), recalibrating funding to include exploring mechanisms of newly recognized targets could bring novel ideas and renewed energy to a field.

- NIH should expand investigator-centered funding mechanisms that sustain promising research directions beyond milestone-driven project awards, and should consider tracking “exploratory yield” metrics, such as the proportion of

grant budgets allocated to unproven mechanistic hypotheses and the rate at which basic-science findings generate unanticipated follow-on discoveries. Dedicated failure-tolerant award mechanisms (e.g., expanded NIH Director's Pioneer Award/EUREKA-style programs) should base renewal criteria on rigor of inquiry rather than positive results.

- Additional indicators of success for foundational science include whether findings later seed new subfields, spin out new research programs, or attract follow-on funding from other agencies or industry, even if an immediate patent or product is absent. NIH institutes should also track "long-horizon citation lag", the time between initial publication and eventual translational impact, as a complement to immediate citation counts, since basic science findings often take a decade or more to yield clinical or technological impact.

- Study sections and promotion and tenure committees should be instructed to credit well-designed studies that produce null or negative results. Penalizing negative outcomes disincentivizes the risk-taking this initiative aims to protect.

- Lastly, feasibility varies by career stage. Early-stage investigators need protected, failure-tolerant funding lines distinct from R01 renewal pressure.

7. Public Impact

NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.

Response:

- One measure of the impact of medical research is the incorporation of findings into clinical guidelines. Research that changes practice guidelines has substantial clinical impact. Research that increases the speed with which findings reach the patient in their physicians' office or at the point of care should be prioritized and incentivized.

- The Translational Science Benefits Model (TSBM) offers a structured framework for identifying and assessing the broader public and societal impacts of translational science beyond traditional publication and citation metrics. The TSBM identifies benefits across four domains, making it a potentially useful model for NIH to consider in assessing public impact:
    - Clinical and Medical Benefits: Person-involved care, procedures and guidelines, tools and products
    - Community and Public Health Benefits: Community engagement and wellness, health activities and products, health care characteristics, health promotion
    - Economic Benefits: Commercial products, financial savings and benefits, healthcare workforce
    - Policy and Legislative Benefits: Advisory activities, community-centered policy, policies and legislation

Sincerely,



Hugh C. Hemmings, Jr, MD, PhD, FRCA
Senior Associate Dean for Research

184. Make Data Count

I am responding: On behalf of an organization

Name: John Chodacki

Name of Organization: Make Data Count

Type of Organization: Non-Profit Research Organization

Role: Organizational Official

Date Submitted: 8/17/2026

1. Rigor and Reproducibility: We encourage NIH to recognize the creation, documentation, stewardship, and reuse of research data as contributions to rigor and reproducibility. Useful indicators could include structured data citations and documented examples of datasets being used for replication, validation, aggregation, methodological development, or comparison with new findings.

Because a citation or download alone does not reveal whether data contributed to reproducibility, we recommend combining quantitative evidence such as standardized dataset citations, views, and downloads with qualitative descriptions of how the data were used and contextual information about the discipline, dataset type, access conditions, intended users, career stage, and contributor roles.

NIH could explore these approaches through pilots across several Institutes and Centers, disciplines, career stages, repository types, and institutional settings. These pilots could examine the completeness of the underlying information, whether reviewers interpret it consistently, whether particular communities are disadvantaged, and whether the indicators create incentives for gaming or superficial compliance.

We also encourage NIH to account for controlled-access and sensitive human-subject data. Such datasets may contribute substantially to replication and validation even when their use cannot be measured through public downloads or other conventional open-access indicators.

2. Data, Software, and Model Sharing: We recommend that NIH treat research data as first-class research outputs and explicitly recognize their creation, curation, documentation, sharing, stewardship, and reuse throughout the grant lifecycle. Researchers could be enabled to identify datasets as distinct contributions in applications, biosketches, progress reports, and final reports. Recognition could extend beyond data producers to those who curate, document, steward, and enable the responsible reuse of data.

Evaluation could consider the quality and significance of a data contribution, rather than focusing only on whether data were deposited or made publicly available. A portfolio of evidence might include:
• Quantitative measures, including standardized dataset citations, views, and downloads;
• Qualitative evidence of replication, validation, subsequent research, clinical application, policy use, education, or collaborations enabled by the data; and

• Context about the discipline, dataset type, access conditions, intended users, career stage, and contributor roles.

We caution against creating a composite data-impact score or using raw citation, view, or download counts to rank researchers. These measures can provide useful evidence, but are most meaningful when interpreted in the appropriate disciplinary and research context.

We encourage NIH-supported repositories to assign persistent identifiers to datasets; provide complete, machine-readable metadata; identify contributor roles and relationships among research outputs; support structured data citation; collect views and downloads using common standards such as the COUNTER Code of Practice for Research Data; and make citation and usage metadata openly available for aggregation and auditing.

The NIH-supported Generalist Repository Ecosystem Initiative provides a practical foundation for this work. Its recommendations for collecting, exposing, and aggregating data citations could serve as a starting point. NIH could also draw upon the open Data Citation Corpus as an evidence and learning resource, while recognizing that citation coverage and metadata completeness continue to vary across repositories and disciplines.

3. Training and Mentorship: N/A

4. Collaboration: We encourage NIH to recognize the full range of roles involved in producing and sustaining reusable research data, including data creation, curation, documentation, quality assurance, stewardship, repository management, and support for responsible reuse. These contributions are essential to team science but are frequently invisible in publication-centered assessment.

Repositories and reporting systems could capture contributor roles in machine-readable metadata and connect contributors, datasets, grants, publications, software, and other outputs through persistent identifiers. Researchers could also be given opportunities to describe their specific data-related contributions in biosketches, applications, progress reports, and final reports.

Assessment could consider both structured contribution information and qualitative evidence of collaborations enabled by shared data. We caution against inferring individual credit solely from citation or download counts. Pilot programs could examine whether proposed indicators appropriately recognize team-based contributions and work performed by data stewards, curators, technical specialists, and other research professionals at different career stages.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: We encourage NIH to consider not only whether research data were accessed, but also how they were subsequently used. Evidence of public impact could include documented uses of data in clinical applications, public-health decisions, policy development, education, community resources, methodological improvements, and research that produces subsequent health or societal benefits.

NIH could support the development and adoption of contextual classifications, such as the Make Data

Count Typology of Data Uses, to distinguish among replication, validation, aggregation, methodological development, education, clinical application, and policy development. Connecting data citations and other usage evidence to these purposes could help NIH understand how its investments contribute to health, scientific progress, and public benefit.

Because many benefits emerge slowly or cannot be reduced to numerical counts, we recommend supplementing quantitative indicators with concise qualitative impact narratives. We also encourage recognition of controlled-access, specialized, and smaller-community datasets whose public benefit may be substantial despite relatively low citation or download totals.

8. Other Areas: Before data-related indicators are used in consequential funding or evaluation decisions, we recommend that NIH explore them through pilots across disciplines, career stages, repository types, Institutes and Centers, and institutional settings. These pilots could evaluate data quality and coverage, disciplinary differences, equity effects, recognition of team contributions, reviewer consistency, and the potential for gaming.

We encourage the involvement of researchers, data stewards, repositories, infrastructure providers, research administrators, evaluators, and affected communities in designing and evaluating these pilots. NIH could also support the open infrastructure and research needed to improve metadata completeness, establish appropriate contextual benchmarks, and make the resulting evidence transparent and auditable.

The objective would not be to replace publication-based metrics with another narrow set of counts, but to create a more complete, responsible, and transparent account of the many contributions through which modern biomedical science advances.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/NOT-OD-26-087-RFI-on-Measuring-and-Rewarding-Scientific-Impact.pdf

Description: Make Data Count response to NOT-OD-26-087, Request for Information on Measuring and Rewarding Scientific Impact. The attached letter provides recommendations for recognizing research data as first-class outputs; combining quantitative indicators with qualitative and contextual evidence; building on open standards and infrastructure; identifying the purposes and outcomes of data reuse; and piloting these approaches before using them in consequential decisions.

August 10, 2026

Re: NOT-OD-26-087, Request for Information on Measuring and Rewarding Scientific Impact

Dear NIH Office of Science Policy,

Make Data Count welcomes NIH's effort to identify better ways of measuring and rewarding scientific impact. Our response focuses on data sharing and reuse, an area in which NIH has an opportunity to recognize important research contributions that are largely invisible in current assessment systems.

Make Data Count is a global, community-led initiative that has been working to develop open, transparent, and meaningful measures of how research data are used since 2014. Following a rigorous selection process, the Global Sustainability Coalition for Open Science Services (SCOSS) selected Make Data Count as infrastructure essential to open science. Make Data Count has worked over the years with NIH on recommendations and implementations of best practices, including chairing a virtual conference on data metrics[1] and publishing in an NIH data policy special issue of Harvard Data Science Review[2]. Our central recommendation is that NIH treat data as first-class research outputs and explicitly recognize the creation, curation, documentation, sharing, stewardship, and reuse of data in funding and evaluation processes in ways that incentivize researcher behavior.

We offer the following recommendations:

1. **Recognize data contributions throughout the grant lifecycle**

We recommend that NIH enable researchers to identify datasets as distinct research contributions in applications, biosketches, progress reports, and final reports. These processes should recognize not only data producers, but also those who curate, document, steward, and enable the responsible reuse of data.

Evaluation should consider the quality and significance of the contribution, not simply whether data were made publicly available. Controlled-access data, sensitive human-subject data, dynamic datasets, and data created for small or specialized communities may have substantial value even when their use cannot be captured through conventional public-access measures.

2. **Combine quantitative measures with qualitative evidence and context**

[1] https://datascience.nih.gov/sites/g/files/mnhszr336/files/01_Opening_Remarks_Dawei_Lin_508.pdf
[2] Lowenberg, D. (2022). Recognizing Our Collective Responsibility in the Prioritization of Open Data Metrics. Harvard Data Science Review, 4(3). https://doi.org/10.1162/99608f92.c71c3479

No individual indicator can adequately represent the impact of a dataset. A portfolio approach combining the following would give NIH a more complete picture: :

- Quantitative measures, including standardized dataset citations, views, and downloads;
- Qualitative evidence, such as descriptions of replication, validation, new research, clinical application, policy use, education, or collaborations enabled by the data; and
- Contextual information about the discipline, dataset type, access conditions, intended users, career stage, and contributor roles.

Our preprint, *Make Data Count: Driving metrics for the meaningful evaluation of data*[3], describes why all three components are needed for data metrics to be meaningful, comparable, and trusted.

We caution against a composite data-impact score or use raw citation, view, or download counts to rank researchers. Such counts may provide useful evidence, but they should be interpreted within the relevant research and disciplinary context.

3. **Build upon open standards and infrastructure**

Indicators should be derived from transparent, auditable information rather than proprietary or opaque systems. NIH-supported repositories should be encouraged or required to:

- Assign persistent identifiers to datasets;
- Provide complete, machine-readable metadata, including contributor roles and relationships between datasets and other research outputs;
- Support structured citation of data;
- Collect views and downloads using common standards, such as the COUNTER Code of Practice for Research Data[4]; and
- Make citation and usage metadata openly available for aggregation, analysis, and auditing.

The NIH-supported Generalist Repository Ecosystem Initiative (GREI)[5] has already provided a practical foundation for this work. The GREI recommendations for collecting, exposing, and aggregating data citations[6] could serve as a starting point for broader NIH implementation. NIH can also draw upon the open Data Citation Corpus[7] to explore how NIH-funded data are cited and reused. The Corpus should initially be treated as an evidence and learning resource and not as a comprehensive or definitive ranking system because citation coverage and metadata completeness continue to vary across repositories and disciplines.

4. **Capture how data are used, not merely whether they are used.**

A citation, view, or download does not by itself explain the purpose or significance of data use. We encourage NIH to support the development and adoption of contextual classifications, such

---

[3] Puebla, I.& Chodacki, J. (2024). Make Data Count: Driving metrics for the meaningful evaluation of data. Zenodo. https://doi.org/10.5281/zenodo.14261211

[4] https://coprd.countermetrics.org/

[5] https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative

[6] https://makedatacount.org/read-our-blog/grei-data-citation-recommendations/

[7] https://makedatacount.org/find-a-tool/data-citation-corpus-documentation/

as the Typology of Data Uses[8], to distinguish uses including replication, validation, aggregation, methodological development, education, clinical application, and policy development. Capturing these different pathways would help NIH connect data reuse more directly to rigor, reproducibility, collaboration, translation, and public benefit.

5. **Pilot, evaluate, and improve these approaches before using them in consequential decisions**

We recommend that NIH begin with pilots across several Institutes, Centers, disciplines, career stages, repository types, and institutional settings. These pilots should assess:

- Whether the underlying information is sufficiently complete and reliable;
- Whether indicators disadvantage particular disciplines, institutions, career stages, or research communities;
- Whether measures appropriately recognize team-based contributions;
- Whether measures create incentives for gaming or superficial compliance; and
- Whether reviewers can interpret the information consistently and responsibly.

These pilots should involve researchers, data stewards, repositories, infrastructure providers, research administrators, evaluators, and affected communities in the design and evaluation of these pilots. It should also fund the open infrastructure and research necessary to improve coverage, establish appropriate benchmarks, and understand disciplinary differences.

Better recognition of data will strengthen incentives for rigorous data stewardship and sharing while helping NIH understand how its investments enable subsequent discovery and public benefit. The objective should not be to replace publication-based metrics with another narrow set of counts. It should be to create a more complete, responsible, and transparent account of the many contributions through which modern science advances.

Make Data Count would welcome the opportunity to share its experience, infrastructure, standards, and community resources as NIH develops and tests these approaches.

Sincerely,

[redacted]

Matt Buys
Executive Director, DataCite

John Chodacki
Director, UC Curation Center
California Digital Library

Britta Dreyer
Finance & Operations Director,
DataCite

Co-Chairs, Make Data Count Steering Committee
On behalf of the Make Data Count Initiative
info@makedatacount.org
makedatacount.org

[8] https://makedatacount.org/read-our-blog/typology-of-data-uses/

185. Digital Science & Research Solutions Inc.

I am responding: On behalf of an organization

Name: Digital Science

Name of Organization: Digital Science & Research Solutions Inc.

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Rigorous and reproducible science is most effectively assessed through observable and documented research practices rather than publication outcomes alone. This directly reflects the priorities set out in Executive Order 14303, Restoring Gold Standard Science, which identifies reproducibility, transparency, communication of uncertainty, and a willingness to subject scientific findings and assumptions to continued scrutiny as foundational tenets of trustworthy federal research.

Successful replication remains one of the strongest indicators of scientific robustness. However, replication is inherently retrospective, resource-intensive and feasible for only a fraction of published research. NIH should therefore complement direct replication with measurable indicators of the research practices and outputs that make findings independently assessable, reproducible, and ultimately replicable.

Use connected research to identify where replication is needed most
Persistent identifiers, interoperable metadata and connected research information have fundamentally changed what can be observed across the research lifecycle. Publications can increasingly be linked to grants, datasets, software, protocols, biological resources, organizations, researchers, and subsequent research outputs. This provides NIH with a substantially broader evidence base for assessing scientific rigor than publication counts and citation metrics alone. These evidence-based indicators reflect transparency, methodological rigor and the stewardship of research outputs.

This same connected evidence base can also help address a critical question: Where would additional replication effort have the greatest scientific value?
Rather than prioritizing replication primarily in response to highly visible failures or individual nominations, NIH could develop an evidence-based framework for identifying research areas where independent verification is most warranted. Such framework could combine multiple signals, including:
availability and accessibility of quality underlying datasets and analytical code;
availability of sufficiently detailed protocols and methods;
use of trusted repositories and persistent identifiers;
preregistration where scientifically appropriate;
transparent reporting of methods, including negative and confirmatory findings;
citation and demonstrated reuse of datasets, software and research resources; and
independent validation or replication of published results.

Natural language processing and structured metadata applied to the full text of published research can now detect the presence or absence of specific transparency markers at scale, including disclosed funding, stated ethical approval, data and code availability, declared competing interests, validated across multiple fields to avoid disciplinary bias. This allows NIH to move from a largely reactive and subjective model of replication toward a more systematic, evidence-informed model: identifying consequential findings or research areas that have limited independent verification, weaker reproducibility signals, or substantial downstream scientific dependence. NIH could use such evidence to help prioritize replication investments, including expansion of existing efforts to independently replicate significant lines of NIH-supported research.

Reward the work required to make science reliable
Identifying where replication is needed does not resolve why so little of it happens. Academic incentive structures are built almost entirely around individual achievement, including first or last authorship, novel-discovery outputs, principal investigator status, publication volume, grant funding, and citation-based indicators. These measures can undervalue the collaborative and often less glamorous and less visible work required to verify, reproduce, challenge, or strengthen existing findings. Researchers undertaking rigorous replication largely absorb the professional cost of that work personally, while the resulting benefit of a more reliable scientific evidence base accrues broadly to the field as a whole.

NIH can help change this incentive structure by making contributions to reproducibility and replication visible, attributable, and creditable. When replication studies are linked to the original finding through persistent identifiers a contributor's specific role in it, replication design, execution, data verification, methodology, software, resource development, and experimental execution are captured through structured contributor taxonomies such as CRediT, those contributions can become part of an investigator’s documented scientific record independent of authorship position high citation counts.

NIH could extend this further by formally recognizing verification and replication contributions in biosketches, progress reports, grant review criteria, and other assessment processes and by continuing to explore dedicated funding and recognition mechanisms for high-value replication research.

Measure practices and outcomes, not compliance alone
These indicators should prioritize evidence of accessibility, reuse, verification and scientific utility, rather than simple compliance with reporting requirements. A dataset that is discoverable, interoperable and reused by multiple research communities provides considerably stronger evidence of scientific value than one that is merely deposited.

Potential measurable indicators could include:
proportion of applicable research outputs with accessible underlying datasets;
proportion with accessible analytical code or computational workflows;
availability and completeness of protocols and methodological documentation;
use of trusted repositories and persistent identifiers;
preregistration where scientifically appropriate;
transparent reporting of methods, limitations, negative findings, and confirmatory findings;

demonstrated reuse of datasets, software, protocols, and research resources;
successful computational or methodological reproduction by independent researchers;
documented independent validation or replication;
links between original findings and subsequent replication or validation studies; and
continued accessibility of research outputs over time.

Create infrastructure for negative and confirmatory findings
Transparent reporting of negative and null findings, named explicitly above as an indicator, and named directly among the Gold Standard Science tenets as a marker of trustworthy research culture, depends on more than researcher willingness. It requires infrastructure through which these outputs can be persistently identified, discovered, cited, connected to the original research, and recognized as legitimate scientific contributions regardless of journal publication.

Repository infrastructure capable of assigning persistent identifiers to diverse research outputs can provide a durable mechanism for making negative confirmatory findings part of the scientific record. Connecting these outputs to grants, investigators, datasets, protocols and origination studies would also allow NIH to evaluate their subsequent attention, reuse, and scientific influence.

This is important because an incentive to report negative findings will have limited effect if those findings remain effectively invisible within research assessment systems.

Minimize reporting burden through interoperability
Implementation should build upon existing community standards, including DOIs, ORCID iDs, ROR identifiers, RRIDs and grant identifiers, and structured contributor roles. Wherever possible, NIH should derive indicators from existing research infrastructure rather than requiring investigators to repeatedly report on information already available elsewhere.
The objective should be to create a connected evidence layer around NIH-supported research, not another administrative reporting layer.

Protect the independence of replication review
Replication also presents distinct review-integrity considerations. A proposal to replicate an influential finding may be reviewed by researchers who collaborated with, trained with, competed with, or have other significant scientific relationships to investigators associated with the original work.
For replication-focused funding, NIH could therefore consider conflict-of-interest screening that examines relevant relationships to both the replication applicant and the research being replicated. Connected publication, funding, institutional, and collaboration information could assist NIH in identifying relationships requiring additional review while preserving appropriate scientific expertise on review panels.

Distinguish transparency, reproducibility, and replication
Finally, NIH should avoid collapsing transparency, reproducibility, and replication into a single composite measure.

They represent related but distinct dimensions of scientific quality:

Transparency asks whether the evidence, methods, materials, assumptions, and limitations necessary to evaluate the research are available.
Reproducibility asks whether the reported result can be obtained using the original data, methods, or analytical procedures, as appropriate to the research domain.
Replication asks whether a finding can be independently tested and supported using new data, experiments, samples, or other appropriate independent approaches.
Each provides different information about the reliability of a scientific claim. NIH assessment frameworks should preserve these distinctions while recognizing their complementary value.

Research assessment itself should also be treated as an evidence-based scientific endeavor. Any indicators NIH adopts should be evaluated for construct validity, disciplinary and methodological bias, susceptibility to gaming, administrative burden, and unintended behavioral consequences, and should evolve as evidence accumulates.

The opportunity for NIH is therefore broader than developing another set of research metrics. By connecting information already generated throughout the research lifecycle, NIH can build an evidence-informed system that helps determine which findings most warrant independent replication, which practices make research more reproducible, and which researchers and teams are contributing to the collective reliability of biomedical science. Such a framework would allow NIH to direct replication investment based on evidence rather than visibility while creating meaningful incentives for researchers to produce science that can be independently evaluated, reproduced, and trusted.

2. Data, Software, and Model Sharing: Scientific publications are increasingly one component of a much broader ecosystem of research objects. The publication should be viewed as an entry point into the research record, rather than the research record itself. Contemporary biomedical discovery depends upon interconnected datasets, software, computational models, protocols, workflows and other digital research assets that enable verification, reuse and cumulative knowledge generation.

NIH assessment frameworks should recognize these outputs as first-class scientific contributions whose value extends well beyond their role in supporting individual publications.

The central objective should not simply be to incentivize sharing, but to incentivize useful and reusable sharing. A binary measure of whether a dataset, software package, or model was deposited provides limited information about whether the resource can be discovered, understood, accessed, reused, and built upon by other researchers.
The more meaningful question is:
Did the shared research asset continue to generate scientific value after the originating project ended?
This distinction is important for maximizing the return on NIH's research investment.

Move from measuring deposition to measuring reuse
Persistent identifiers, interoperable metadata, repositories, and connected research information increasingly make it possible to observe how research objects move through the research ecosystem, from creation and deposit to discovery, citation, integration, reuse, and subsequent scientific outcomes.

NIH therefore has an opportunity to evolve from measuring whether sharing occurred toward measuring

whether shared assets are findable, accessible,understandable, reusable, and demonstrably used by others.

Potential measurable indicators could include:
deposition in appropriate trusted repositories;
assignment and continued resolution of persistent identifiers;
completeness and quality of metadata;
linkage to originating grants, publications, investigators, and institutions;
adherence to relevant FAIR principles and community standards;
availability of documentation sufficient to understand and reuse the resource;
software versioning, documentation, dependency information, and maintenance;
model documentation, including relevant assumptions, parameters, provenance, limitations, and versioning;
use of appropriate licenses or terms that clearly communicate conditions of reuse;
citation or formal acknowledgment of datasets, software, models, protocols, and other research assets;
evidence of reuse in subsequent publications or research projects;
reuse by investigators or institutions independent of the originating research team;
integration into subsequent datasets, software, models, clinical research, public resources, or other research infrastructure;
evidence of continued maintenance and accessibility over time; and
breadth and diversity of downstream reuse across research communities.

These indicators should not necessarily be combined into a single score. Different research objects and disciplines have different reuse patterns, and assessment should account for those differences.

Reward stewardship, not compliance alone
NIH should distinguish deposit, stewardship, and reuse.
Deposit establishes that an output has been made available.

Stewardship establishes that the output remains accessible, documented, interpretable, interoperable, and usable.
Reuse demonstrates that the output has generated value beyond the originating research project.
These are related but distinct dimensions of responsible research sharing.

A dataset deposited to satisfy a policy requirement but lacking sufficient metadata or documentation to permit meaningful reuse should not be treated as equivalent to a well-curated dataset that is independently reused across multiple research projects. Similarly, software that is deposited but cannot be executed because dependencies, documentation, or version information are absent provides different scientific value from software that remains maintained, documented, and adopted by other research teams.

NIH incentives should recognize this continuum.

Give researchers credit for creating reusable research assets

Researchers continue to receive substantially more professional recognition for publications than for the creation and stewardship of reusable research resources. This creates a structural incentive problem: producing high-quality datasets, maintaining software, documenting models, curating metadata, and supporting reusable resources require significant scientific effort, yet those contributions may remain largely invisible in traditional research assessment.
NIH can help change this by making these contributions visible and creditable.

Dataset, software, model, and resource contributions could be incorporated more explicitly into NIH biosketches, progress reporting, grant review, and other assessment processes. Where contributor roles can be reliably attributed, NIH could recognize activities such as data curation, software development, validation, visualization, methodology development, and resource stewardship alongside more traditional publication-based evidence of contribution.
Importantly, credit should not depend solely upon citation counts. Dataset and software citation practices remain inconsistent across disciplines. NIH should therefore consider multiple forms of evidence, including formal citation, persistent-identifier linkages, documented reuse, integration into downstream research, and other appropriate indicators of adoption.

Treat software and computational models as scientific methodology
Software and computational models warrant particular attention because they are increasingly integral to biomedical methodology.
Where software materially contributes to scientific findings, its documentation, versioning, provenance, dependencies, accessibility, and maintenance directly affect the ability of other researchers to evaluate or reproduce the work. Similarly, computational and AI/ML models may require documentation not only of the model itself but also of relevant training or validation data, versioning, parameters, assumptions, limitations, and appropriate conditions of use.

Assessment frameworks should therefore recognize software and model stewardship as part of scientific rigor rather than treating these outputs as secondary technical products.

NIH should also recognize that maintenance itself is a scientific contribution. A widely used software package or research resource can continue enabling NIH-funded science for many years after its original grant has concluded. Traditional assessment approaches may substantially underestimate this continuing return on federal investment.

Use connected evidence to understand downstream value
A major opportunity is the ability to connect shared research assets to subsequent scientific activity. Persistent identifiers and connected metadata can increasingly link a dataset or software resource to its originating grant and publication and then trace subsequent citations, publications, grants, investigators, institutions, clinical research, or other outputs that make use of it.

This creates the possibility of measuring research asset lineage: not simply whether an asset exists, but where it came from, who subsequently used it, and what additional scientific activity it helped enable.

For NIH, this could provide a new way to understand return on research investment. A dataset generated

through one NIH award may subsequently support research conducted by dozens of investigators, across multiple institutions and funding programs. Those downstream uses represent scientific value created by the original investment but are largely invisible in publication-centered assessment.

NIH could therefore consider indicators such as:
Reuse rate: whether and how often a resource is independently reused.
Reuse breadth: the number and diversity of investigators, institutions, disciplines, or projects using it.
Reuse longevity: whether the resource continues to generate research activity over time.
Research lineage: the subsequent publications, grants, clinical studies, resources, software, or other outputs connected to the original asset.
Independent reuse: whether reuse extends beyond the originating investigators or institution.
Together, these measures would provide a substantially richer view of scientific impact than deposit counts alone.

Build upon NIH's work on measuring data sharing
NIH has already explored this challenge through efforts such as the Data Sharing Index (S-Index) Challenge, which examined approaches for moving beyond a binary determination of whether data were shared toward more scalable measures of data sharing and reuse.

Experience from that effort demonstrates both the importance and feasibility of connecting repository information, persistent identifiers, research metadata, and downstream scholarly activity to better understand the lifecycle of shared data.

Digital Science's participation in NIH's S-Index Challenge demonstrated the practical feasibility of this approach. Our submission combined repository infrastructure supporting persistent identification and FAIR-aligned metadata with connected research information capable of tracing relationships among datasets, publications, grants, researchers, institutions, and subsequent research activity. Importantly, these capabilities were based on established infrastructure already operating at scale rather than systems created solely for the Challenge. That experience suggests that many of the indicators described above can be derived from existing infrastructure and metadata, rather than requiring NIH to create new reporting systems or impose additional reporting requirements on researchers.

Existing research infrastructure already indexes datasets distributed across many repositories and connects research objects with publications, grants, researchers, organizations, and other outputs. This provides a practical foundation for evaluating not only whether data have been deposited, but whether those assets are subsequently discovered, cited, reused, and connected to additional scientific activity.

NIH can therefore build upon existing community and commercial infrastructure, standards, persistent identifiers, and metadata to develop and validate indicators before introducing additional reporting requirements.

Reduce burden through interoperability
Implementation should prioritize interoperability and automation.
DOIs, ORCID iDs, ROR identifiers, grant identifiers, software identifiers, repository metadata, and other

established standards can provide much of the connective infrastructure required to understand research asset creation and reuse.

Where information can be derived reliably from existing systems, researchers should not be required to report it again.
This principle is particularly important if NIH expands assessment beyond publications. The goal should be to recognize a broader range of scientific contributions without creating a correspondingly broader administrative burden.

NIH should also recognize that no single indicator will adequately capture the value of all data, software, or models. Indicators should be validated for disciplinary differences, field maturity, resource type, susceptibility to gaming, and unintended incentives. Measures of reuse should complement, rather than replace, qualitative scientific judgment.

Ultimately, NIH has an opportunity to shift the incentive structure from "Did you share it?" to "Did you make it reusable, and did it continue to create value?"

That shift would better recognize the researchers who create and steward high-value scientific resources, provide NIH with a more complete picture of the return generated by federal research investments, and encourage a research ecosystem in which data, software, and models remain active scientific assets long after the projects that created them have ended.

3. Training and Mentorship: The continued strength of the biomedical research ecosystem depends not only on generating new knowledge but on developing the people, skills and research practices, and cultures that make future discovery possible. Training and mentorship should therefore be assessed not simply as activities that occurred, but by the scientific capacity they create over time.

Traditional indicators such as the number of trainees supported, fellowships completed, or training activities delivered provide important administrative information but only a partial measure of impact. They establish that training occurred; they do not establish what that training produced.

NIH has an opportunity to shift from measuring training activity toward measuring researcher development, mentorship effectiveness, and sustained scientific capacity.

Follow outcomes beyond the training period
The impact of NIH-supported training frequently becomes apparent years after an award or fellowship ends. Assessment should therefore follow researcher trajectories over longer periods rather than relying predominantly on outcomes observable within a single grant cycle.
Potential longitudinal indicators could include:
progression to subsequent career stages;
time to research independence, where independence is an appropriate career objective;
subsequent participation or leadership in funded research;
movement into scientific leadership positions;
continued participation in biomedical research across academia, healthcare, government, industry, nonprofit organizations, and other sectors;

participation in interdisciplinary and multi-institutional research;
development of sustained collaborative research networks;
subsequent contributions to data, software, methods, protocols, and shared research resources;
adoption and leadership of rigorous and reproducible research practices;
subsequent mentoring and supervision of other researchers; and
continued contribution to the biomedical research ecosystem over time.

Importantly, successful training should not be defined solely as progression toward a tenure-track faculty position or principal investigator status. NIH-supported researchers contribute to biomedical science through many pathways, including industry, clinical research, regulatory science, public health, government, research infrastructure, data science, software engineering, and other technical and collaborative roles.

An assessment framework that defines success too narrowly risks categorizing valuable scientific careers as attrition when they may instead represent substantial returns on NIH's training investment.

Measure mentorship by what it enables
Mentorship itself should be recognized as a scientific contribution.
Effective mentors do more than supervise research activity. They help researchers develop scientific judgment, methodological rigor, professional networks, collaborative skills, responsible research practices, and the ability to mentor others.

NIH could therefore consider indicators of mentorship effectiveness that examine the subsequent development of trainees over time. These might include trainee progression, continued participation in biomedical research, transition into appropriate leadership or independent roles, development of collaborative networks, contributions to rigorous and reproducible research, and evidence that former trainees subsequently become effective mentors themselves.
This creates an important potential indicator of long-term impact: mentorship propagation.

Rather than viewing the completion of an NIH-supported fellowship or training award as the endpoint of measurement, NIH could examine the longer arc of the investment:
NIH-supported training → mentorship → researcher development → continued contribution to the biomedical research ecosystem → creation of new research, data, software, methods, or other scientific resources → subsequent mentorship and development of additional researchers.

Viewed in this way, the impact of NIH-supported training is not limited to the career outcome of a single trainee. It can propagate through subsequent research contributions, collaborations, institutions, sectors, and generations of researchers. A training investment made in one individual may ultimately contribute to the development of an extended scientific network whose collective impact substantially exceeds what can be observed at the conclusion of the original award.

Persistent identifiers and connected research information increasingly make portions of this longitudinal pathway observable. By linking training awards and mentors to subsequent grants, research outputs, affiliations, collaborative networks, contribution roles, and later mentorship relationships, NIH could

begin to understand not simply whether an individual completed training, but how that investment propagated through the biomedical research ecosystem over time.

Such evidence should not be reduced to a simplistic mentor ranking or score. Trainee outcomes are influenced by discipline, institution, career goals, economic conditions, personal choices, and many factors beyond an individual mentor's control. Longitudinal evidence should therefore inform qualitative assessment rather than substitute for it.

Use longitudinal research information to understand career trajectories
Advances in research information infrastructure make it increasingly possible to understand researcher development over longer periods.
Persistent researcher identifiers, grant information, publications, affiliations, contribution metadata, and collaborative relationships can help connect an individual's research activity across career stages. With appropriate safeguards and validation, these data can provide NIH with a portfolio-level view of what happens to researchers following NIH-supported training.

For example, connecting persistent researcher information across successive grants, research outputs, affiliations, and sectors can help NIH understand whether trainees subsequently participate in academic research, industry, government, healthcare, public health, regulatory science, or other parts of the biomedical ecosystem; whether they develop sustained research collaborations; and where appropriate, whether they progress toward independent funding or scientific leadership.

This would allow NIH to ask questions that are difficult to answer from administrative completion data alone:
What kinds of careers do NIH-supported trainees build?
Which training environments consistently produce sustained participation in biomedical research?
Which programs produce researchers who subsequently develop other researchers?
Where does NIH-supported talent remain within the biomedical research ecosystem, and where is scientific capacity being lost?
Which training approaches are associated with stronger long-term outcomes across different career pathways?
The objective should not be to prescribe a preferred career destination, but to understand whether NIH training investments generate durable scientific capability.

Recognize the full range of contributions that sustain research careers
Many activities that underpin high-quality biomedical research, including mentorship, data stewardship, software development, research infrastructure, peer review, methodological support, and collaborative leadership remain less visible within publication and PI-centered assessment frameworks.
This can particularly disadvantage researchers whose scientific contribution occurs through technical, collaborative, or infrastructure leadership rather than through conventional principal investigator roles.

Structured contributor information and recognized contribution taxonomies can help make these roles more visible. Rather than inferring contribution predominantly from authorship position or PI status, NIH could recognize documented roles in areas such as data curation, software development, methodology,

validation, project administration, supervision, and other contributions essential to modern team science.

This is relevant to workforce stability because NIH cannot effectively understand retention if its assessment systems recognize only a subset of the careers through which people contribute to biomedical research.

A researcher who moves from an NIH-supported training program into scientific software development, data stewardship, research infrastructure, regulatory science, or another research-enabling role should not automatically disappear from NIH's understanding of the workforce simply because that individual does not subsequently become a PI.

Recognize contributions to training culture
Mentorship also occurs beyond traditional one-to-one supervisory relationships.
Researchers contribute to workforce development by creating openly available training resources, developing interdisciplinary training programs, teaching responsible research practices, supporting reproducibility initiatives, participating in peer review, building research communities, and creating shared technical or methodological resources.

Potential indicators could therefore include:
development and demonstrated use of openly available educational or training resources;
contributions to interdisciplinary or cross-institutional training programs;
leadership in research integrity, responsible conduct of research, and reproducibility initiatives;
documented mentorship and supervision contributions;
contributions to shared research methods, infrastructure, data, or software that enable other researchers;
sustained collaborative relationships involving junior researchers;
participation in peer review and other forms of scientific community stewardship; and
evidence that trainees subsequently contribute to mentorship and workforce development themselves.

Where possible, assessment should emphasize evidence of use and downstream impact, rather than simply counting activities. For example, an openly available training resource that is adopted across multiple institutions provides stronger evidence of training impact than the existence of the resource alone.

Build on existing information rather than creating new reporting burdens
Much of the information needed to develop these indicators already exists across research systems. Persistent identifiers and connected research information can increasingly link researchers with grants, publications, affiliations, research outputs, collaborative networks, and documented contribution roles over time. These data can complement NIH administrative information and targeted qualitative evidence to provide a richer understanding of researcher development.

The goal should be to derive longitudinal evidence from existing information wherever it can be done reliably, rather than requiring mentors and trainees to repeatedly report information NIH or the broader

research ecosystem already holds.

NIH should also be cautious about the limitations of these indicators. Persistent identifier coverage varies; affiliation and career information can be incomplete; some sectors produce fewer visible scholarly outputs; career trajectories are not linear; and quantitative indicators may systematically underrepresent important forms of scientific contribution.

For these reasons, longitudinal data, structured contribution statements, trainee and mentor perspectives, and narrative evidence should complement one another.

Measure durable scientific capacity
Ultimately, NIH should assess training and mentorship according to the durable scientific capacity created by its investment.
A successful training investment is not simply a completed fellowship. It is a researcher who develops skills, contributes meaningfully to biomedical science, participates in productive scientific networks, and potentially develops the next generation of researchers.

A successful mentor is not simply someone who supervises a large number of trainees. It is someone whose mentorship contributes to the development of capable researchers and strengthens the culture and practice of science.

And a stable biomedical workforce should not be defined solely by the number of researchers who become principal investigators. It should encompass the diverse scientific, technical, clinical, public-sector, and collaborative careers required for modern biomedical research.

By following researcher trajectories over time, recognizing the full range of scientific contributions, and examining how mentorship propagates through subsequent careers and generations, NIH can move from counting training activities to understanding whether its investments are building and sustaining the biomedical research capacity required for its mission.

4. Collaboration: Authorship remains one of the dominant mechanisms for recognizing scientific contribution, yet it captures only a fraction of what contemporary team-based science requires. As biomedical research becomes increasingly interdisciplinary, computational, collaborative, and cross-sector, NIH assessment frameworks should recognize collaboration as the integration of complementary expertise and contributions toward a scientific objective, not simply as co-authorship.

This distinction is increasingly important. Scientific teams now depend upon specialized contributions spanning experimental science, clinical expertise, statistics, data science, software development, methodology, research infrastructure, project leadership, community engagement, and other capabilities. Assessment systems that focus predominantly on first or last authorship risk overlooking many of the contributions upon which modern biomedical research depends.
NIH therefore has an opportunity to address two related but distinct questions:
What did each person contribute to the team?
and
Did the collaboration create scientific capability or value that the participants would have been unlikely

to achieve independently?
Both should inform the assessment and recognition of team science.

Recognize contribution, not simply authorship position
Author order is an imperfect proxy for scientific contribution and becomes increasingly inadequate as teams grow larger and roles become more specialized.
NIH could place greater emphasis on structured contributor information, including recognized taxonomies such as CRediT, to make individual roles within collaborative research more visible. Contributions such as methodology development, software, data curation, validation, statistical analysis, visualization, supervision, project administration, resource development, and other activities should be recognizable as scientific contributions in their own right where appropriate.

This would be particularly valuable for researchers whose contributions are essential to multiple projects but who may rarely appear as first or last author or principal investigator.

The objective should not be to replace authorship, but to supplement it with evidence that better reflects who contributed what.
Over time, structured contribution information could enable NIH to recognize patterns of scientific contribution across multiple projects. A researcher who repeatedly provides essential methodological, computational, clinical, data, or infrastructure expertise across collaborative programs may be making an important contribution to NIH's mission even if traditional publication metrics systematically understate that role.

Distinguish collaboration from scientific integration
NIH should also distinguish between the size of a collaboration and the quality or depth of scientific integration.
More investigators, institutions, disciplines, or sectors do not necessarily indicate more effective collaboration. Indicators that simply reward larger teams could unintentionally encourage collaboration for its own sake.

More meaningful evidence would examine whether collaborators combine complementary expertise, share research resources, create sustained intellectual or technical relationships, and produce outputs or capabilities that cross traditional disciplinary or organizational boundaries.

Potential indicators could include:
diversity and complementarity of documented contributor roles;
interdisciplinary participation appropriate to the scientific question;
sustained collaboration across projects or funding periods;
cross-institutional or cross-sector research relationships;
shared datasets, software, methods, protocols, infrastructure, or other scientific resources;
joint development of community research resources;
evidence of knowledge exchange across disciplines or sectors;
subsequent reuse or adoption of collaborative outputs;
development of new collaborations or research communities arising from the original team;

translation of research across scientific, clinical, public-sector, or commercial environments where appropriate; and
evidence that the collaboration creates capabilities or outputs beyond those attributable to a single participating group.

These indicators should complement expert assessment rather than be combined Mechanically into a collaboration score.

Understand collaboration as a network, not an event
Collaboration should also be assessed longitudinally.
A co-authored publication provides evidence that researchers worked together on the output. It provides considerably less information about whether the relationship developed into a sustained and productive scientific partnership.

Persistent identifiers and connected research information increasingly make it possible to examine collaborative relationships across grants, publications, datasets, software, clinical studies, organizations and other research activity over time.

This allows collaboration to be understood as a network that develops and evolves, rather than a characteristic of an individual publication.
For example, NIH could examine whether a funded interdisciplinary collaboration:
NIH investment → brings together complementary expertise → produces shared research and resources → develops sustained collaborative relationships → generates additional research or capabilities → expands into new institutions, disciplines, or sectors → produces downstream scientific or translational value.
This provides a richer conception of collaborative impact than counting co-authored publications or participating institutions.

It also allows NIH to distinguish between first-order and downstream effects of collaborative investments. The immediate output may be a publication, dataset, software resource, or clinical finding. The longer-term return may include a durable interdisciplinary network, shared infrastructure, subsequent jointly funded research, movement of methods between fields, or new collaborations that continue after the original award ends.

Evaluate whether funded collaborations become genuinely integrated
Connected research evidence could also help NIH evaluate collaborative programs, centers, and consortia at the portfolio level.
Relationships among funding, researchers, organizations, publications, datasets, software, research resources, and documented contributor roles can provide evidence about whether collaborations develop scientific connections beyond the administrative structure of the original award.

For example, NIH could examine whether participating groups subsequently produce joint outputs, share or reuse research resources, establish new cross-institutional relationships, develop interconnected research programs, or continue collaborating after the initial funding period.

This would not by itself determine whether a collaboration is scientifically successful; expert judgment remains essential. It could, however, help NIH distinguish between collaborations showing evidence of deep and sustained scientific integration and those showing comparatively limited evidence of interaction.
For large collaborative investments, this could provide valuable evidence for program evaluation and future funding decisions.

Recognize cross-sector collaboration
Team science increasingly extends beyond academic institutions.
Biomedical research may bring together academic researchers, healthcare organizations, government laboratories, industry, nonprofit organizations, patient or community groups, data and technology organizations, and other participants.

Assessment frameworks should recognize the value of these collaborations without assuming that all sectors produce the same kinds of outputs.
A successful academic-industry collaboration, for example, may generate software, data resources, methods, technologies, patents, clinical applications, or other outputs that are not adequately represented by publication counts. Similarly, collaborations involving healthcare or public-sector organizations may produce implementation, policy, clinical, or population-health effects that require different forms of evidence.

NIH should therefore assess collaboration according to the scientific objective and intended contribution of the participants rather than imposing a single publication-centered definition of collaborative success.

Use connected information to help form teams as well as assess them
The same research information that supports assessment can also help NIH facilitate future collaboration.
Disambiguated researcher profiles connected to publications, grants, research topics, methods, datasets, affiliations, and previous collaborations can help identify relevant expertise beyond established professional networks.

This creates an opportunity for NIH not simply to measure collaboration after it occurs, but to help identify complementary expertise before teams are formed.
At a portfolio level, such approaches could help identify researchers or institutions working on related problems from different disciplinary perspectives, reveal capabilities that are missing from existing research networks, and identify opportunities to connect investigators who might not otherwise encounter one another.

Similar information could assist with identifying potential reviewers with appropriate expertise while supporting conflict-of-interest screening, although such systems should augment rather than replace established NIH review procedures.

Minimize burden and avoid unintended incentives

Implementation should build upon persistent identifiers, structured contributor roles, existing grant and publication information, and interoperable research infrastructure wherever possible.

Researchers should not be required to produce extensive new collaboration reporting when relevant evidence can be derived reliably from information already generated through the research process.

NIH should also be cautious about creating incentives that reward collaboration numerically. Number of collaborators, institutions, disciplines, or countries may describe a research team but should not automatically be interpreted as evidence of quality.

Network-based measures should likewise be used carefully. Highly connected researchers are not necessarily the most valuable collaborators, and important specialists may contribute deeply to relatively small numbers of projects.

Quantitative evidence should therefore support expert assessment of contribution, complementarity, integration, durability, and scientific value, rather than become an end in itself.

Recognize collaboration as a scientific capability
Ultimately, NIH should move from treating collaboration primarily as an authorship pattern toward recognizing it as a scientific capability.
The relevant questions are not simply how many researchers participated or how many institutions appeared on an award. They are:
Were complementary capabilities brought together?
Can the contributions of individual team members be recognized?
Did participants genuinely integrate their expertise and resources?
Did the collaboration create something that would have been difficult to achieve independently?
Did the relationships, resources, or capabilities created through the collaboration continue to generate value?

A funding agency that cannot distinguish integrated collaboration from a shared cover page cannot reliably reward the former or understand the return generated by its collaborative investments.

By combining structured contribution information, persistent identifiers, longitudinal research information, network evidence, and expert scientific assessment, NIH can develop a more defensible framework for recognizing the individuals who make team science possible while also identifying the collaborative structures that produce genuine scientific integration and sustained biomedical value.

5. Entrepreneurship and Translation: Patents, licenses, company formation, and commercialization are among the most visible and readily measurable outcomes of biomedical innovation. They are important indicators, but they capture only a portion of the process through which NIH-supported discoveries ultimately improve human health.

NIH should therefore assess not only translational outcomes, but also contributions to translational pathways: the cumulative, often non-linear process through which knowledge moves from fundamental discovery toward clinical, commercial, public-health, policy, or other forms of application.

This distinction matters because biomedical translation rarely results from a single publication, investigator, institution, or funding award. It frequently develops through successive contributions from researchers, organizations, funding programs, healthcare systems, government, industry, and other participants over extended periods.
A framework that recognizes those contributions would better reflect how biomedical innovation actually occurs and allow NIH to understand more fully how its investments contribute to downstream health, economic, and societal value.

Measure pathways, not simply endpoints
Traditional translational indicators tend to emphasize observable endpoints: patents, licenses, spinouts, products, or clinical applications.
NIH should retain these measures while also examining the intermediate steps through which research progresses toward application.

Potential indicators could include:
progression of NIH-supported research into subsequent translational or development funding;
patent applications and granted patents linked to NIH-supported research;
licensing activity and technology transfer, where appropriate;
formation of companies or other commercialization vehicles;
industry partnerships and cross-sector research relationships;
progression into clinical studies or clinical trials;
incorporation into clinical guidelines or standards of care;
citations or incorporation into regulatory documents;
incorporation into public policy or public-health guidance;
development and adoption of diagnostic tools;
development and adoption of research or clinical software;
development of standards, methods, protocols, or shared research resources;
use within healthcare systems or clinical practice;
development of public-health interventions;
subsequent investment or funding that builds upon the original research; and
demonstrated adoption or use by organizations beyond the originating research team.

These indicators should be interpreted according to the scientific field and intended pathway to impact. Not all valuable NIH-funded research should result in a patent, company, clinical trial, or commercial product.

Understand translation as a connected pathway
Advances in connected research information increasingly make it possible to observe translational pathways empirically.
Persistent identifiers and linked research information can connect publications with grants, patents, clinical trials, datasets, software, researchers, organizations, policy documents, clinical guidelines, and other downstream activity.

Rather than viewing these outputs independently, NIH could increasingly examine the pathways connecting them:
NIH investment → fundamental discovery → subsequent research and validation → translational development → cross-sector engagement → clinical testing or other application → adoption → health, economic, or societal value.

Importantly, these pathways will not always be linear.

A fundamental discovery may contribute to multiple subsequent research programs. A dataset may enable a diagnostic technology. Software may move directly into clinical research without generating a patent. A methodological advance may be incorporated into clinical guidelines. An NIH-supported discovery may contribute to a commercial product years later through research conducted by investigators and organizations that were not part of the original award.
Understanding these connections provides a more realistic model of translation than attempting to attribute an eventual societal outcome to a single publication or investigator.

Distinguish contribution from attribution
NIH should be cautious about attempting to assign exclusive credit for downstream impact.
Translation is cumulative. A successful therapeutic, diagnostic, clinical practice, or public-health intervention may depend upon decades of foundational research funded through multiple programs and performed across numerous institutions.
The appropriate question is often not: “Which investigator caused this outcome?” but rather: “What NIH-supported research contributed to the pathway that made this outcome possible?”

Connected evidence can help identify these contributions without implying simplistic causality.
This approach would also allow NIH to recognize foundational research that creates important translational value even when the originating investigators never participate directly in commercialization.

Identify translational inflection points earlier
The same evidence used to evaluate translation retrospectively could potentially help NIH identify emerging translational opportunities prospectively.
Today, important translational signals may appear across disconnected systems. A body of NIH-supported research may begin generating patent activity, follow-on funding, industry collaboration, clinical investigation, new software or diagnostic applications, or increasing attention from organizations outside the originating research community.

Individually, these signals provide incomplete information. Considered together, they may indicate that a line of research is approaching a translational inflection point.
NIH could explore whether combinations of signals such as: “research maturation + independent validation + follow-on funding + patent activity + cross-sector engagement + clinical activity” can help identify research areas where targeted translational support, expertise, partnership, or additional funding might accelerate progress.
This would move the assessment beyond asking “What translated?” toward also asking “What may be

ready to translate next?”
Such approaches should be validated carefully and used to support, rather than automate, scientific and programmatic judgment. Translational potential is inherently uncertain, and indicators should not create pressure to commercialize research prematurely.

Recognize the people who enable translation
Entrepreneurial and translational activity should also be recognized as a legitimate scientific contribution.
Researchers may contribute to translation through invention, company formation, technology development, industry collaboration, clinical development, standards development, software creation, regulatory science, implementation, or other activities that move knowledge toward use.
These contributions can be difficult to see in publication-centered assessment.

NIH could recognize documented translational contributions within biosketches, progress reports, portfolio assessment, and other appropriate review processes. Structured contributor information and connected evidence could help make these activities more visible.

Prior entrepreneurial and translational experience may also provide useful contextual evidence, for example, previous involvement in technology transfer, industry collaboration, clinical development, licensing, or company formation. However, such experience should not become a prerequisite or simplistic score for evaluating future potential.

Researchers at institutions with mature technology-transfer ecosystems have substantially different opportunities from researchers at smaller or less-resourced institutions. Similarly, some biomedical fields offer clearer commercial pathways than others. Assessment should recognize entrepreneurial contribution without creating a structural advantage for investigators who already have access to established commercialization networks.

Recognize multiple routes to impact
Commercialization is one route through which NIH-supported research creates value, but it is not the only one.
Translation may occur through clinical practice, healthcare delivery, public-health interventions, regulatory science, government policy, clinical guidelines, standards, open-source software, shared research infrastructure, diagnostic methods, or widespread adoption of a new research technique.

NIH should therefore avoid equating entrepreneurship exclusively with company formation or translation exclusively with commercialization.
Different types of research should be assessed against pathways appropriate to their intended contribution.

This is particularly important for foundational research. Basic science may have enormous long-term translational significance despite having little immediate commercial potential. An assessment framework overly weighted toward near-term translation could inadvertently discourage the fundamental discovery upon which future biomedical innovation depends.

Measure time to translation carefully
Translation also operates across very different timescales.
Some computational tools or methods can move rapidly into use. Therapeutic development may take many years. Foundational discoveries may generate little visible translational activity for decades before becoming critical to a new technology or treatment.

NIH should therefore consider time-to-translation and translational persistence as contextual rather than purely performance indicators.
Longitudinal analysis can help NIH understand how different types of research move toward application, where pathways commonly stall, and which interventions are associated with successful progression.

This could help answer important portfolio-level questions:
Where do promising NIH-supported discoveries most commonly stall?
Which types of support are associated with successful movement between translational stages?
Where are collaborations with industry, healthcare, or other sectors accelerating translation?
Which foundational investments ultimately contribute to multiple downstream innovations?
Which NIH programs are generating translational value that conventional publication metrics fail to capture?

Use connected evidence for portfolio intelligence
At sufficient scale, these indicators could provide NIH with more than an assessment framework. They could provide portfolio intelligence.

Connected evidence across grants, publications, patents, clinical trials, organizations, researchers, and downstream outputs could help program officers and institute leadership identify emerging areas of translational activity across an entire portfolio rather than relying exclusively on individual investigators or technology-transfer offices to surface opportunities.

This could help NIH identify clusters of related discoveries, emerging cross-sector engagement, technologies progressing toward clinical investigation, and areas where additional translational support may be warranted.

It could also help NIH understand the comparative pathways through which different institutes, programs, fields, or funding mechanisms generate downstream value.
The objective should not be to predict commercial success. It should be to provide decision-makers with additional evidence about where translation appears to be occurring, where it may be emerging, and where potentially valuable research is encountering barriers.

Build upon existing research information infrastructure
Many of the relationships required for this type of assessment can already be observed through existing research information infrastructure.
Large-scale connected research databases link publications with grants, patents, clinical trials, researchers, institutions, and other research outputs, making it increasingly possible to examine

translational relationships across entire research portfolios rather than reconstructing them manually one project at a time.

Digital Science's experience connecting grants, publications, patents, clinical trials, researchers, and organizations demonstrates the practical feasibility of examining these pathways at research-system scale. Dimensions, for example, links research activity across these different stages, allowing relationships between foundational funding, subsequent research, patent activity, and clinical investigation to be examined within a common evidence environment. This type of existing commercial infrastructure provides NIH with an opportunity to test and validate translational indicators using evidence already available today, rather than first constructing an entirely new reporting or tracking system.

This infrastructure provides a practical starting point for NIH to test translational indicators, examine research pathways retrospectively, and evaluate whether combinations of signals can reliably inform prospective portfolio analysis.

Implementation should build upon existing identifiers, standards, and data sources wherever possible, supplementing these with NIH administrative information and expert assessment rather than creating additional investigator reporting requirements.

Assess contribution while preserving discovery
Ultimately, NIH should expand its assessment of entrepreneurship and translation from counting visible endpoints toward understanding the pathways through which federally funded knowledge becomes useful.

The relevant questions are not simply:
Was there a patent?
Was a company formed?
Was a product commercialized?

They also include:
What research made the eventual application possible?
Who contributed to moving that research toward use?
What intermediate steps indicate meaningful translational progress?
Where are promising discoveries encountering barriers?
Can NIH recognize emerging translational opportunities early enough to help accelerate them?

A discovery that becomes visible as translational only when it reaches a technology-transfer office may already have traveled a substantial portion of its translational pathway.

By connecting evidence across the research lifecycle, NIH can develop a more complete understanding of how its investments contribute to innovation, identify translational momentum earlier, recognize the researchers and teams who enable it, and support pathways that deliver sustained health, economic,

and societal value, while continuing to protect the foundational science upon which future translation depends.

6. Foundational Scientific Exploration: Any indicator built around short-term productivity will systematically penalize the research most likely to produce paradigm-shifting results. NIH's challenge is therefore not to predict which individual foundational or high-risk research project will ultimately transform a field, but to protect sufficient space for exploration within a research portfolio that also rewards more predictable scientific progress.

Transformative advances frequently arise from fundamental research, enabling technologies, methodological innovation, exploratory investigation, unexpected findings, and new combinations of ideas that may initially appear incremental, uncertain, or disconnected from immediate application.

These contributions expand the scientific landscape by creating new questions, methods, resources, and capabilities rather than necessarily producing immediate translational outcomes.

Assessment of foundational scientific exploration should therefore recognize both the contributions created by exploratory research and NIH's broader success in maintaining a portfolio in which scientific risk and uncertainty can be sustained.

Do not require exploratory research to predict its own impact
One of the fundamental difficulties in assessing foundational research is that its eventual importance may be unknowable when the research is proposed or even when the original award concludes.

A discovery, method, dataset, biological resource, or enabling technology may acquire significance only years later, when another scientific development makes it newly useful. Research initially regarded as highly specialized may become foundational to an emerging field. A negative result may eliminate an unproductive direction and redirect subsequent research. A tool created for one purpose may eventually enable discoveries its creators never anticipated.

NIH should therefore avoid assessment frameworks that require foundational research to demonstrate short-term downstream impact as evidence of success.
Instead, potential indicators could include:
development of novel methodologies, computational approaches, or instrumentation;
creation of foundational datasets, biological resources, software, or other enabling infrastructure;
introduction of new concepts, hypotheses, or research questions;
development of novel combinations of methods or disciplines;
establishment of new interdisciplinary research directions;
production and dissemination of negative or null findings that meaningfully refine scientific understanding;
subsequent adoption of methods, resources, or concepts by researchers outside the originating team;
diffusion of methods or resources across disciplines;
emergence of new research communities building upon foundational work; and
evidence, over longer timescales, that exploratory work enabled subsequent scientific advances.

These indicators should describe contribution and influence without implying that every exploratory project must ultimately generate a transformative outcome.

Evaluate foundational contribution over longer timescales
Foundational research frequently operates on timescales poorly aligned with conventional research assessment.
By the time the importance of a foundational contribution is apparent through citation counts, clinical applications, patents, or other downstream indicators, the funding decisions that enabled it may have occurred many years earlier.

Longitudinal research information provides an opportunity to examine these contributions differently. Persistent identifiers and connected evidence can increasingly link grants, publications, methods, datasets, software, patents, clinical studies, researchers, institutions, and other outputs over extended periods. This can help reveal how an initially exploratory investment subsequently contributes to scientific capability.

For example:
NIH exploratory investment → new concept, method, resource, or capability → adoption by other researchers → diffusion across research communities → new questions and combinations → subsequent discoveries → potentially transformative scientific or translational impact.

The important point is that the original project's value should not be judged solely by whether it reaches the end of this pathway.
Creating a method or resource that opens new scientific possibilities may itself represent a valuable outcome of foundational research.

Identify emerging science before conventional metrics can see it
Connected research evidence also creates an opportunity to move beyond retrospective assessment.

Citation-based measures are inherently lagging indicators. By the time an emerging research area accumulates substantial citation activity, its importance may already be widely recognized.

NIH could complement conventional assessment with horizon-scanning approaches designed to identify nascent scientific directions while they are still forming.
Potential signals could include:
emergence of new concepts or terminology;
rapid development or adoption of new methods;
unusual combinations of previously separate research areas;
growth of new interdisciplinary research networks;
migration of methods or technologies between fields;
increasing creation or reuse of specialized datasets, software, or research resources;
formation of new clusters of investigators around emerging questions; and
changes in funding, publication, or collaboration patterns that suggest the development of a new research direction.

Natural language processing, topic analysis, network analysis, and connected research information can help identify such patterns across large bodies of scientific activity.
These signals should not be interpreted as automatic indicators of scientific quality or future breakthrough potential. Emergence is not synonymous with importance, and novelty is not synonymous with scientific merit. Expert scientific judgment remains essential.

Their value is instead in helping NIH see where the scientific landscape may be changing early enough to investigate further.

Use horizon scanning to inform funding strategy
Applied prospectively, these approaches could provide NIH with a form of scientific horizon scanning.

Rather than waiting until an emerging field has accumulated years of conventional evidence, program officers and institute leadership could receive evidence about where new concepts, methods, technologies, interdisciplinary combinations, and research communities appear to be forming.

This could help NIH ask:
Which potentially important scientific directions are emerging?
Which established fields are being disrupted by new methods or technologies?
Where are previously separate disciplines beginning to converge?
Which enabling technologies are beginning to diffuse across biomedical research?
Where is scientific activity emerging faster than NIH investment?
Are there potentially important areas of exploration that fall between existing programmatic structures?
The objective should not be to automate funding decisions or predict the next breakthrough. It should be to expand the evidence available to scientific leaders deciding where additional exploration may be warranted.

Assess NIH's portfolio risk, not just investigator risk
Perhaps the more important opportunity is to apply these approaches to NIH's own funding portfolio.

High-risk, high-reward research is inherently uncertain at the individual-project level. Some projects will fail to produce their intended results. That is a characteristic of genuine scientific risk, not necessarily evidence of poor funding decisions.

Assessment is therefore more meaningful at the portfolio level.

NIH could examine the distribution of investment across established, developing, and emerging research directions and how that distribution changes over time.
Potential portfolio-level indicators could include:
concentration of funding within established research areas;
proportion of funding supporting emerging or exploratory areas;
entry of new investigators or disciplines into established research fields;
funding for novel combinations of methods or disciplines;

diversity of scientific approaches supported within a portfolio;
changes in the distribution of funding as scientific fields mature;
persistence of support for exploratory areas during periods of uncertainty;
subsequent development of research communities around previously exploratory investments; and
evidence that earlier exploratory investments created methods, resources, or capabilities subsequently used across NIH-supported science.
These measures could help NIH evaluate an important question that cannot be answered by examining individual grant productivity:
Is NIH maintaining a research portfolio with sufficient capacity for scientific exploration and responsible risk-taking?

Detect portfolio drift toward predictability
Portfolio analysis could also help identify risk drift.
Funding systems can gradually become more concentrated around established investigators, methods, topics, or scientific approaches even without an explicit decision to become more conservative.

Each individual funding decision may be scientifically defensible while the cumulative portfolio nevertheless becomes increasingly concentrated around areas with established evidence, established communities, and predictable outputs.
NIH could therefore examine changes in portfolio concentration and scientific novelty longitudinally.

For example: Portfolio at time 1 → distribution across established and emerging areas → subsequent funding decisions → changing concentration of investment → portfolio at time 2.
This would allow NIH to examine whether the agency, an institute, or a program is maintaining its intended balance between exploitation of established scientific opportunity and exploration of less certain directions.

Importantly, such analysis should not assume that a particular percentage of funding to emerging areas is inherently optimal. Appropriate risk tolerance will vary across institutes, programs, scientific opportunities, and public-health needs.
The objective is to make the portfolio's risk characteristics visible enough to be managed intentionally.

Learn from unsuccessful high-risk research
NIH should also distinguish between scientific failure and unproductive investment.

A high-risk project that does not confirm its original hypothesis may nevertheless produce important negative findings, eliminate an unproductive scientific direction, generate a useful method or dataset, reveal methodological limitations, or create knowledge that informs subsequent research.

If high-risk programs are assessed primarily according to whether projects achieve their original anticipated outcomes, researchers will have a rational incentive to propose projects whose outcomes are more predictable.

NIH should therefore recognize informative failure as a legitimate potential outcome of exploratory

research, provided the work was rigorous and the resulting knowledge is made available to the scientific community.

This is also why transparent reporting of negative and null results, addressed elsewhere in this RFI, is particularly important for high-risk research.

Protect unconventional combinations of expertise
Transformative research may also arise at the boundaries between established disciplines.
Connected research information can help identify unusual combinations of expertise, methods, concepts, or research communities that differ from established collaboration patterns.
NIH could examine whether funding mechanisms intended to support innovation are actually enabling these unconventional combinations or predominantly supporting teams whose structures resemble already successful research networks.

Again, unusual combinations should not automatically be rewarded. Novelty without scientific rationale has little value.
But the ability to observe disciplinary and methodological combinations provides NIH with another way to assess whether programs intended to encourage convergence and exploration are actually broadening the scientific landscape.

Build on existing evidence infrastructure
Many of the signals required for this type of analysis already exist within research information systems.
Large-scale connected evidence across grants, publications, researchers, organizations, datasets, software, methods, patents, clinical research, and other outputs can support longitudinal analysis of emerging fields, knowledge diffusion, interdisciplinary convergence, and funding concentration.
Natural language processing and network analysis can further help identify changes in concepts and relationships that are difficult to detect through traditional publication and citation measures alone.

NIH could use this existing evidence base to test and validate horizon-scanning and portfolio-risk indicators retrospectively before applying them prospectively.
For example, NIH could examine historical research areas now recognized as transformative and ask: Would the proposed indicators have detected their emergence before conventional metrics did?

Similarly, NIH could examine previous high-risk funding portfolios to determine whether proposed measures meaningfully distinguish exploratory investment from established research.

This type of retrospective validation would be important before incorporating new indicators into funding strategy.

Protect the conditions for discovery
Ultimately, foundational scientific exploration presents a different assessment problem from research intended to produce predictable near-term outcomes.
The objective should not be to develop a metric that predicts which high-risk project will succeed. Doing so would risk eliminating the uncertainty that makes the research high-risk in the first place.

Instead, NIH should ask:
Is exploratory research creating new scientific possibilities?
Can NIH identify emerging directions early enough to investigate them?
Is the agency maintaining sufficient diversity of scientific approaches?
Are programs intended to support high-risk research actually funding research that differs meaningfully from established patterns?
Is NIH's portfolio gradually becoming more conservative without intending to do so?
Does the system recognize rigorous, informative failure as part of scientific exploration?
A portfolio that cannot see its own risk profile cannot knowingly protect it.

By combining longitudinal research information, horizon scanning, portfolio analysis, and expert scientific judgment, NIH can make the conditions for exploration more visible without attempting to make uncertain science artificially predictable. This would help preserve the scientific risk, curiosity, and methodological innovation from which future paradigm-shifting discoveries are most likely to emerge.

7. Public Impact: Public impact is too often measured by visibility rather than uptake. Media attention is not the same as influence on clinical practice or policy. NIH's assessment framework should draw that distinction explicitly.

Scientific knowledge creates public value through diverse and interconnected pathways, including clinical practice, public health, regulatory science, public policy, education, industrial innovation, healthcare delivery, accessibility of knowledge, and public engagement. These pathways often interact, reinforce one another, and develop over extended periods.

NIH should therefore assess public impact as a pathway from attention to uptake, application, and ultimately societal benefit, rather than attempting to represent it through a single impact metric.

Distinguish attention, uptake, application, and outcome
Different indicators provide evidence of fundamentally different things.
Attention indicates that research has been noticed or discussed.

Uptake indicates that research has been incorporated into a professional, scientific, policy, clinical, educational, or other decision-making context.
Application indicates that knowledge has influenced a practice, intervention, technology, policy, guideline, or other activity.
Outcome indicates evidence of resulting health, economic, accessibility, or broader societal benefit.

These should not be treated as equivalent.
For example: NIH-supported research → publication → clinical guideline citation → adoption in clinical practice → change in patient care → improved health outcome.
Similarly: NIH-supported research → evidence base → public-health guidance → implementation → population-level intervention → measurable health benefit.
Or: NIH-supported research → enabling discovery → technology development → commercial adoption → economic activity → broader health or societal benefit.

A media mention may provide evidence of public attention. A citation in a clinical guideline provides stronger evidence that research has entered professional decision-making. Demonstrated implementation of that guideline represents a further stage of impact. Evidence that implementation contributed to improved outcomes represents something different again.

NIH should preserve these distinctions rather than collapsing them into a single measure of "impact."

Measure multiple pathways to public value
Because NIH supports an extraordinarily diverse biomedical research portfolio, no single pathway will adequately describe public impact.
Potential measurable indicators could include:

Clinical impact
incorporation of NIH-supported research into clinical guidelines or standards of care;
citations in clinical decision-support resources;
progression of research into clinical studies or trials;
adoption of evidence-based interventions within healthcare systems;
development or adoption of diagnostics, therapeutics, devices, software, or other clinical tools;
changes in clinical practice associated with research findings; and
where evidence permits, contribution to improved patient outcomes.

Public-health and policy impact
incorporation into public-health guidance;
citation or use in federal, state, local, or international policy documents;
incorporation into regulatory science or regulatory decision-making;
contribution to standards or technical guidance;
adoption within public-health programs or interventions; and
evidence of changes in population-health practice associated with the research.

Economic and innovation impact
patents and licensing associated with NIH-supported research;
company formation and entrepreneurial activity;
industrial research partnerships;
subsequent private-sector investment;
development of products, technologies, services, or research tools;
employment or economic activity associated with NIH-supported innovation; and
downstream commercial use of NIH-supported knowledge or resources.

Accessibility and knowledge diffusion
public accessibility of NIH-supported research outputs;
availability and reuse of datasets, software, protocols, models, and other research resources;
use of research in educational or professional training materials;
adoption of research resources by communities beyond the originating investigators;
accessibility of findings to patients, practitioners, policymakers, and other intended audiences; and

sustained availability of publicly funded research resources.

Public engagement and trust
evidence of responsible communication of scientific findings, including uncertainty and limitations;
engagement with patients, communities, and relevant public stakeholders;
use of NIH-supported evidence in trusted public-information resources;
transparent correction or updating of scientific findings as evidence changes; and
where methodologically defensible, changes in understanding or confidence associated with sustained engagement.
These categories should remain distinct. A research project should not be expected to demonstrate impact across all of them.

Measure contribution, not simplistic attribution
NIH should distinguish between attribution and contribution when assessing public impact.

Most improvements in health and society cannot credibly be attributed to a single publication, investigator, grant, or institution. A change in clinical practice may depend upon decades of fundamental research, validation studies, clinical trials, systematic reviews, professional guidelines, healthcare implementation, and public-health activity.

The appropriate question is therefore often not: "Did this NIH award cause this societal outcome?" but: "What role did NIH-supported research play in the pathway that contributed to this outcome?"
Connected research evidence can help reconstruct these pathways without implying causality that the evidence cannot support.
For example: multiple NIH investments → body of evidence → clinical synthesis → professional guideline → healthcare implementation → patient benefit.
This provides a more scientifically defensible account of public impact while recognizing the cumulative nature of biomedical progress.

Trace public impact longitudinally
Public impact frequently occurs years or decades after the original research is funded.
Assessment should therefore connect evidence across longer timescales rather than requiring investigators to demonstrate societal impact within the period of an individual award.

Persistent identifiers and connected research information increasingly make it possible to link grants, publications, datasets, software, patents, clinical trials, policy documents, clinical guidelines, researchers, and organizations over time.

This creates an opportunity for NIH to examine impact pathways across its portfolio.

Rather than relying predominantly on individual success stories, NIH could ask:
Which NIH-supported research contributes to clinical guidelines?
Which research areas are subsequently reflected in public-health or regulatory decision-making?
Which NIH investments contribute to technologies or industries?

Which datasets, software, and research resources continue to create public value after the originating grant concludes?
How long does it typically take different forms of NIH-supported research to move from discovery to uptake and application?
Where does research appear to reach an uptake stage but fail to progress toward implementation?
This would provide a more systematic understanding of the return generated by NIH investment.

Assess impact at the portfolio level
Portfolio-level assessment is particularly important for public impact.
A single project may contribute only one component of a much larger pathway. At the level of an institute, disease area, program, or NIH as a whole, however, the cumulative relationship between federal investment and downstream activity becomes more visible.

NIH could therefore develop public-impact profiles for different portfolios rather than attempting to reduce all impact to one score.
For example, one portfolio may demonstrate substantial influence on clinical practice. Another may primarily create foundational resources used throughout biomedical research. Another may generate significant economic and industrial activity. Another may contribute heavily to public-health policy.

These are different forms of mission value, not necessarily different levels of value.
This approach would also enable NIH to demonstrate impact consistently across Institutes and Centers and research areas rather than depending primarily on highly visible individual examples.

Treat public trust as an outcome requiring particular care
Public trust warrants special treatment because it is both central to NIH's mission and particularly difficult to measure.
NIH has explicitly connected rigor and reproducibility with strengthening confidence and trust in biomedical research.
However, public trust should not be reduced to popularity, positive sentiment, media volume, or social-media engagement.

Trustworthy science and popular science are not necessarily the same thing.

More defensible indicators may include evidence of:
transparent communication of methods, limitations, and uncertainty;
accessibility of underlying research and supporting evidence;
correction of the scientific record when findings change;
responsible communication of negative or conflicting findings;
meaningful engagement with affected communities;
accessibility of research findings to intended audiences; and
demonstrable use of NIH-supported evidence by trusted clinical, public-health, and professional institutions.

Where surveys or public-opinion measures are used, they should be interpreted carefully and should not

create incentives for researchers or NIH to optimize scientific communication for popularity rather than accuracy.

NIH's own current replication and reproducibility efforts recognize that rigorous, transparent, reliable science is foundational to public confidence. Public-impact assessment should reinforce that relationship.

Distinguish public communication from public impact
Media, news, online attention, and other public-facing sources can provide useful evidence, but they should be interpreted according to what they actually measure.
Media attention can identify research entering public discourse.

It does not establish that the research changed policy, clinical practice, health outcomes, or public understanding.
NIH could therefore use public-attention indicators as pathway signals rather than endpoint measures. For example: “research → public attention → stakeholder engagement → policy consideration → policy adoption” may represent a legitimate pathway worth examining.
But attention alone should not be rewarded as equivalent to demonstrated uptake.

This distinction would also reduce incentives for sensationalism and help ensure that responsible communication of uncertainty remains compatible with NIH's impact framework.

Identify where evidence fails to become impact
One of the most valuable applications of a connected public-impact framework may be identifying translation and implementation gaps.

NIH traditionally asks whether research generated important knowledge. Connected evidence could increasingly help it ask what happened next.
For example:
strong scientific evidence → no guideline uptake
guideline incorporation → limited clinical adoption
effective intervention → limited accessibility
valuable public-health evidence → inconsistent implementation

These gaps matter because generating knowledge is only one part of delivering value to the American public.

NIH could use portfolio-level analysis to identify areas where strong research evidence exists but downstream uptake appears limited, providing an evidence base for considering whether additional implementation research, dissemination, partnerships, accessibility efforts, or other interventions may be warranted.
This would transform public-impact assessment from a retrospective reporting exercise into a tool for identifying where additional NIH action could help research generate greater public value.

Build upon existing evidence infrastructure
Many components required for this framework can already be observed through existing research information infrastructure.

Research can increasingly be connected across grants, publications, clinical trials, patents, policy documents, clinical guidelines, researchers, institutions, and public-facing sources. Combining these evidence streams provides a practical starting point for reconstructing pathways from federal investment toward downstream uptake.

Digital Science's experience working with research funders and institutions demonstrates the feasibility of this type of cross-source analysis at portfolio scale. Connected research information can link funding and scholarly outputs with downstream evidence such as patents, clinical activity, policy references, and public attention, helping organizations move beyond individual impact anecdotes toward systematic evidence of where research is being taken up and used.

This type of infrastructure could provide NIH with an existing evidence base for testing and validating public-impact indicators across different Institutes, Centers, disease areas, disciplines, and timescales rather than requiring a new reporting system to be created before assessment can begin.

Avoid creating an "impact burden"
NIH should be cautious about requiring every investigator to produce an extensive narrative demonstrating public impact.
Such an approach could create substantial administrative burden, advantage investigators with access to professional communications or research-impact staff, and encourage overstatement of downstream effects.

Where possible, NIH should derive evidence from existing research information, persistent identifiers, administrative systems, and downstream sources and allow investigators to provide contextual or qualitative evidence where automated information is incomplete.

Assessment should also recognize differences among disciplines, institution types, career stages, and research objectives. Early-career researchers and investigators conducting foundational science should not be disadvantaged because the societal effects of their work may not become visible for many years.
Measure value to the public, not simply visibility

Ultimately, public impact assessment should help NIH answer a straightforward but difficult question: What happened because the knowledge NIH supported entered the world?

Answering that question requires more than counting publications, citations, patents, news stories, or social-media mentions.

NIH should examine whether research was noticed, taken up, applied, and ultimately associated with meaningful health, economic, clinical, accessibility, or societal benefit, while recognizing that different forms of research travel through different pathways and on different timescales.

Public trust in NIH-funded science also depends upon NIH being able to demonstrate these contributions consistently, not only through a small number of highly visible success stories, but across the breadth of its research portfolio.

By connecting evidence from funding, scholarship, clinical practice, policy, innovation, public health, and public engagement, NIH can develop a more rigorous account of the value generated by federal research investment: one that distinguishes attention from impact, contribution from attribution, and anecdote from systematic evidence.

8. Other Areas: One overarching opportunity cuts across each of the areas addressed in this RFI: the need to shift from metric-based assessment toward evidence-based assessment.
For many decades, research evaluation has relied heavily on a relatively small number of publication- and citation-based indicators because these represented some of the most consistently available evidence of scientific activity. The research landscape has changed fundamentally.

Contemporary biomedical research now generates rich, interconnected evidence spanning grants, publications, datasets, software, protocols, methods, clinical trials, patents, policy documents, clinical guidelines, contributor roles, researchers, organizations, and collaborative relationships.

The opportunity for NIH is therefore not to identify a new metric to replace publication counts or citations. It is to integrate complementary evidence streams to develop a more complete understanding of how scientific knowledge is generated, validated, shared, reused, extended, translated, and ultimately contributes to public benefit.
Move from counting outputs to understanding pathways

Across the topics addressed in this RFI, a common principle emerges: individual outputs provide only a partial view of scientific contribution.

NIH increasingly has the opportunity to understand the pathways connecting its investments to subsequent scientific activity and value.
For example:
Rigor and reproducibility: research → transparency → reproducibility → independent verification → replication → more reliable knowledge
Data, software, and models: creation → sharing → stewardship → discovery → reuse → new research and scientific value
Training and mentorship: training investment → mentorship → researcher development → continued scientific contribution → development of subsequent researchers
Collaboration: investment → complementary expertise → scientific integration → shared outputs and capabilities → durable research networks → downstream value
Entrepreneurship and translation: fundamental discovery → validation → translational development → cross-sector engagement → application → adoption
Foundational scientific exploration: exploratory investment → new concept, method, resource, or capability → diffusion → new scientific directions → subsequent discovery

Public impact: research → attention → uptake → application → health, economic, clinical, or societal benefit

These pathways are neither universal nor linear, and they should not become rigid models against which every research project is evaluated. Their value is conceptual: they illustrate why assessment focused predominantly on individual outputs can miss much of the value created by federal research investment. NIH should increasingly ask not only "What was produced?" but "What did it enable?"

Assess contribution without overstating attribution

A second principle running across these topics is the importance of distinguishing contribution from attribution.

Biomedical progress is cumulative. Important discoveries frequently depend upon decades of foundational research, multiple investigators and institutions, shared datasets and infrastructure, methodological advances, validation studies, collaborations, funding programs, and subsequent translational activity.

Assessment systems should resist the temptation to attribute complex outcomes to individual publications, grants, or investigators simply because those relationships are easiest to measure.

Connected evidence instead provides an opportunity to identify credible contributions to longer scientific pathways.

This distinction is particularly important when assessing societal impact. NIH should be able to demonstrate that federally funded research contributed to advances in health, clinical practice, innovation, policy, and economic activity without implying causal attribution where the available evidence cannot support it.

Recognize the scientific contributions that conventional metrics miss

Scientific progress emerges through the interaction of people, infrastructure, methods, data, software, organizations, and institutions over extended periods.

Many activities that make this system function remain poorly represented in conventional assessment, including:

stewardship of shared datasets and research resources;
development and maintenance of scientific software;
methodological and statistical contributions;
creation and maintenance of research infrastructure;
mentorship and supervision;
peer review;
replication and validation;
development of standards and protocols;
technical and collaborative leadership; and
community building and scientific stewardship.

These activities should not automatically receive credit simply because they can be documented. Rather, NIH should develop mechanisms for recognizing when they make meaningful contributions to research

quality, capability, efficiency, reliability, or impact.

Persistent identifiers, structured contributor roles, interoperable metadata, and connected research information increasingly make these contributions observable.
The challenge is to convert that increased visibility into responsible evidence for assessment, rather than simply creating additional things to count.

Treat research assessment as a scientific endeavor
NIH should also treat research assessment itself as an object of scientific investigation.

Any indicator creates incentives. Poorly designed indicators can change researcher behavior in ways that undermine the objective the indicator was intended to support.
New assessment approaches should therefore be treated as hypotheses to be tested rather than metrics to be institutionalized immediately.

Candidate indicators should be evaluated for:
construct validity: whether the indicator actually measures the contribution it is intended to represent;
robustness across disciplines, research types, institutions, and career stages;
sensitivity to differences in scientific timescale;
completeness and quality of the underlying data;
susceptibility to gaming or strategic behavior;
potential inequities or structural biases;
administrative burden;
unintended behavioral consequences; and
incremental value beyond information already available to expert reviewers and program staff.

Where possible, NIH should test proposed indicators retrospectively before using them prospectively. For example, NIH could ask whether proposed indicators of emerging science would have identified historically important research directions before conventional metrics did; whether proposed translational indicators reliably distinguish meaningful progression from incidental activity; or whether proposed mentorship indicators accurately reflect successful researcher development across different career pathways.

This would allow NIH to validate the assessment system before attaching significant incentives to it.

Use indicators as evidence, not answers
No quantitative indicator should substitute for scientific judgment.

Connected evidence, natural language processing, network analysis, persistent identifiers, and other analytical approaches can make patterns visible that would be impossible for an individual reviewer or program officer to observe across an entire research system.

Their purpose should be to improve the evidence available to decision-makers, not to automate decisions that require scientific judgment.

This distinction is particularly important as the volume and sophistication of research information increase.
The appropriate model is: data → evidence → interpretation → expert judgment → decision
rather than: data → score → decision.

NIH should be cautious about composite scores that collapse fundamentally different dimensions of scientific contribution into a single ranking.
A researcher, project, program, or portfolio can be strong in one dimension and intentionally different in another. Foundational research should not be penalized because it lacks immediate translational impact. A high-value dataset should not be judged as though it were a publication. A collaborative specialist should not be assessed according to PI-oriented measures. A high-risk research portfolio should not be expected to produce the same success pattern as a program designed around predictable deliverables.

Evidence-based assessment should preserve those distinctions.

Evaluate portfolios as well as people
Many of the questions raised by this RFI cannot be answered adequately at the level of an individual investigator or award.
NIH should therefore complement researcher- and project-level assessment with portfolio-level assessment.

At portfolio scale, NIH can examine questions such as:
Are investments producing reusable research resources?
Are training programs generating durable scientific capacity?
Are collaborative programs producing genuinely integrated research networks?
Where are translational pathways progressing or stalling?
Is strong scientific evidence reaching clinical practice and public-health application?
Are emerging scientific areas receiving appropriate attention?
Is the balance between established and exploratory research changing over time?
Which NIH investments continue generating value long after the originating awards have ended?

This changes assessment from an exercise focused primarily on evaluating researchers into a capability NIH can also use to evaluate and improve its own funding strategy.
That distinction is important. Responsible research assessment should not only ask whether researchers are producing the right behaviors and outcomes. It should also allow the funding agency to ask whether its own policies, programs, incentives, and portfolio decisions are producing the research ecosystem it intends to create.

Build upon shared infrastructure and community standards
Continued investment in open scholarly infrastructure, persistent identifiers, trusted repositories, and community standards remains essential.
Digital Science has participated in this work directly, including through NIH's Generalist Repository Ecosystem Initiative (GREI). Initiatives of this kind demonstrate the value of interoperable infrastructure that enables research outputs to remain discoverable, accessible, connected, and reusable beyond the

systems or projects in which they originated.

Shared standards such as DOIs, ORCID iDs, ROR identifiers, RRIDs, grant identifiers, structured contributor roles, and interoperable metadata improve not only assessment but also reproducibility, transparency, collaboration, data stewardship, and research efficiency.

NIH should continue supporting this foundational infrastructure while taking advantage of the broader public, community, institutional, and commercial research-information ecosystem that has developed around it.

Where evidence can be obtained reliably from existing infrastructure, researchers should not be required to report it again.

The goal should be a richer evidence base with a lower reporting burden, not a larger assessment bureaucracy.

Recognize that research is an ecosystem
The questions in this RFI span researchers, funders, institutions, publishers, repositories, healthcare organizations, government, industry, and the public because modern biomedical research operates across all of them.

Assessment frameworks designed from the perspective of any one participant risk capturing only part of that system.

Digital Science's perspective on these questions has been shaped by more than two decades of working across this research ecosystem—including with researchers, universities and research institutions, publishers, research funders, governments, and commercial organizations, including pharmaceutical and biotechnology companies involved in translating scientific discovery toward application.

That cross-ecosystem perspective informs a central theme of this response: scientific value frequently becomes visible in the connections between systems, organizations, outputs, and stages of the research lifecycle, rather than within any single source of information.

Connected research infrastructure can increasingly make those relationships observable at scale. But infrastructure alone is not an assessment framework. The evidence it provides must be interpreted in the context of scientific purpose, disciplinary differences, timescale, uncertainty, and the intended mission of the research.
From metrics to evidence

The challenge facing research assessment is no longer simply one of measurement. It is increasingly one of interpretation.

As the evidence describing the research ecosystem becomes richer and more connected, NIH should resist replacing one dominant metric with another.

Instead, the opportunity is to develop an assessment framework that asks:
What contribution are we trying to understand?
What evidence would credibly demonstrate that contribution?
Over what timescale should it be observed?
What contextual information is required to interpret it?
What behaviors might measuring it incentivize?
What can quantitative evidence tell us, and what still requires expert scientific judgment?

Across the areas addressed in this RFI, this approach would allow NIH to move from counting isolated outputs toward understanding the pathways through which research creates scientific and public value.

The resulting framework would not eliminate metrics. It would put them in their proper role: as individual pieces of evidence within a broader, scientifically grounded assessment system.

That distinction is essential.

NIH does not need a new definition of scientific success expressed through a different score. It needs an evidence framework capable of recognizing the many ways rigorous, collaborative, exploratory, reusable, translational, and publicly valuable science contributes to its mission, and capable of evolving as science itself changes.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Digital-Science_NIH-Request-for-Information-on-Measuring-and-Rewarding-Scientific-Impact-1.pdf

Description: Digital Science response to the NIH Request for Information on Measuring and Rewarding Scientific Impact.

DIGITAL SCIENCE RESPONSE TO NIH REQUEST FOR INFORMATION ON MEASURING AND REWARDING SCIENTIFIC IMPACT

I am responding: On behalf of an organisation

Name: Digital Science

Name of Organisation: Digital Science & Research Solutions Inc.

Type of Organisation: Commercial Entity

Role: Organizational Official

• **Rigor and Reproducibility:** ***Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings.***

Digital Science Response:

Rigorous and reproducible science is most effectively assessed through observable and documented research practices rather than publication outcomes alone. This directly reflects the priorities set out in Executive Order 14303, Restoring Gold Standard Science, which identifies reproducibility, transparency, communication of uncertainty, and a willingness to subject scientific findings and assumptions to continued scrutiny as foundational tenets of trustworthy federal research.

Successful replication remains one of the strongest indicators of scientific robustness. However, replication is inherently retrospective, resource-intensive and feasible for only a fraction of published research. NIH should therefore complement direct replication with measurable indicators of the research practices and outputs that make findings independently assessable, reproducible, and ultimately replicable.

<u>Use connected research to identify where replication is needed most</u>

Persistent identifiers, interoperable metadata and connected research information have fundamentally changed what can be observed across the research lifecycle. Publications can increasingly be linked to grants, datasets, software, protocols, biological resources, organizations, researchers, and subsequent research outputs. This provides NIH with a substantially broader evidence base for assessing scientific rigor than publication counts and citation metrics alone. These evidence-based indicators reflect transparency, methodological rigor and the stewardship of research outputs.

This same connected evidence base can also help address a critical question: Where would additional replication effort have the greatest scientific value?

Rather than prioritizing replication primarily in response to highly visible failures or individual nominations, NIH could develop an evidence-based framework for identifying research areas where independent verification is most warranted. Such framework could combine multiple signals, including:

- availability and accessibility of quality underlying datasets and analytical code;
- availability of sufficiently detailed protocols and methods;
- use of trusted repositories and persistent identifiers;
- preregistration where scientifically appropriate;
- transparent reporting of methods, including negative and confirmatory findings;
- citation and demonstrated reuse of datasets, software and research resources; and
- independent validation or replication of published results.

Natural language processing and structured metadata applied to the full text of published research can now detect the presence or absence of specific transparency markers at scale, including disclosed funding, stated ethical approval, data and code availability, declared competing interests, validated across multiple fields to avoid disciplinary bias. This allows NIH to move from a largely reactive and subjective model of replication toward a more systematic, evidence-informed model: identifying consequential findings or research areas that have limited independent verification, weaker reproducibility signals, or substantial downstream scientific dependence. NIH could use such evidence to help prioritize replication investments, including expansion of existing efforts to independently replicate significant lines of NIH-supported research.

Reward the work required to make science reliable

Identifying where replication is needed does not resolve why so little of it happens. Academic incentive structures are built almost entirely around individual achievement, including first or last authorship, novel-discovery outputs, principal investigator status, publication volume, grant funding, and citation-based indicators. These measures can undervalue the collaborative and often less glamorous and less visible work required to verify, reproduce, challenge, or strengthen existing findings. Researchers undertaking rigorous replication largely absorb the professional cost of that work personally, while the resulting benefit of a more reliable scientific evidence base accrues broadly to the field as a whole.

NIH can help change this incentive structure by making contributions to reproducibility and replication visible, attributable, and creditable. When replication studies are linked to the original finding through persistent identifiers a contributor's specific role in it, replication design, execution, data verification, methodology, software, resource development, and experimental execution are captured through structured contributor taxonomies such as CRediT, those contributions can become part of an investigator’s

documented scientific record independent of authorship position high citation counts.

NIH could extend this further by formally recognizing verification and replication contributions in biosketches, progress reports, grant review criteria, and other assessment processes and by continuing to explore dedicated funding and recognition mechanisms for high-value replication research.

Measure practices and outcomes, not compliance alone

These indicators should prioritize evidence of accessibility, reuse, verification and scientific utility, rather than simple compliance with reporting requirements. A dataset that is discoverable, interoperable and reused by multiple research communities provides considerably stronger evidence of scientific value than one that is merely deposited.

Potential measurable indicators could include:

- proportion of applicable research outputs with accessible underlying datasets;
- proportion with accessible analytical code or computational workflows;
- availability and completeness of protocols and methodological documentation;
- use of trusted repositories and persistent identifiers;
- preregistration where scientifically appropriate;
- transparent reporting of methods, limitations, negative findings, and confirmatory findings;
- demonstrated reuse of datasets, software, protocols, and research resources;
- successful computational or methodological reproduction by independent researchers;
- documented independent validation or replication;
- links between original findings and subsequent replication or validation studies; and
- continued accessibility of research outputs over time.

Create infrastructure for negative and confirmatory findings

Transparent reporting of negative and null findings, named explicitly above as an indicator, and named directly among the Gold Standard Science tenets as a marker of trustworthy research culture, depends on more than researcher willingness. It requires infrastructure through which these outputs can be persistently identified, discovered, cited, connected to the original research, and recognized as legitimate scientific contributions regardless of journal publication.

Repository infrastructure capable of assigning persistent identifiers to diverse research outputs can provide a durable mechanism for making negative confirmatory findings part of the scientific record. Connecting these outputs to grants, investigators, datasets, protocols and origination studies would also allow NIH to evaluate their subsequent attention, reuse, and scientific influence.

This is important because an incentive to report negative findings will have limited effect if those findings remain effectively invisible within research assessment systems.

Minimize reporting burden through interoperability

Implementation should build upon existing community standards, including DOIs, ORCID iDs, ROR identifiers, RRIDs and grant identifiers, and structured contributor roles. Wherever possible, NIH should derive indicators from existing research infrastructure rather than requiring investigators to repeatedly report on information already available elsewhere.

The objective should be to create a connected evidence layer around NIH-supported research, not another administrative reporting layer.

Protect the independence of replication review

Replication also presents distinct review-integrity considerations. A proposal to replicate an influential finding may be reviewed by researchers who collaborated with, trained with, competed with, or have other significant scientific relationships to investigators associated with the original work.

For replication-focused funding, NIH could therefore consider conflict-of-interest screening that examines relevant relationships to both the replication applicant and the research being replicated. Connected publication, funding, institutional, and collaboration information could assist NIH in identifying relationships requiring additional review while preserving appropriate scientific expertise on review panels.

Distinguish transparency, reproducibility, and replication

Finally, NIH should avoid collapsing transparency, reproducibility, and replication into a single composite measure.

They represent related but distinct dimensions of scientific quality:

Transparency asks whether the evidence, methods, materials, assumptions, and limitations necessary to evaluate the research are available.

Reproducibility asks whether the reported result can be obtained using the original data, methods, or analytical procedures, as appropriate to the research domain.

Replication asks whether a finding can be independently tested and supported using new data, experiments, samples, or other appropriate independent approaches.

Each provides different information about the reliability of a scientific claim. NIH assessment frameworks should preserve these distinctions while recognizing their

complementary value.

Research assessment itself should also be treated as an evidence-based scientific endeavor. Any indicators NIH adopts should be evaluated for construct validity, disciplinary and methodological bias, susceptibility to gaming, administrative burden, and unintended behavioral consequences, and should evolve as evidence accumulates.

The opportunity for NIH is therefore broader than developing another set of research metrics. By connecting information already generated throughout the research lifecycle, NIH can build an evidence-informed system that helps determine which findings most warrant independent replication, which practices make research more reproducible, and which researchers and teams are contributing to the collective reliability of biomedical science. Such a framework would allow NIH to direct replication investment based on evidence rather than visibility while creating meaningful incentives for researchers to produce science that can be independently evaluated, reproduced, and trusted.

• **Data, Software, and Model Sharing:** ***Sharing scientific data and tools accelerates biomedical research discovery, in part, by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies. NIH seeks to incentivize and reward these activities to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.***

Digital Science Response:

Scientific publications are increasingly one component of a much broader ecosystem of research objects. The publication should be viewed as an entry point into the research record, rather than the research record itself. Contemporary biomedical discovery depends upon interconnected datasets, software, computational models, protocols, workflows and other digital research assets that enable verification, reuse and cumulative knowledge generation.

NIH assessment frameworks should recognize these outputs as first-class scientific contributions whose value extends well beyond their role in supporting individual publications.

The central objective should not simply be to incentivize sharing, but to incentivize useful and reusable sharing. A binary measure of whether a dataset, software package, or model was deposited provides limited information about whether the resource can be discovered, understood, accessed, reused, and built upon by other researchers.

The more meaningful question is:

Did the shared research asset continue to generate scientific value after the originating project ended?

This distinction is important for maximizing the return on NIH's research investment.

Move from measuring deposition to measuring reuse

Persistent identifiers, interoperable metadata, repositories, and connected research information increasingly make it possible to observe how research objects move through the research ecosystem, from creation and deposit to discovery, citation, integration, reuse, and subsequent scientific outcomes.

NIH therefore has an opportunity to evolve from measuring whether sharing occurred toward measuring whether shared assets are findable, accessible,understandable, reusable, and demonstrably used by others.

Potential measurable indicators could include:

- deposition in appropriate trusted repositories;
- assignment and continued resolution of persistent identifiers;
- completeness and quality of metadata;
- linkage to originating grants, publications, investigators, and institutions;
- adherence to relevant FAIR principles and community standards;
- availability of documentation sufficient to understand and reuse the resource;
- software versioning, documentation, dependency information, and maintenance;
- model documentation, including relevant assumptions, parameters, provenance, limitations, and versioning;
- use of appropriate licenses or terms that clearly communicate conditions of reuse;
- citation or formal acknowledgment of datasets, software, models, protocols, and other research assets;
- evidence of reuse in subsequent publications or research projects;
- reuse by investigators or institutions independent of the originating research team;
- integration into subsequent datasets, software, models, clinical research, public resources, or other research infrastructure;
- evidence of continued maintenance and accessibility over time; and
- breadth and diversity of downstream reuse across research communities.

These indicators should not necessarily be combined into a single score. Different research objects and disciplines have different reuse patterns, and assessment should account for those differences.

Reward stewardship, not compliance alone

NIH should distinguish deposit, stewardship, and reuse.

Deposit establishes that an output has been made available.

Stewardship establishes that the output remains accessible, documented, interpretable, interoperable, and usable.

Reuse demonstrates that the output has generated value beyond the originating research project.

These are related but distinct dimensions of responsible research sharing.

A dataset deposited to satisfy a policy requirement but lacking sufficient metadata or documentation to permit meaningful reuse should not be treated as equivalent to a well-curated dataset that is independently reused across multiple research projects. Similarly, software that is deposited but cannot be executed because dependencies, documentation, or version information are absent provides different scientific value from software that remains maintained, documented, and adopted by other research teams.

NIH incentives should recognize this continuum.

Give researchers credit for creating reusable research assets

Researchers continue to receive substantially more professional recognition for publications than for the creation and stewardship of reusable research resources. This creates a structural incentive problem: producing high-quality datasets, maintaining software, documenting models, curating metadata, and supporting reusable resources require significant scientific effort, yet those contributions may remain largely invisible in traditional research assessment.

NIH can help change this by making these contributions visible and creditable.

Dataset, software, model, and resource contributions could be incorporated more explicitly into NIH biosketches, progress reporting, grant review, and other assessment processes. Where contributor roles can be reliably attributed, NIH could recognize activities such as data curation, software development, validation, visualization, methodology development, and resource stewardship alongside more traditional publication-based evidence of contribution.

Importantly, credit should not depend solely upon citation counts. Dataset and software citation practices remain inconsistent across disciplines. NIH should therefore consider multiple forms of evidence, including formal citation, persistent-identifier linkages, documented reuse, integration into downstream research, and other appropriate indicators of adoption.

Treat software and computational models as scientific methodology

Software and computational models warrant particular attention because they are increasingly integral to biomedical methodology.

Where software materially contributes to scientific findings, its documentation, versioning, provenance, dependencies, accessibility, and maintenance directly affect the ability of other researchers to evaluate or reproduce the work. Similarly, computational and AI/ML models may require documentation not only of the model itself but also of relevant training or validation data, versioning, parameters, assumptions, limitations, and appropriate conditions of use.

Assessment frameworks should therefore recognize software and model stewardship as part of scientific rigor rather than treating these outputs as secondary technical products.

NIH should also recognize that maintenance itself is a scientific contribution. A widely used software package or research resource can continue enabling NIH-funded science for many years after its original grant has concluded. Traditional assessment approaches may substantially underestimate this continuing return on federal investment.

Use connected evidence to understand downstream value

A major opportunity is the ability to connect shared research assets to subsequent scientific activity.

Persistent identifiers and connected metadata can increasingly link a dataset or software resource to its originating grant and publication and then trace subsequent citations, publications, grants, investigators, institutions, clinical research, or other outputs that make use of it.

This creates the possibility of measuring research asset lineage: not simply whether an asset exists, but where it came from, who subsequently used it, and what additional scientific activity it helped enable.

For NIH, this could provide a new way to understand return on research investment. A dataset generated through one NIH award may subsequently support research conducted by dozens of investigators, across multiple institutions and funding programs. Those downstream uses represent scientific value created by the original investment but are largely invisible in publication-centered assessment.

NIH could therefore consider indicators such as:

Reuse rate: whether and how often a resource is independently reused.

Reuse breadth: the number and diversity of investigators, institutions, disciplines, or projects using it.

Reuse longevity: whether the resource continues to generate research activity over time.

Research lineage: the subsequent publications, grants, clinical studies, resources, software, or other outputs connected to the original asset.

Independent reuse: whether reuse extends beyond the originating investigators or institution.

Together, these measures would provide a substantially richer view of scientific impact than deposit counts alone.

Build upon NIH's work on measuring data sharing

NIH has already explored this challenge through efforts such as the Data Sharing Index (S-Index) Challenge, which examined approaches for moving beyond a binary determination of whether data were shared toward more scalable measures of data sharing and reuse.

Experience from that effort demonstrates both the importance and feasibility of connecting repository information, persistent identifiers, research metadata, and downstream scholarly activity to better understand the lifecycle of shared data.

Digital Science's participation in NIH's S-Index Challenge demonstrated the practical feasibility of this approach. Our submission combined repository infrastructure supporting persistent identification and FAIR-aligned metadata with connected research information capable of tracing relationships among datasets, publications, grants, researchers, institutions, and subsequent research activity. Importantly, these capabilities were based on established infrastructure already operating at scale rather than systems created solely for the Challenge. That experience suggests that many of the indicators described above can be derived from existing infrastructure and metadata, rather than requiring NIH to create new reporting systems or impose additional reporting requirements on researchers.

Existing research infrastructure already indexes datasets distributed across many repositories and connects research objects with publications, grants, researchers, organizations, and other outputs. This provides a practical foundation for evaluating not only whether data have been deposited, but whether those assets are subsequently discovered, cited, reused, and connected to additional scientific activity.

NIH can therefore build upon existing community and commercial infrastructure, standards, persistent identifiers, and metadata to develop and validate indicators before introducing additional reporting requirements.

Reduce burden through interoperability

Implementation should prioritize interoperability and automation.

DOIs, ORCID iDs, ROR identifiers, grant identifiers, software identifiers, repository metadata, and other established standards can provide much of the connective infrastructure required to understand research asset creation and reuse.

Where information can be derived reliably from existing systems, researchers should not be required to report it again.

This principle is particularly important if NIH expands assessment beyond publications. The goal should be to recognize a broader range of scientific contributions without creating a correspondingly broader administrative burden.

NIH should also recognize that no single indicator will adequately capture the value of all data, software, or models. Indicators should be validated for disciplinary differences, field maturity, resource type, susceptibility to gaming, and unintended

incentives. Measures of reuse should complement, rather than replace, qualitative scientific judgment.

Ultimately, NIH has an opportunity to shift the incentive structure from "Did you share it?" to "Did you make it reusable, and did it continue to create value?"

That shift would better recognize the researchers who create and steward high-value scientific resources, provide NIH with a more complete picture of the return generated by federal research investments, and encourage a research ecosystem in which data, software, and models remain active scientific assets long after the projects that created them have ended.

• **Training and Mentorship:** ***NIH recognizes that advancing its mission requires promoting the growth and stability of the biomedical research workforce. While NIH supports training and protects time through existing policies and programs, additional indicators are needed to assess and quantify the impact of these efforts.***

Digital Science Response:

The continued strength of the biomedical research ecosystem depends not only on generating new knowledge but on developing the people, skills and research practices, and cultures that make future discovery possible. Training and mentorship should therefore be assessed not simply as activities that occurred, but by the scientific capacity they create over time.

Traditional indicators such as the number of trainees supported, fellowships completed, or training activities delivered provide important administrative information but only a partial measure of impact. They establish that training occurred; they do not establish what that training produced.

NIH has an opportunity to shift from measuring training activity toward measuring researcher development, mentorship effectiveness, and sustained scientific capacity.

Follow outcomes beyond the training period

The impact of NIH-supported training frequently becomes apparent years after an award or fellowship ends. Assessment should therefore follow researcher trajectories over longer periods rather than relying predominantly on outcomes observable within a single grant cycle.

Potential longitudinal indicators could include:

- progression to subsequent career stages;
- time to research independence, where independence is an appropriate career objective;
- subsequent participation or leadership in funded research;
- movement into scientific leadership positions;
- continued participation in biomedical research across academia, healthcare, government, industry, nonprofit organizations, and other sectors;
- participation in interdisciplinary and multi-institutional research;
- development of sustained collaborative research networks;
- subsequent contributions to data, software, methods, protocols, and shared research resources;
- adoption and leadership of rigorous and reproducible research practices;
- subsequent mentoring and supervision of other researchers; and
- continued contribution to the biomedical research ecosystem over time.

Importantly, successful training should not be defined solely as progression toward a tenure-track faculty position or principal investigator status. NIH-supported researchers contribute to biomedical science through many pathways, including industry, clinical research, regulatory science, public health, government, research infrastructure, data science, software engineering, and other technical and collaborative roles.

An assessment framework that defines success too narrowly risks categorizing valuable scientific careers as attrition when they may instead represent substantial returns on NIH's training investment.

Measure mentorship by what it enables

Mentorship itself should be recognized as a scientific contribution.

Effective mentors do more than supervise research activity. They help researchers develop scientific judgment, methodological rigor, professional networks, collaborative skills, responsible research practices, and the ability to mentor others.

NIH could therefore consider indicators of mentorship effectiveness that examine the subsequent development of trainees over time. These might include trainee progression, continued participation in biomedical research, transition into appropriate leadership or independent roles, development of collaborative networks, contributions to rigorous and reproducible research, and evidence that former trainees subsequently become effective

mentors themselves.

This creates an important potential indicator of long-term impact: mentorship propagation.

Rather than viewing the completion of an NIH-supported fellowship or training award as the endpoint of measurement, NIH could examine the longer arc of the investment:

NIH-supported training ➜ mentorship ➜ researcher development ➜ continued contribution to the biomedical research ecosystem ➜ creation of new research, data, software, methods, or other scientific resources ➜ subsequent mentorship and development of additional researchers.

Viewed in this way, the impact of NIH-supported training is not limited to the career outcome of a single trainee. It can propagate through subsequent research contributions, collaborations, institutions, sectors, and generations of researchers. A training investment made in one individual may ultimately contribute to the development of an extended scientific network whose collective impact substantially exceeds what can be observed at the conclusion of the original award.

Persistent identifiers and connected research information increasingly make portions of this longitudinal pathway observable. By linking training awards and mentors to subsequent grants, research outputs, affiliations, collaborative networks, contribution roles, and later mentorship relationships, NIH could begin to understand not simply whether an individual completed training, but how that investment propagated through the biomedical research ecosystem over time.

Such evidence should not be reduced to a simplistic mentor ranking or score. Trainee outcomes are influenced by discipline, institution, career goals, economic conditions, personal choices, and many factors beyond an individual mentor's control. Longitudinal evidence should therefore inform qualitative assessment rather than substitute for it.

Use longitudinal research information to understand career trajectories

Advances in research information infrastructure make it increasingly possible to understand researcher development over longer periods.

Persistent researcher identifiers, grant information, publications, affiliations, contribution metadata, and collaborative relationships can help connect an individual's research activity across career stages. With appropriate safeguards and validation, these data can provide NIH with a portfolio-level view of what happens to researchers following NIH-supported training.

For example, connecting persistent researcher information across successive grants, research outputs, affiliations, and sectors can help NIH understand whether trainees subsequently participate in academic research, industry, government, healthcare, public health, regulatory science, or other parts of the biomedical ecosystem; whether they develop sustained research collaborations; and where appropriate, whether they progress

toward independent funding or scientific leadership.

This would allow NIH to ask questions that are difficult to answer from administrative completion data alone:

What kinds of careers do NIH-supported trainees build?

Which training environments consistently produce sustained participation in biomedical research?

Which programs produce researchers who subsequently develop other researchers?

Where does NIH-supported talent remain within the biomedical research ecosystem, and where is scientific capacity being lost?

Which training approaches are associated with stronger long-term outcomes across different career pathways?

The objective should not be to prescribe a preferred career destination, but to understand whether NIH training investments generate durable scientific capability.

Recognize the full range of contributions that sustain research careers

Many activities that underpin high-quality biomedical research, including mentorship, data stewardship, software development, research infrastructure, peer review, methodological support, and collaborative leadership remain less visible within publication and PI-centered assessment frameworks.

This can particularly disadvantage researchers whose scientific contribution occurs through technical, collaborative, or infrastructure leadership rather than through conventional principal investigator roles.

Structured contributor information and recognized contribution taxonomies can help make these roles more visible. Rather than inferring contribution predominantly from authorship position or PI status, NIH could recognize documented roles in areas such as data curation, software development, methodology, validation, project administration, supervision, and other contributions essential to modern team science.

This is relevant to workforce stability because NIH cannot effectively understand retention if its assessment systems recognize only a subset of the careers through which people contribute to biomedical research.

A researcher who moves from an NIH-supported training program into scientific software development, data stewardship, research infrastructure, regulatory science, or another research-enabling role should not automatically disappear from NIH's understanding of the workforce simply because that individual does not subsequently become a PI.

Recognize contributions to training culture

Mentorship also occurs beyond traditional one-to-one supervisory relationships.

Researchers contribute to workforce development by creating openly available training resources, developing interdisciplinary training programs, teaching responsible research practices, supporting reproducibility initiatives, participating in peer review, building research communities, and creating shared technical or methodological resources.

Potential indicators could therefore include:

- development and demonstrated use of openly available educational or training resources;
- contributions to interdisciplinary or cross-institutional training programs;
- leadership in research integrity, responsible conduct of research, and reproducibility initiatives;
- documented mentorship and supervision contributions;
- contributions to shared research methods, infrastructure, data, or software that enable other researchers;
- sustained collaborative relationships involving junior researchers;
- participation in peer review and other forms of scientific community stewardship; and
- evidence that trainees subsequently contribute to mentorship and workforce development themselves.

Where possible, assessment should emphasize evidence of **use and downstream impact,** rather than simply counting activities. For example, an openly available training resource that is adopted across multiple institutions provides stronger evidence of training impact than the existence of the resource alone.

Build on existing information rather than creating new reporting burdens

Much of the information needed to develop these indicators already exists across research systems.

Persistent identifiers and connected research information can increasingly link researchers with grants, publications, affiliations, research outputs, collaborative networks, and documented contribution roles over time. These data can complement NIH administrative information and targeted qualitative evidence to provide a richer understanding of researcher development.

The goal should be to derive longitudinal evidence from existing information wherever it can be done reliably, rather than requiring mentors and trainees to repeatedly report information NIH or the broader research ecosystem already holds.

NIH should also be cautious about the limitations of these indicators. Persistent identifier coverage varies; affiliation and career information can be incomplete; some sectors produce fewer visible scholarly outputs; career trajectories are not linear; and quantitative indicators may systematically underrepresent important forms of scientific contribution.

For these reasons, longitudinal data, structured contribution statements, trainee and mentor perspectives, and narrative evidence should complement one another.

Measure durable scientific capacity

Ultimately, NIH should assess training and mentorship according to the durable scientific capacity created by its investment.

A successful training investment is not simply a completed fellowship. It is a researcher who develops skills, contributes meaningfully to biomedical science, participates in productive scientific networks, and potentially develops the next generation of researchers.

A successful mentor is not simply someone who supervises a large number of trainees. It is someone whose mentorship contributes to the development of capable researchers and strengthens the culture and practice of science.

And a stable biomedical workforce should not be defined solely by the number of researchers who become principal investigators. It should encompass the diverse scientific, technical, clinical, public-sector, and collaborative careers required for modern biomedical research.

By following researcher trajectories over time, recognizing the full range of scientific contributions, and examining how mentorship propagates through subsequent careers and generations, NIH can move from counting training activities to understanding whether its investments are building and sustaining the biomedical research capacity required for its mission.

• **Collaboration:** ***As biomedical research becomes increasingly interdisciplinary and convergent, collaboration is growing in both scale and complexity. NIH seeks new approaches for assessing and recognizing contributions to team-based science.***

Digital Science Response:

Authorship remains one of the dominant mechanisms for recognizing scientific contribution, yet it captures only a fraction of what contemporary team-based science requires. As biomedical research becomes increasingly interdisciplinary, computational, collaborative, and cross-sector, NIH assessment frameworks should recognize collaboration as the integration of complementary expertise and contributions toward a scientific objective, not simply as co-authorship.

This distinction is increasingly important. Scientific teams now depend upon specialized contributions spanning experimental science, clinical expertise, statistics, data science, software development, methodology, research infrastructure, project leadership, community engagement, and other capabilities. Assessment systems that focus predominantly on first or last authorship risk overlooking many of the contributions upon which modern biomedical research depends.

NIH therefore has an opportunity to address two related but distinct questions:

What did each person contribute to the team?

and

Did the collaboration create scientific capability or value that the participants would have been unlikely to achieve independently?

Both should inform the assessment and recognition of team science.

Recognize contribution, not simply authorship position

Author order is an imperfect proxy for scientific contribution and becomes increasingly inadequate as teams grow larger and roles become more specialized.

NIH could place greater emphasis on structured contributor information, including recognized taxonomies such as CRediT, to make individual roles within collaborative research more visible. Contributions such as methodology development, software, data curation, validation, statistical analysis, visualization, supervision, project administration, resource development, and other activities should be recognizable as scientific contributions in their own right where appropriate.

This would be particularly valuable for researchers whose contributions are essential to multiple projects but who may rarely appear as first or last author or principal investigator.

The objective should not be to replace authorship, but to supplement it with evidence that better reflects who contributed what.

Over time, structured contribution information could enable NIH to recognize patterns of scientific contribution across multiple projects. A researcher who repeatedly provides essential methodological, computational, clinical, data, or infrastructure expertise across collaborative programs may be making an important contribution to NIH's mission even if traditional publication metrics systematically understate that role.

Distinguish collaboration from scientific integration

NIH should also distinguish between the size of a collaboration and the quality or depth of scientific integration.

More investigators, institutions, disciplines, or sectors do not necessarily indicate more effective collaboration. Indicators that simply reward larger teams could unintentionally encourage collaboration for its own sake.

More meaningful evidence would examine whether collaborators combine complementary expertise, share research resources, create sustained intellectual or technical relationships, and produce outputs or capabilities that cross traditional disciplinary or organizational boundaries.

Potential indicators could include:

- diversity and complementarity of documented contributor roles;
- interdisciplinary participation appropriate to the scientific question;
- sustained collaboration across projects or funding periods;
- cross-institutional or cross-sector research relationships;
- shared datasets, software, methods, protocols, infrastructure, or other scientific resources;
- joint development of community research resources;
- evidence of knowledge exchange across disciplines or sectors;
- subsequent reuse or adoption of collaborative outputs;
- development of new collaborations or research communities arising from the original team;
- translation of research across scientific, clinical, public-sector, or commercial environments where appropriate; and
- evidence that the collaboration creates capabilities or outputs beyond those attributable to a single participating group.

These indicators should complement expert assessment rather than be combined Mechanically into a collaboration score.

Understand collaboration as a network, not an event

Collaboration should also be assessed longitudinally.

A co-authored publication provides evidence that researchers worked together on the output. It provides considerably less information about whether the relationship developed into a sustained and productive scientific partnership.

Persistent identifiers and connected research information increasingly make it possible to examine collaborative relationships across grants, publications, datasets, software, clinical studies, organizations and other research activity over time.

This allows collaboration to be understood as a network that develops and evolves, rather than a characteristic of an individual publication.

For example, NIH could examine whether a funded interdisciplinary collaboration:

NIH investment ➔ brings together complementary expertise ➔ produces shared research and resources ➔ develops sustained collaborative relationships ➔ generates additional research or capabilities ➔ expands into new institutions, disciplines, or sectors ➔ produces downstream scientific or translational value.

This provides a richer conception of collaborative impact than counting co-authored publications or participating institutions.

It also allows NIH to distinguish between first-order and downstream effects of collaborative investments. The immediate output may be a publication, dataset, software resource, or clinical finding. The longer-term return may include a durable interdisciplinary network, shared infrastructure, subsequent jointly funded research, movement of methods between fields, or new collaborations that continue after the original award ends.

Evaluate whether funded collaborations become genuinely integrated

Connected research evidence could also help NIH evaluate collaborative programs, centers, and consortia at the portfolio level.

Relationships among funding, researchers, organizations, publications, datasets, software, research resources, and documented contributor roles can provide evidence about whether collaborations develop scientific connections beyond the administrative structure of the original award.

For example, NIH could examine whether participating groups subsequently produce joint outputs, share or reuse research resources, establish new cross-institutional relationships, develop interconnected research programs, or continue collaborating after the initial funding period.

This would not by itself determine whether a collaboration is scientifically successful; expert judgment remains essential. It could, however, help NIH distinguish between collaborations showing evidence of deep and sustained scientific integration and those showing comparatively limited evidence of interaction.

For large collaborative investments, this could provide valuable evidence for program evaluation and future funding decisions.

Recognize cross-sector collaboration

Team science increasingly extends beyond academic institutions.

Biomedical research may bring together academic researchers, healthcare organizations, government laboratories, industry, nonprofit organizations, patient or community groups, data and technology organizations, and other participants.

Assessment frameworks should recognize the value of these collaborations without assuming that all sectors produce the same kinds of outputs.

A successful academic-industry collaboration, for example, may generate software, data resources, methods, technologies, patents, clinical applications, or other outputs that are not adequately represented by publication counts. Similarly, collaborations involving healthcare or public-sector organizations may produce implementation, policy, clinical, or population-health effects that require different forms of evidence.

NIH should therefore assess collaboration according to the scientific objective and intended contribution of the participants rather than imposing a single publication-centered definition of collaborative success.

Use connected information to help form teams as well as assess them

The same research information that supports assessment can also help NIH facilitate future collaboration.

Disambiguated researcher profiles connected to publications, grants, research topics, methods, datasets, affiliations, and previous collaborations can help identify relevant expertise beyond established professional networks.

This creates an opportunity for NIH not simply to measure collaboration after it occurs, but to help identify complementary expertise before teams are formed.

At a portfolio level, such approaches could help identify researchers or institutions working on related problems from different disciplinary perspectives, reveal capabilities that are missing from existing research networks, and identify opportunities to connect investigators who might not otherwise encounter one another.

Similar information could assist with identifying potential reviewers with appropriate expertise while supporting conflict-of-interest screening, although such systems should augment rather than replace established NIH review procedures.

Minimize burden and avoid unintended incentives

Implementation should build upon persistent identifiers, structured contributor roles, existing grant and publication information, and interoperable research infrastructure

wherever possible.

Researchers should not be required to produce extensive new collaboration reporting when relevant evidence can be derived reliably from information already generated through the research process.

NIH should also be cautious about creating incentives that reward collaboration numerically. Number of collaborators, institutions, disciplines, or countries may describe a research team but should not automatically be interpreted as evidence of quality.

Network-based measures should likewise be used carefully. Highly connected researchers are not necessarily the most valuable collaborators, and important specialists may contribute deeply to relatively small numbers of projects.

Quantitative evidence should therefore support expert assessment of contribution, complementarity, integration, durability, and scientific value, rather than become an end in itself.

Recognize collaboration as a scientific capability

Ultimately, NIH should move from treating collaboration primarily as an authorship pattern toward recognizing it as a scientific capability.

The relevant questions are not simply how many researchers participated or how many institutions appeared on an award. They are:

Were complementary capabilities brought together?

Can the contributions of individual team members be recognized?

Did participants genuinely integrate their expertise and resources?

Did the collaboration create something that would have been difficult to achieve independently?

Did the relationships, resources, or capabilities created through the collaboration continue to generate value?

A funding agency that cannot distinguish integrated collaboration from a shared cover page cannot reliably reward the former or understand the return generated by its collaborative investments.

By combining structured contribution information, persistent identifiers, longitudinal research information, network evidence, and expert scientific assessment, NIH can develop a more defensible framework for recognizing the individuals who make team science possible while also identifying the collaborative structures that produce genuine

scientific integration and sustained biomedical value.

• **Entrepreneurship and Translation:** ***To advance the translation of discoveries into products for people, NIH fosters the maturation of discoveries from bench to bedside and promotes innovation across sectors. More can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission.***

Digital Science Response:

Patents, licenses, company formation, and commercialization are among the most visible and readily measurable outcomes of biomedical innovation. They are important indicators, but they capture only a portion of the process through which NIH-supported discoveries ultimately improve human health.

NIH should therefore assess not only translational outcomes, but also contributions to translational pathways: the cumulative, often non-linear process through which knowledge moves from fundamental discovery toward clinical, commercial, public-health, policy, or other forms of application.

This distinction matters because biomedical translation rarely results from a single publication, investigator, institution, or funding award. It frequently develops through successive contributions from researchers, organizations, funding programs, healthcare systems, government, industry, and other participants over extended periods.

A framework that recognizes those contributions would better reflect how biomedical innovation actually occurs and allow NIH to understand more fully how its investments contribute to downstream health, economic, and societal value.

Measure pathways, not simply endpoints

Traditional translational indicators tend to emphasize observable endpoints: patents, licenses, spinouts, products, or clinical applications.

NIH should retain these measures while also examining the intermediate steps through which research progresses toward application.

Potential indicators could include:

- progression of NIH-supported research into subsequent translational or development funding;
- patent applications and granted patents linked to NIH-supported research;
- licensing activity and technology transfer, where appropriate;
- formation of companies or other commercialization vehicles;
- industry partnerships and cross-sector research relationships;
- progression into clinical studies or clinical trials;
- incorporation into clinical guidelines or standards of care;

- citations or incorporation into regulatory documents;
- incorporation into public policy or public-health guidance;
- development and adoption of diagnostic tools;
- development and adoption of research or clinical software;
- development of standards, methods, protocols, or shared research resources;
- use within healthcare systems or clinical practice;
- development of public-health interventions;
- subsequent investment or funding that builds upon the original research; and
- demonstrated adoption or use by organizations beyond the originating research team.

These indicators should be interpreted according to the scientific field and intended pathway to impact. Not all valuable NIH-funded research should result in a patent, company, clinical trial, or commercial product.

Understand translation as a connected pathway

Advances in connected research information increasingly make it possible to observe translational pathways empirically.

Persistent identifiers and linked research information can connect publications with grants, patents, clinical trials, datasets, software, researchers, organizations, policy documents, clinical guidelines, and other downstream activity.

Rather than viewing these outputs independently, NIH could increasingly examine the pathways connecting them:

NIH investment → fundamental discovery → subsequent research and validation → translational development → cross-sector engagement → clinical testing or other application → adoption → health, economic, or societal value.

Importantly, these pathways will not always be linear.

A fundamental discovery may contribute to multiple subsequent research programs. A dataset may enable a diagnostic technology. Software may move directly into clinical research without generating a patent. A methodological advance may be incorporated into clinical guidelines. An NIH-supported discovery may contribute to a commercial product years later through research conducted by investigators and organizations that were not part of the original award.

Understanding these connections provides a more realistic model of translation than attempting to attribute an eventual societal outcome to a single publication or investigator.

Distinguish contribution from attribution

NIH should be cautious about attempting to assign exclusive credit for downstream impact.

Translation is cumulative. A successful therapeutic, diagnostic, clinical practice, or public-health intervention may depend upon decades of foundational research funded through multiple programs and performed across numerous institutions.

The appropriate question is often not: **“**Which investigator caused this outcome?” but rather: “What NIH-supported research contributed to the pathway that made this outcome possible?”

Connected evidence can help identify these contributions without implying simplistic causality.

This approach would also allow NIH to recognize foundational research that creates important translational value even when the originating investigators never participate directly in commercialization.

Identify translational inflection points earlier

The same evidence used to evaluate translation retrospectively could potentially help NIH identify emerging translational opportunities prospectively.

Today, important translational signals may appear across disconnected systems. A body of NIH-supported research may begin generating patent activity, follow-on funding, industry collaboration, clinical investigation, new software or diagnostic applications, or increasing attention from organizations outside the originating research community.

Individually, these signals provide incomplete information. Considered together, they may indicate that a line of research is approaching a translational inflection point.

NIH could explore whether combinations of signals such as: “research maturation + independent validation + follow-on funding + patent activity + cross-sector engagement + clinical activity” can help identify research areas where targeted translational support, expertise, partnership, or additional funding might accelerate progress.

This would move the assessment beyond asking “What translated?” toward also asking “What may be ready to translate next?”

Such approaches should be validated carefully and used to support, rather than automate, scientific and programmatic judgment. Translational potential is inherently uncertain, and

indicators should not create pressure to commercialize research prematurely.

Recognize the people who enable translation

Entrepreneurial and translational activity should also be recognized as a legitimate scientific contribution.

Researchers may contribute to translation through invention, company formation, technology development, industry collaboration, clinical development, standards development, software creation, regulatory science, implementation, or other activities that move knowledge toward use.

These contributions can be difficult to see in publication-centered assessment.

NIH could recognize documented translational contributions within biosketches, progress reports, portfolio assessment, and other appropriate review processes. Structured contributor information and connected evidence could help make these activities more visible.

Prior entrepreneurial and translational experience may also provide useful contextual evidence, for example, previous involvement in technology transfer, industry collaboration, clinical development, licensing, or company formation. However, such experience should not become a prerequisite or simplistic score for evaluating future potential.

Researchers at institutions with mature technology-transfer ecosystems have substantially different opportunities from researchers at smaller or less-resourced institutions. Similarly, some biomedical fields offer clearer commercial pathways than others. Assessment should recognize entrepreneurial contribution without creating a structural advantage for investigators who already have access to established commercialization networks.

Recognize multiple routes to impact

Commercialization is one route through which NIH-supported research creates value, but it is not the only one.

Translation may occur through clinical practice, healthcare delivery, public-health interventions, regulatory science, government policy, clinical guidelines, standards, open-source software, shared research infrastructure, diagnostic methods, or widespread adoption of a new research technique.

NIH should therefore avoid equating entrepreneurship exclusively with company formation or translation exclusively with commercialization.

Different types of research should be assessed against pathways appropriate to their intended contribution.

This is particularly important for foundational research. Basic science may have enormous long-term translational significance despite having little immediate commercial potential. An assessment framework overly weighted toward near-term translation could inadvertently discourage the fundamental discovery upon which future biomedical innovation depends.

Measure time to translation carefully

Translation also operates across very different timescales.

Some computational tools or methods can move rapidly into use. Therapeutic development may take many years. Foundational discoveries may generate little visible translational activity for decades before becoming critical to a new technology or treatment.

NIH should therefore consider time-to-translation and translational persistence as contextual rather than purely performance indicators.

Longitudinal analysis can help NIH understand how different types of research move toward application, where pathways commonly stall, and which interventions are associated with successful progression.

This could help answer important portfolio-level questions:

- Where do promising NIH-supported discoveries most commonly stall?
- Which types of support are associated with successful movement between translational stages?
- Where are collaborations with industry, healthcare, or other sectors accelerating translation?
- Which foundational investments ultimately contribute to multiple downstream innovations?
- Which NIH programs are generating translational value that conventional publication metrics fail to capture?

Use connected evidence for portfolio intelligence

At sufficient scale, these indicators could provide NIH with more than an assessment framework. They could provide portfolio intelligence.

Connected evidence across grants, publications, patents, clinical trials, organizations, researchers, and downstream outputs could help program officers and institute leadership identify emerging areas of translational activity across an entire portfolio rather than relying exclusively on individual investigators or technology-transfer offices to surface opportunities.

This could help NIH identify clusters of related discoveries, emerging cross-sector engagement, technologies progressing toward clinical investigation, and areas where additional translational support may be warranted.

It could also help NIH understand the comparative pathways through which different institutes, programs, fields, or funding mechanisms generate downstream value.

The objective should not be to predict commercial success. It should be to provide decision-makers with additional evidence about where translation appears to be occurring, where it may be emerging, and where potentially valuable research is encountering barriers.

Build upon existing research information infrastructure

Many of the relationships required for this type of assessment can already be observed through existing research information infrastructure.

Large-scale connected research databases link publications with grants, patents, clinical trials, researchers, institutions, and other research outputs, making it increasingly possible to examine translational relationships across entire research portfolios rather than reconstructing them manually one project at a time.

Digital Science's experience connecting grants, publications, patents, clinical trials, researchers, and organizations demonstrates the practical feasibility of examining these pathways at research-system scale. Dimensions, for example, links research activity across these different stages, allowing relationships between foundational funding, subsequent research, patent activity, and clinical investigation to be examined within a common evidence environment. This type of existing commercial infrastructure provides NIH with an opportunity to test and validate translational indicators using evidence already available today, rather than first constructing an entirely new reporting or tracking system.

This infrastructure provides a practical starting point for NIH to test translational indicators, examine research pathways retrospectively, and evaluate whether combinations of signals can reliably inform prospective portfolio analysis.

Implementation should build upon existing identifiers, standards, and data sources wherever possible, supplementing these with NIH administrative information and expert assessment rather than creating additional investigator reporting requirements.

Assess contribution while preserving discovery

Ultimately, NIH should expand its assessment of entrepreneurship and translation from counting visible endpoints toward understanding the pathways through which federally

funded knowledge becomes useful.

The relevant questions are not simply:

- Was there a patent?
- Was a company formed?
- Was a product commercialized?

They also include:

- What research made the eventual application possible?
- Who contributed to moving that research toward use?
- What intermediate steps indicate meaningful translational progress?
- Where are promising discoveries encountering barriers?
- Can NIH recognize emerging translational opportunities early enough to help accelerate them?

A discovery that becomes visible as translational only when it reaches a technology-transfer office may already have traveled a substantial portion of its translational pathway.

By connecting evidence across the research lifecycle, NIH can develop a more complete understanding of how its investments contribute to innovation, identify translational momentum earlier, recognize the researchers and teams who enable it, and support pathways that deliver sustained health, economic, and societal value, while continuing to protect the foundational science upon which future translation depends.

• **Foundational Scientific Exploration:** ***Basic research and high-risk, high-reward initiatives represent foundational scientific exploration with an inherent risk of failure but the potential for paradigm-shifting breakthroughs. Further embedding these approaches within the academic enterprise is essential to ensuring continued support for bold and innovative research directions.***

Digital Science Response:

Any indicator built around short-term productivity will systematically penalize the research most likely to produce paradigm-shifting results. NIH's challenge is therefore not to predict which individual foundational or high-risk research project will ultimately transform a field, but to protect sufficient space for exploration within a research portfolio that also rewards more predictable scientific progress.

Transformative advances frequently arise from fundamental research, enabling technologies, methodological innovation, exploratory investigation, unexpected findings, and new combinations of ideas that may initially appear incremental, uncertain, or disconnected from immediate application.

These contributions expand the scientific landscape by creating new questions, methods, resources, and capabilities rather than necessarily producing immediate translational outcomes.

Assessment of foundational scientific exploration should therefore recognize both the contributions created by exploratory research and NIH's broader success in maintaining a portfolio in which scientific risk and uncertainty can be sustained.

Do not require exploratory research to predict its own impact

One of the fundamental difficulties in assessing foundational research is that its eventual importance may be unknowable when the research is proposed or even when the original award concludes.

A discovery, method, dataset, biological resource, or enabling technology may acquire significance only years later, when another scientific development makes it newly useful. Research initially regarded as highly specialized may become foundational to an emerging field. A negative result may eliminate an unproductive direction and redirect subsequent research. A tool created for one purpose may eventually enable discoveries its creators never anticipated.

NIH should therefore avoid assessment frameworks that require foundational research to demonstrate short-term downstream impact as evidence of success.

Instead, potential indicators could include:

- development of novel methodologies, computational approaches, or instrumentation;
- creation of foundational datasets, biological resources, software, or other enabling infrastructure;
- introduction of new concepts, hypotheses, or research questions;
- development of novel combinations of methods or disciplines;
- establishment of new interdisciplinary research directions;
- production and dissemination of negative or null findings that meaningfully refine scientific understanding;
- subsequent adoption of methods, resources, or concepts by researchers outside the originating team;
- diffusion of methods or resources across disciplines;
- emergence of new research communities building upon foundational work; and
- evidence, over longer timescales, that exploratory work enabled subsequent scientific advances.

These indicators should describe contribution and influence without implying that every exploratory project must ultimately generate a transformative outcome.

Evaluate foundational contribution over longer timescales

Foundational research frequently operates on timescales poorly aligned with conventional research assessment.

By the time the importance of a foundational contribution is apparent through citation counts, clinical applications, patents, or other downstream indicators, the funding decisions that enabled it may have occurred many years earlier.

Longitudinal research information provides an opportunity to examine these contributions differently.

Persistent identifiers and connected evidence can increasingly link grants, publications, methods, datasets, software, patents, clinical studies, researchers, institutions, and other outputs over extended periods. This can help reveal how an initially exploratory investment subsequently contributes to scientific capability.

For example:

NIH exploratory investment ➜ new concept, method, resource, or capability ➜ adoption by other researchers ➜ diffusion across research communities ➜ new questions and combinations ➜ subsequent discoveries ➜ potentially transformative scientific or translational impact.

The important point is that the original project's value should not be judged solely by whether it reaches the end of this pathway.

Creating a method or resource that opens new scientific possibilities may itself represent a valuable outcome of foundational research.

Identify emerging science before conventional metrics can see it

Connected research evidence also creates an opportunity to move beyond retrospective assessment.

Citation-based measures are inherently lagging indicators. By the time an emerging research area accumulates substantial citation activity, its importance may already be widely recognized.

NIH could complement conventional assessment with horizon-scanning approaches designed to identify nascent scientific directions while they are still forming.

Potential signals could include:

- emergence of new concepts or terminology;
- rapid development or adoption of new methods;
- unusual combinations of previously separate research areas;
- growth of new interdisciplinary research networks;

- migration of methods or technologies between fields;
- increasing creation or reuse of specialized datasets, software, or research resources;
- formation of new clusters of investigators around emerging questions; and
- changes in funding, publication, or collaboration patterns that suggest the development of a new research direction.

Natural language processing, topic analysis, network analysis, and connected research information can help identify such patterns across large bodies of scientific activity.

These signals should not be interpreted as automatic indicators of scientific quality or future breakthrough potential. Emergence is not synonymous with importance, and novelty is not synonymous with scientific merit. Expert scientific judgment remains essential.

Their value is instead in helping NIH see where the scientific landscape may be changing early enough to investigate further.

Use horizon scanning to inform funding strategy

Applied prospectively, these approaches could provide NIH with a form of scientific horizon scanning.

Rather than waiting until an emerging field has accumulated years of conventional evidence, program officers and institute leadership could receive evidence about where new concepts, methods, technologies, interdisciplinary combinations, and research communities appear to be forming.

This could help NIH ask:

- Which potentially important scientific directions are emerging?
- Which established fields are being disrupted by new methods or technologies?
- Where are previously separate disciplines beginning to converge?
- Which enabling technologies are beginning to diffuse across biomedical research?
- Where is scientific activity emerging faster than NIH investment?
- Are there potentially important areas of exploration that fall between existing programmatic structures?

The objective should not be to automate funding decisions or predict the next breakthrough. It should be to expand the evidence available to scientific leaders deciding where additional exploration may be warranted.

Assess NIH's portfolio risk, not just investigator risk

Perhaps the more important opportunity is to apply these approaches to NIH's own funding portfolio.

High-risk, high-reward research is inherently uncertain at the individual-project level. Some projects will fail to produce their intended results. That is a characteristic of genuine scientific risk, not necessarily evidence of poor funding decisions.

Assessment is therefore more meaningful at the portfolio level.

NIH could examine the distribution of investment across established, developing, and emerging research directions and how that distribution changes over time.

Potential portfolio-level indicators could include:

- concentration of funding within established research areas;
- proportion of funding supporting emerging or exploratory areas;
- entry of new investigators or disciplines into established research fields;
- funding for novel combinations of methods or disciplines;
- diversity of scientific approaches supported within a portfolio;
- changes in the distribution of funding as scientific fields mature;
- persistence of support for exploratory areas during periods of uncertainty;
- subsequent development of research communities around previously exploratory investments; and
- evidence that earlier exploratory investments created methods, resources, or capabilities subsequently used across NIH-supported science.

These measures could help NIH evaluate an important question that cannot be answered by examining individual grant productivity:

Is NIH maintaining a research portfolio with sufficient capacity for scientific exploration and responsible risk-taking?

Detect portfolio drift toward predictability

Portfolio analysis could also help identify risk drift.

Funding systems can gradually become more concentrated around established investigators, methods, topics, or scientific approaches even without an explicit decision to become more conservative.

Each individual funding decision may be scientifically defensible while the cumulative portfolio nevertheless becomes increasingly concentrated around areas with established evidence, established communities, and predictable outputs.

NIH could therefore examine changes in portfolio concentration and scientific novelty longitudinally.

For example: Portfolio at time 1 ➔ distribution across established and emerging areas ➔ subsequent funding decisions ➔ changing concentration of investment ➔ portfolio at time 2.

This would allow NIH to examine whether the agency, an institute, or a program is maintaining its intended balance between exploitation of established scientific opportunity and exploration of less certain directions.

Importantly, such analysis should not assume that a particular percentage of funding to emerging areas is inherently optimal. Appropriate risk tolerance will vary across institutes, programs, scientific opportunities, and public-health needs.

The objective is to make the portfolio's risk characteristics visible enough to be managed intentionally.

Learn from unsuccessful high-risk research

NIH should also distinguish between scientific failure and unproductive investment.

A high-risk project that does not confirm its original hypothesis may nevertheless produce important negative findings, eliminate an unproductive scientific direction, generate a useful method or dataset, reveal methodological limitations, or create knowledge that informs subsequent research.

If high-risk programs are assessed primarily according to whether projects achieve their original anticipated outcomes, researchers will have a rational incentive to propose projects whose outcomes are more predictable.

NIH should therefore recognize informative failure as a legitimate potential outcome of exploratory research, provided the work was rigorous and the resulting knowledge is made available to the scientific community.

This is also why transparent reporting of negative and null results, addressed elsewhere in this RFI, is particularly important for high-risk research.

Protect unconventional combinations of expertise

Transformative research may also arise at the boundaries between established disciplines.

Connected research information can help identify unusual combinations of expertise, methods, concepts, or research communities that differ from established collaboration patterns.

NIH could examine whether funding mechanisms intended to support innovation are actually enabling these unconventional combinations or predominantly supporting teams whose structures resemble already successful research networks.

Again, unusual combinations should not automatically be rewarded. Novelty without scientific rationale has little value.

But the ability to observe disciplinary and methodological combinations provides NIH with another way to assess whether programs intended to encourage convergence and exploration are actually broadening the scientific landscape.

Build on existing evidence infrastructure

Many of the signals required for this type of analysis already exist within research information systems.

Large-scale connected evidence across grants, publications, researchers, organizations, datasets, software, methods, patents, clinical research, and other outputs can support longitudinal analysis of emerging fields, knowledge diffusion, interdisciplinary convergence, and funding concentration.

Natural language processing and network analysis can further help identify changes in concepts and relationships that are difficult to detect through traditional publication and citation measures alone.

NIH could use this existing evidence base to test and validate horizon-scanning and portfolio-risk indicators retrospectively before applying them prospectively.

For example, NIH could examine historical research areas now recognized as transformative and ask: **Would the proposed indicators have detected their emergence before conventional metrics did?**

Similarly, NIH could examine previous high-risk funding portfolios to determine whether proposed measures meaningfully distinguish exploratory investment from established research.

This type of retrospective validation would be important before incorporating new indicators into funding strategy.

Protect the conditions for discovery

Ultimately, foundational scientific exploration presents a different assessment problem from research intended to produce predictable near-term outcomes.

The objective should not be to develop a metric that predicts which high-risk project will succeed. Doing so would risk eliminating the uncertainty that makes the research high-risk in the first place.

Instead, NIH should ask:

- Is exploratory research creating new scientific possibilities?
- Can NIH identify emerging directions early enough to investigate them?

- Is the agency maintaining sufficient diversity of scientific approaches?
- Are programs intended to support high-risk research actually funding research that differs meaningfully from established patterns?
- Is NIH's portfolio gradually becoming more conservative without intending to do so?
- Does the system recognize rigorous, informative failure as part of scientific exploration?

A portfolio that cannot see its own risk profile cannot knowingly protect it.

By combining longitudinal research information, horizon scanning, portfolio analysis, and expert scientific judgment, NIH can make the conditions for exploration more visible without attempting to make uncertain science artificially predictable. This would help preserve the scientific risk, curiosity, and methodological innovation from which future paradigm-shifting discoveries are most likely to emerge.

• **Public Impact:** ***NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.***

Digital Science Response:

Public impact is too often measured by visibility rather than uptake. Media attention is not the same as influence on clinical practice or policy. NIH's assessment framework should draw that distinction explicitly.

Scientific knowledge creates public value through diverse and interconnected pathways, including clinical practice, public health, regulatory science, public policy, education, industrial innovation, healthcare delivery, accessibility of knowledge, and public engagement. These pathways often interact, reinforce one another, and develop over extended periods.

NIH should therefore assess public impact as a pathway from attention to uptake, application, and ultimately societal benefit, rather than attempting to represent it through a single impact metric.

Distinguish attention, uptake, application, and outcome

Different indicators provide evidence of fundamentally different things.

Attention indicates that research has been noticed or discussed.

Uptake indicates that research has been incorporated into a professional, scientific, policy, clinical, educational, or other decision-making context.

Application indicates that knowledge has influenced a practice, intervention, technology, policy, guideline, or other activity.

Outcome indicates evidence of resulting health, economic, accessibility, or broader societal benefit.

These should not be treated as equivalent.

For example: NIH-supported research → publication → clinical guideline citation → adoption in clinical practice → change in patient care → improved health outcome.

Similarly: NIH-supported research → evidence base → public-health guidance → implementation → population-level intervention → measurable health benefit.

Or: NIH-supported research → enabling discovery → technology development → commercial adoption → economic activity → broader health or societal benefit.

A media mention may provide evidence of public attention. A citation in a clinical guideline provides stronger evidence that research has entered professional decision-making. Demonstrated implementation of that guideline represents a further stage of impact. Evidence that implementation contributed to improved outcomes represents something different again.

NIH should preserve these distinctions rather than collapsing them into a single measure of "impact."

Measure multiple pathways to public value

Because NIH supports an extraordinarily diverse biomedical research portfolio, no single pathway will adequately describe public impact.

Potential measurable indicators could include:

Clinical impact

- incorporation of NIH-supported research into clinical guidelines or standards of care;
- citations in clinical decision-support resources;
- progression of research into clinical studies or trials;
- adoption of evidence-based interventions within healthcare systems;
- development or adoption of diagnostics, therapeutics, devices, software, or other clinical tools;
- changes in clinical practice associated with research findings; and
- where evidence permits, contribution to improved patient outcomes.

Public-health and policy impact

- incorporation into public-health guidance;
- citation or use in federal, state, local, or international policy documents;
- incorporation into regulatory science or regulatory decision-making;
- contribution to standards or technical guidance;

- adoption within public-health programs or interventions; and
- evidence of changes in population-health practice associated with the research.

Economic and innovation impact

- patents and licensing associated with NIH-supported research;
- company formation and entrepreneurial activity;
- industrial research partnerships;
- subsequent private-sector investment;
- development of products, technologies, services, or research tools;
- employment or economic activity associated with NIH-supported innovation; and
- downstream commercial use of NIH-supported knowledge or resources.

Accessibility and knowledge diffusion

- public accessibility of NIH-supported research outputs;
- availability and reuse of datasets, software, protocols, models, and other research resources;
- use of research in educational or professional training materials;
- adoption of research resources by communities beyond the originating investigators;
- accessibility of findings to patients, practitioners, policymakers, and other intended audiences; and
- sustained availability of publicly funded research resources.

Public engagement and trust

- evidence of responsible communication of scientific findings, including uncertainty and limitations;
- engagement with patients, communities, and relevant public stakeholders;
- use of NIH-supported evidence in trusted public-information resources;
- transparent correction or updating of scientific findings as evidence changes; and
- where methodologically defensible, changes in understanding or confidence associated with sustained engagement.

These categories should remain distinct. A research project should not be expected to demonstrate impact across all of them.

Measure contribution, not simplistic attribution

NIH should distinguish between attribution and contribution when assessing public impact.

Most improvements in health and society cannot credibly be attributed to a single publication, investigator, grant, or institution. A change in clinical practice may depend upon decades of fundamental research, validation studies, clinical trials, systematic reviews, professional guidelines, healthcare implementation, and public-health activity.

The appropriate question is therefore often not: "Did this NIH award cause this societal outcome?" but: "What role did NIH-supported research play in the pathway that contributed to this outcome?"

Connected research evidence can help reconstruct these pathways without implying causality that the evidence cannot support.

For example: multiple NIH investments ➜ body of evidence ➜ clinical synthesis ➜ professional guideline ➜ healthcare implementation ➜ patient benefit.

This provides a more scientifically defensible account of public impact while recognizing the cumulative nature of biomedical progress.

Trace public impact longitudinally

Public impact frequently occurs years or decades after the original research is funded.

Assessment should therefore connect evidence across longer timescales rather than requiring investigators to demonstrate societal impact within the period of an individual award.

Persistent identifiers and connected research information increasingly make it possible to link grants, publications, datasets, software, patents, clinical trials, policy documents, clinical guidelines, researchers, and organizations over time.

This creates an opportunity for NIH to examine impact pathways across its portfolio.

Rather than relying predominantly on individual success stories, NIH could ask:

- Which NIH-supported research contributes to clinical guidelines?
- Which research areas are subsequently reflected in public-health or regulatory decision-making?
- Which NIH investments contribute to technologies or industries?
- Which datasets, software, and research resources continue to create public value after the originating grant concludes?
- How long does it typically take different forms of NIH-supported research to move from discovery to uptake and application?
- Where does research appear to reach an uptake stage but fail to progress toward implementation?

This would provide a more systematic understanding of the return generated by NIH investment.

Assess impact at the portfolio level

Portfolio-level assessment is particularly important for public impact.

A single project may contribute only one component of a much larger pathway. At the level of an institute, disease area, program, or NIH as a whole, however, the cumulative relationship between federal investment and downstream activity becomes more visible.

NIH could therefore develop public-impact profiles for different portfolios rather than attempting to reduce all impact to one score.

For example, one portfolio may demonstrate substantial influence on clinical practice. Another may primarily create foundational resources used throughout biomedical research. Another may generate significant economic and industrial activity. Another may contribute heavily to public-health policy.

These are different forms of mission value, not necessarily different levels of value.

This approach would also enable NIH to demonstrate impact consistently across Institutes and Centers and research areas rather than depending primarily on highly visible individual examples.

Treat public trust as an outcome requiring particular care

Public trust warrants special treatment because it is both central to NIH's mission and particularly difficult to measure.

NIH has explicitly connected rigor and reproducibility with strengthening confidence and trust in biomedical research.

However, public trust should not be reduced to popularity, positive sentiment, media volume, or social-media engagement.

Trustworthy science and popular science are not necessarily the same thing.

More defensible indicators may include evidence of:

- transparent communication of methods, limitations, and uncertainty;
- accessibility of underlying research and supporting evidence;
- correction of the scientific record when findings change;
- responsible communication of negative or conflicting findings;
- meaningful engagement with affected communities;
- accessibility of research findings to intended audiences; and
- demonstrable use of NIH-supported evidence by trusted clinical, public-health, and professional institutions.

Where surveys or public-opinion measures are used, they should be interpreted carefully and should not create incentives for researchers or NIH to optimize scientific communication for popularity rather than accuracy.

NIH's own current replication and reproducibility efforts recognize that rigorous, transparent, reliable science is foundational to public confidence. Public-impact assessment should reinforce that relationship.

Distinguish public communication from public impact

Media, news, online attention, and other public-facing sources can provide useful evidence, but they should be interpreted according to what they actually measure.

Media attention can identify research entering public discourse.

It does not establish that the research changed policy, clinical practice, health outcomes, or public understanding.

NIH could therefore use public-attention indicators as pathway signals rather than endpoint measures.

For example: “research ➜ public attention ➜ stakeholder engagement ➜ policy consideration ➜ policy adoption” may represent a legitimate pathway worth examining.

But attention alone should not be rewarded as equivalent to demonstrated uptake.

This distinction would also reduce incentives for sensationalism and help ensure that responsible communication of uncertainty remains compatible with NIH's impact framework.

Identify where evidence fails to become impact

One of the most valuable applications of a connected public-impact framework may be identifying translation and implementation gaps.

NIH traditionally asks whether research generated important knowledge. Connected evidence could increasingly help it ask what happened next.

For example:

- strong scientific evidence ➜ no guideline uptake
- guideline incorporation ➜ limited clinical adoption
- effective intervention ➜ limited accessibility
- valuable public-health evidence ➜ inconsistent implementation

These gaps matter because generating knowledge is only one part of delivering value to the American public.

NIH could use portfolio-level analysis to identify areas where strong research evidence exists but downstream uptake appears limited, providing an evidence base for considering whether additional implementation research, dissemination, partnerships, accessibility efforts, or other interventions may be warranted.

This would transform public-impact assessment from a retrospective reporting exercise into a tool for identifying where additional NIH action could help research generate greater public value.

Build upon existing evidence infrastructure

Many components required for this framework can already be observed through existing research information infrastructure.

Research can increasingly be connected across grants, publications, clinical trials, patents, policy documents, clinical guidelines, researchers, institutions, and public-facing sources. Combining these evidence streams provides a practical starting point for reconstructing pathways from federal investment toward downstream uptake.

Digital Science's experience working with research funders and institutions demonstrates the feasibility of this type of cross-source analysis at portfolio scale. Connected research information can link funding and scholarly outputs with downstream evidence such as patents, clinical activity, policy references, and public attention, helping organizations move beyond individual impact anecdotes toward systematic evidence of where research is being taken up and used.

This type of infrastructure could provide NIH with an existing evidence base for testing and validating public-impact indicators across different Institutes, Centers, disease areas, disciplines, and timescales rather than requiring a new reporting system to be created before assessment can begin.

Avoid creating an "impact burden"

NIH should be cautious about requiring every investigator to produce an extensive narrative demonstrating public impact.

Such an approach could create substantial administrative burden, advantage investigators with access to professional communications or research-impact staff, and encourage overstatement of downstream effects.

Where possible, NIH should derive evidence from existing research information, persistent identifiers, administrative systems, and downstream sources and allow investigators to

provide contextual or qualitative evidence where automated information is incomplete.

Assessment should also recognize differences among disciplines, institution types, career stages, and research objectives. Early-career researchers and investigators conducting foundational science should not be disadvantaged because the societal effects of their work may not become visible for many years.

Measure value to the public, not simply visibility

Ultimately, public impact assessment should help NIH answer a straightforward but difficult question: What happened because the knowledge NIH supported entered the world?

Answering that question requires more than counting publications, citations, patents, news stories, or social-media mentions.

NIH should examine whether research was noticed, taken up, applied, and ultimately associated with meaningful health, economic, clinical, accessibility, or societal benefit, while recognizing that different forms of research travel through different pathways and on different timescales.

Public trust in NIH-funded science also depends upon NIH being able to demonstrate these contributions consistently, not only through a small number of highly visible success stories, but across the breadth of its research portfolio.

By connecting evidence from funding, scholarship, clinical practice, policy, innovation, public health, and public engagement, NIH can develop a more rigorous account of the value generated by federal research investment: one that distinguishes attention from impact, contribution from attribution, and anecdote from systematic evidence.

• **Other Areas:**

Digital Science Response:

One overarching opportunity cuts across each of the areas addressed in this RFI: the need to shift from metric-based assessment toward evidence-based assessment.

For many decades, research evaluation has relied heavily on a relatively small number of publication- and citation-based indicators because these represented some of the most consistently available evidence of scientific activity. The research landscape has changed fundamentally.

Contemporary biomedical research now generates rich, interconnected evidence spanning grants, publications, datasets, software, protocols, methods, clinical trials, patents, policy documents, clinical guidelines, contributor roles, researchers,

organizations, and collaborative relationships.

The opportunity for NIH is therefore not to identify a new metric to replace publication counts or citations. It is to integrate complementary evidence streams to develop a more complete understanding of how scientific knowledge is generated, validated, shared, reused, extended, translated, and ultimately contributes to public benefit.

Move from counting outputs to understanding pathways

Across the topics addressed in this RFI, a common principle emerges: individual outputs provide only a partial view of scientific contribution.

NIH increasingly has the opportunity to understand the pathways connecting its investments to subsequent scientific activity and value.

For example:

Rigor and reproducibility: research → transparency → reproducibility → independent verification → replication → more reliable knowledge

Data, software, and models: creation → sharing → stewardship → discovery → reuse → new research and scientific value

Training and mentorship: training investment → mentorship → researcher development → continued scientific contribution → development of subsequent researchers

Collaboration: investment → complementary expertise → scientific integration → shared outputs and capabilities → durable research networks → downstream value

Entrepreneurship and translation: fundamental discovery → validation → translational development → cross-sector engagement → application → adoption

Foundational scientific exploration: exploratory investment → new concept, method, resource, or capability → diffusion → new scientific directions → subsequent discovery

Public impact: research → attention → uptake → application → health, economic, clinical, or societal benefit

These pathways are neither universal nor linear, and they should not become rigid models against which every research project is evaluated. Their value is conceptual: they illustrate why assessment focused predominantly on individual outputs can miss much of the value created by federal research investment.

NIH should increasingly ask not only "What was produced?" but "What did it enable?"

Assess contribution without overstating attribution

A second principle running across these topics is the importance of distinguishing contribution from attribution.

Biomedical progress is cumulative. Important discoveries frequently depend upon decades of foundational research, multiple investigators and institutions, shared datasets and infrastructure, methodological advances, validation studies, collaborations, funding programs, and subsequent translational activity.

Assessment systems should resist the temptation to attribute complex outcomes to individual publications, grants, or investigators simply because those relationships are easiest to measure.

Connected evidence instead provides an opportunity to identify credible contributions to longer scientific pathways.

This distinction is particularly important when assessing societal impact. NIH should be able to demonstrate that federally funded research contributed to advances in health, clinical practice, innovation, policy, and economic activity without implying causal attribution where the available evidence cannot support it.

Recognize the scientific contributions that conventional metrics miss

Scientific progress emerges through the interaction of people, infrastructure, methods, data, software, organizations, and institutions over extended periods.

Many activities that make this system function remain poorly represented in conventional assessment, including:

- stewardship of shared datasets and research resources;
- development and maintenance of scientific software;
- methodological and statistical contributions;
- creation and maintenance of research infrastructure;
- mentorship and supervision;
- peer review;
- replication and validation;
- development of standards and protocols;
- technical and collaborative leadership; and
- community building and scientific stewardship.

These activities should not automatically receive credit simply because they can be documented. Rather, NIH should develop mechanisms for recognizing when they make meaningful contributions to research quality, capability, efficiency, reliability, or impact.

Persistent identifiers, structured contributor roles, interoperable metadata, and connected research information increasingly make these contributions observable.

The challenge is to convert that increased visibility into responsible evidence for assessment, rather than simply creating additional things to count.

Treat research assessment as a scientific endeavor

NIH should also treat research assessment itself as an object of scientific investigation.

Any indicator creates incentives. Poorly designed indicators can change researcher behavior in ways that undermine the objective the indicator was intended to support.

New assessment approaches should therefore be treated as hypotheses to be tested rather than metrics to be institutionalized immediately.

Candidate indicators should be evaluated for:

- construct validity: whether the indicator actually measures the contribution it is intended to represent;
- robustness across disciplines, research types, institutions, and career stages;
- sensitivity to differences in scientific timescale;
- completeness and quality of the underlying data;
- susceptibility to gaming or strategic behavior;
- potential inequities or structural biases;
- administrative burden;
- unintended behavioral consequences; and
- incremental value beyond information already available to expert reviewers and program staff.

Where possible, NIH should test proposed indicators retrospectively before using them prospectively.

For example, NIH could ask whether proposed indicators of emerging science would have identified historically important research directions before conventional metrics did; whether proposed translational indicators reliably distinguish meaningful progression from incidental activity; or whether proposed mentorship indicators accurately reflect successful researcher development across different career pathways.

This would allow NIH to validate the assessment system before attaching significant incentives to it.

Use indicators as evidence, not answers

No quantitative indicator should substitute for scientific judgment.

Connected evidence, natural language processing, network analysis, persistent identifiers,

and other analytical approaches can make patterns visible that would be impossible for an individual reviewer or program officer to observe across an entire research system.

Their purpose should be to improve the evidence available to decision-makers, not to automate decisions that require scientific judgment.

This distinction is particularly important as the volume and sophistication of research information increase.

The appropriate model is: data ➜ evidence ➜ interpretation ➜ expert judgment ➜ decision rather than: data ➜ score ➜ decision.

NIH should be cautious about composite scores that collapse fundamentally different dimensions of scientific contribution into a single ranking.

A researcher, project, program, or portfolio can be strong in one dimension and intentionally different in another. Foundational research should not be penalized because it lacks immediate translational impact. A high-value dataset should not be judged as though it were a publication. A collaborative specialist should not be assessed according to PI-oriented measures. A high-risk research portfolio should not be expected to produce the same success pattern as a program designed around predictable deliverables.

Evidence-based assessment should preserve those distinctions.

Evaluate portfolios as well as people

Many of the questions raised by this RFI cannot be answered adequately at the level of an individual investigator or award.

NIH should therefore complement researcher- and project-level assessment with portfolio-level assessment.

At portfolio scale, NIH can examine questions such as:

- Are investments producing reusable research resources?
- Are training programs generating durable scientific capacity?
- Are collaborative programs producing genuinely integrated research networks?
- Where are translational pathways progressing or stalling?
- Is strong scientific evidence reaching clinical practice and public-health application?
- Are emerging scientific areas receiving appropriate attention?
- Is the balance between established and exploratory research changing over time?
- Which NIH investments continue generating value long after the originating awards have ended?

This changes assessment from an exercise focused primarily on evaluating researchers into a capability NIH can also use to evaluate and improve its own funding strategy.

That distinction is important. Responsible research assessment should not only ask whether researchers are producing the right behaviors and outcomes. It should also allow the funding agency to ask whether its own policies, programs, incentives, and portfolio decisions are producing the research ecosystem it intends to create.

Build upon shared infrastructure and community standards

Continued investment in open scholarly infrastructure, persistent identifiers, trusted repositories, and community standards remains essential.

Digital Science has participated in this work directly, including through NIH's Generalist Repository Ecosystem Initiative (GREI). Initiatives of this kind demonstrate the value of interoperable infrastructure that enables research outputs to remain discoverable, accessible, connected, and reusable beyond the systems or projects in which they originated.

Shared standards such as DOIs, ORCID iDs, ROR identifiers, RRIDs, grant identifiers, structured contributor roles, and interoperable metadata improve not only assessment but also reproducibility, transparency, collaboration, data stewardship, and research efficiency.

NIH should continue supporting this foundational infrastructure while taking advantage of the broader public, community, institutional, and commercial research-information ecosystem that has developed around it.

Where evidence can be obtained reliably from existing infrastructure, researchers should not be required to report it again.

The goal should be a richer evidence base with a lower reporting burden, not a larger assessment bureaucracy.

Recognize that research is an ecosystem

The questions in this RFI span researchers, funders, institutions, publishers, repositories, healthcare organizations, government, industry, and the public because modern biomedical research operates across all of them.

Assessment frameworks designed from the perspective of any one participant risk capturing only part of that system.

Digital Science's perspective on these questions has been shaped by more than two decades of working across this research ecosystem—including with researchers, universities and research institutions, publishers, research funders, governments, and commercial organizations, including pharmaceutical and biotechnology companies involved in translating scientific discovery toward application.

That cross-ecosystem perspective informs a central theme of this response: scientific

value frequently becomes visible in the connections between systems, organizations, outputs, and stages of the research lifecycle, rather than within any single source of information.

Connected research infrastructure can increasingly make those relationships observable at scale. But infrastructure alone is not an assessment framework. The evidence it provides must be interpreted in the context of scientific purpose, disciplinary differences, timescale, uncertainty, and the intended mission of the research.

From metrics to evidence

The challenge facing research assessment is no longer simply one of measurement. It is increasingly one of interpretation.

As the evidence describing the research ecosystem becomes richer and more connected, NIH should resist replacing one dominant metric with another.

Instead, the opportunity is to develop an assessment framework that asks:

- What contribution are we trying to understand?
- What evidence would credibly demonstrate that contribution?
- Over what timescale should it be observed?
- What contextual information is required to interpret it?
- What behaviors might measuring it incentivize?

What can quantitative evidence tell us, and what still requires expert scientific judgment?

Across the areas addressed in this RFI, this approach would allow NIH to move from counting isolated outputs toward understanding the pathways through which research creates scientific and public value.

The resulting framework would not eliminate metrics. It would put them in their proper role: as individual pieces of evidence within a broader, scientifically grounded assessment system.

That distinction is essential.

NIH does not need a new definition of scientific success expressed through a different score. It needs an evidence framework capable of recognizing the many ways rigorous, collaborative, exploratory, reusable, translational, and publicly valuable science contributes to its mission, and capable of evolving as science itself changes.

186. STM (The International Association

I am responding: On behalf of an organization

Name: Caroline Sutton

Name of Organization: STM (The International Association

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: STM welcomes NIH's efforts to strengthen rigor, reproducibility, and replication across the biomedical research enterprise. These goals are central to public trust in science and to NIH's mission of generating knowledge that enhances health, lengthens life, and reduces illness and disability. NIH's Replication and Reproducibility Initiative recognizes the need to elevate replication and reproducibility studies as foundational to "gold standard science" and identifies data standardization, public availability of research findings, independent replication research, and sustained infrastructure as part of NIH's approach.

These goals also mirror concrete efforts and contributions by STM and its member publishers across the research ecosystem. For example, STM's research integrity work supports editorial standards, integrity checks, stewardship of the Version of Record, post-publication corrections and retractions, and transparent communication when concerns arise. STM's Integrity Hub provides a shared infrastructure to help publishers detect paper-mill activity, duplicate submissions, references to retracted work, and other research-integrity risks and ensure this material does not enter the scientific record. STM's research data initiatives catalyze action to advance FAIR data practices by promoting data availability policies, links between articles and datasets, persistent identifiers, community metadata standards, repository workflows, and best practices for data citation and reuse. These publisher-led activities are complemented by third-party frameworks such as COPE guidance on publication ethics and stewardship of the scholarly record, the Center for Open Science TOP Guidelines on transparency and verifiability, and responsible research-assessment initiatives such as DORA and CoARA. Together, these efforts demonstrate that reproducibility depends not only on actions by individual researchers, but also on the shared systems, standards, editorial practices, and infrastructure that make research findings reliable, discoverable, correctable, reusable, and trustworthy. They also demonstrate that publications are a key part of this system.

STM encourages NIH to adopt a broad, evidence-based framework for recognizing and supporting activities that improve the reliability, transparency, and trustworthiness of biomedical research. Potential indicators include:

• Publication of replication studies, negative results, null findings, and confirmatory research. Many publishers provide options to support the communication of such reports.

• Use of preregistration, registered reports, reporting guidelines, and other transparency-enhancing practices. Such practices are increasingly supported and enabled by publishers.

• Availability of underlying data, code, protocols, and materials, where appropriate. STM, in coordination with others in the research community, have provided best practices for supporting code and data

sharing.
• Adherence to field-specific reporting standards and methodological best practices. These are often communicated and stewarded by journal communities sustained by publishers.
• Reliance on the integrity-enhancing practices of publishers, including peer review, editorial assessment, research integrity screening, corrections, retractions, and post-publication stewardship of the scientific record.
• Documented efforts to reproduce, validate, or extend prior research findings, which could be included in journal articles or elsewhere.

NIH should be careful not to replace one narrow assessment framework with another, nor should NIH ignore the importance of publications in achieving NIH's core mission of supporting high-quality research and ensuring its impact. While overreliance on publication counts, citation metrics, or other individual-level indicators can distort incentives, under reliance on the scholarly record as an integral contributor to the impact of research would be inappropriate.

With both traditional and alternative metrics, care must be taken to ensure that any metric is not overly rigid or treated as compliance targets rather than contextual evidence of scientific contribution. As noted in DORA's guidance on responsible use of quantitative indicators, indicators are inherently reductive and require careful contextualization when used in research assessment. CoARA similarly seeks to reform assessment systems, promoting a broader understanding of research quality and contribution, but not ignoring the importance of publications as expressions of research quality and impact.

Integrity depends on coordinated action across the research ecosystem, with publishers, funders, institutions, policymakers, and researchers fulfilling distinct but interconnected responsibilities. Accordingly, NIH should recognize not only researchers who produce reproducible findings, but also contributions that strengthen the systems through which reproducible science is evaluated, communicated, corrected, preserved, and reused. This is particularly important because replication and reproducibility depend on access to reliable, well-documented, and discoverable research outputs. Research cannot be reproduced or validated if findings, methods, data, and related materials are not effectively communicated and maintained. This again emphasizes the importance of publication to achieving NIH's core goals.

NIH should also account for disciplinary variation. Replication may look different in clinical research, laboratory science, computational biology, population health, epidemiology, and translational research. A flexible framework that recognizes field-specific norms and methodological constraints will be more effective than a uniform checklist.

STM and its members would welcome opportunities to collaborate with NIH on incentives and infrastructure that achieve our shared goals of advancing trusted research.

References and resources:

STM Research Integrity: https://stm-assoc.org/what-we-do/strategic-areas/research-integrity/
STM Integrity Hub: https://stm-assoc.org/what-we-do/strategic-areas/research-integrity/integrity-hub/
STM Research Data Program: https://stm-researchdata.org/
Joint Statement of Best Practices for Research Data Sharing: https://stm-assoc.org/what-we-do/strategic-areas/open-research/research-data/best-practices/
COPE Guidance and Tools: https://publicationethics.org/guidance-and-tools
Center for Open Science TOP Guidelines: https://www.cos.io/initiatives/top-guidelines
DORA Guidance on Responsible Use of Quantitative Indicators in Research Assessment:

https://sfdora.org/resource/guidance-on-the-responsible-use-of-quantitative-indicators-in-research-assessment/
CoARA Commitments: https://www.coara.org/agreement/the-commitments/

2. Data, Software, and Model Sharing: STM supports NIH's efforts to incentivize and reward responsible sharing of data, software, models, and related research objects. Effective sharing can accelerate discovery, enable validation, promote reuse, and increase the return on public investment in biomedical research. NIH's Public Access materials state that access to NIH-supported research can accelerate research, generate higher quality scientific results, encourage greater scientific integrity, and enable future inquiry, discovery, and translation. STM agrees, and this is a key focus of STM's research data initiatives and many of our members' commitments to support data sharing, linking, citation, and use. As NIH seeks to incentivize and reward these activities, it should again take a comprehensive approach, incentivizing not just sharing but appropriate sharing, citation, linking, and use. The STM, DataCite, and Crossref's "12 Best Practices for Research Data Sharing" provides a roadmap for assessment of whether data is shared in ways that make them findable, accessible, interoperable, reusable, and appropriately stewarded. Potential indicators include:

- Deposition of datasets, software, models, protocols, and related materials in trusted repositories and linking to the primary literature.
- Use of persistent identifiers, metadata, and community standards to improve findability and reuse.
- Evidence that shared resources have been accessed, reused, cited, and incorporated into downstream research.
- Clear documentation, versioning, licensing, and maintenance practices.
- Use of appropriate safeguards for privacy, confidentiality, security, intellectual property, and ethical obligations.
- Contributions to infrastructure, standards, tools, and workflows that make data, software, and model sharing more reliable and useful.

NIH should avoid incentives that reward superficial compliance. For example, measuring only the number of datasets deposited could encourage quantity over quality. A dataset that is incomplete, poorly documented, difficult to access, unconnected from scholarly literature, or not maintained may do little to advance scientific impact. NIH should instead encourage meaningful sharing that supports validation, reuse, and long-term stewardship.
NIH should also recognize that not all data, models, or software can or should be shared in the same manner. Legal, ethical, privacy, security, proprietary, and participant-protection considerations may limit what can be made publicly available. Researchers should not be penalized when they apply appropriate protections or when responsible sharing requires controlled access rather than unrestricted access.
Scholarly publishers increasingly support these goals through data availability policies, repository integrations, links between articles and datasets, persistent identifiers, metadata improvement, software citation practices, reporting standards, and editorial checks. These practices help connect publications to the broader research objects that underlie them, making research more transparent, reproducible, and reusable.
STM's Research Data Program (started in 2020) helped some publishers implement these practices and still provides guidance to promote the sharing, linking, and citation of research data alongside scholarly articles. STM and its members heavily contributed to the development of the Scholix framework for linking publications with datasets by the Research Data Alliance (RDA) and continues to champion data linking and citation to ensure findability and utilization of research data. STM and its members are also

coordinating with related initiatives, such as the State of Open Data project, which tracks the evolution of open science practices globally.

As it considers changes to assessing research, NIH should consider how its assessment framework and policies align with its public access and metadata initiatives. NIH has stated that its efforts to increase public access are intended to make it easier for researchers, clinicians, students, and the public to find and use research results, including through metadata and persistent identifiers (PIDs). Publishers are critical to ensuring that articles and data related to NIH research are reliable, usable, findable, and preserved, including through peer review, editorial assessment, research integrity checks, transparency standards, corrections and retractions, metadata enhancement, PIDs, preservation, and maintenance of the Version of Record. It is therefore critical that NIH policy continue to ensure support for the publishing enterprise as well as assessment of data, software, and model sharing: impact depends not only on availability, but on usability, discoverability, integrity, and stewardship.

STM is developing voluntary recommendations to help journals and publishers clarify data sharing guidance for medical and health science researchers, where generic data sharing policies may be insufficient because of privacy, regulation, and ethics considerations. These recommendations are expected to be released later this year. STM and its members would welcome additional discussion on how we can support NIH's efforts to incentivize and reward these activities in line with our shared commitment to increase the impact of NIH-funded research on public health and improve the return on investments, particularly in the health and medical space.

References and resources:

STM Research Data Program: https://stm-researchdata.org/

Joint Statement of Best Practices for Research Data Sharing: https://stm-assoc.org/what-we-do/strategic-areas/open-research/research-data/best-practices/

State of Open Data: https://stateofopendata.com/

Scholix Framework: https://casrai.org/guides/scholix-scholarly-link-exchange-framework

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: STM agrees that public impact is a critical dimension of scientific success. NIH-funded research generates public value when it contributes to better health outcomes, improved clinical practice, innovation, economic growth, public trust, workforce development, and broader societal benefit. However, public impact should be assessed carefully, over appropriate time horizons, and with recognition that many important discoveries produce benefits only after years of cumulative research, validation, dissemination, and translation.

NIH should explicitly recognize the role of trusted scholarly communication in enabling public impact. Scientific discoveries produce public benefit only when they can be found, evaluated, trusted, corrected when necessary, and reused. Peer-reviewed publication provides an essential mechanism for making biomedical knowledge visible, scrutinized, structured, preserved, and connected to the broader scientific record.

NIH's own public access policies strongly support this point. In its Public Access Policy, NIH states that

increasing access to publications resulting from NIH funding offers benefits to the scientific community and the public, including patients, families, healthcare providers, researchers, students, and members of the public. NIH's Office of Science Policy adds that public access can accelerate research, generate higher quality scientific results, encourage greater scientific integrity, and enable future inquiry, discovery, and translation. These statements reflect NIH's longstanding recognition that publication and dissemination are not peripheral to impact; they are core pathways through which research investments generate public benefit.

The peer-reviewed Version of Record is especially important for public impact. It provides a stable, curated, and citable record of research findings; links findings to underlying evidence; supports metadata and indexing; enables discovery through trusted databases; and provides mechanisms for correction, expression of concern, and retraction when needed. These functions are essential to public trust and to the responsible use of biomedical research in clinical, policy, and innovation contexts.

Public impact should therefore be measured not only by immediate translational outcomes, commercial products, or media attention, but also by contributions to the trusted knowledge infrastructure that makes later public benefits possible. Articles reporting on basic research, replication studies, negative results, data resources, methods papers, standards development, and editorial or peer-review contributions may all generate significant public value even when their effects are indirect or long-term.

NIH should be particularly cautious about short-term indicators of public impact. Measures such as media coverage, downloads, social media attention, or public engagement may provide useful signals in some contexts, but they should not be treated as proxies for scientific quality, reliability, or health benefit. Public impact should be assessed through a portfolio of evidence that recognizes both immediate and long-term contributions.

Finally, NIH should recognize that the public benefits described above depend on sustained and appropriate support for the publishing enterprise. Peer review management, editorial assessment, research-integrity screening, metadata creation, indexing, preservation, data and article linking, public access workflows, corrections, retractions, and long-term stewardship of the Version of Record all require ongoing investment. Publication charges, subscriptions, and other sustainable publishing models are mechanisms through which the research ecosystem supports these services and infrastructure. If NIH wishes to maximize the public impact of NIH-funded research, it should avoid policies or assessment frameworks that inadvertently undermine the financial sustainability of trusted scholarly communication and should instead recognize reasonable publishing costs as part of the infrastructure required to make biomedical knowledge reliable, discoverable, reusable, and beneficial to the public.

References and resources:

STM's submission to NIH on limiting publication charges: https://stm-assoc.org/document/stm-submission-to-nih-rfi-allowable-publishing-costs-sept-2025/

STM's statement on proposed OMB revisions affecting publication and dissemination of federally funded research: https://stm-assoc.org/stm-statement-on-proposed-omb-revisions/

NIH Public Access Pages: https://grants.nih.gov/policy-and-compliance/policy-topics/public-access/nih-public-access-policy-overview and https://osp.od.nih.gov/policies/public-access/

8. Other Areas: STM encourages NIH to ensure that any future framework for measuring scientific impact explicitly recognizes publication and dissemination as fundamental components of the research process and as essential pathways for achieving NIH's mission. Fundamentally, NIH's mission is not accomplished when an experiment is completed. It is advanced when knowledge generated through

NIH-supported research can be evaluated, trusted, disseminated, replicated, translated, and ultimately used to improve health. Publication is one of the central mechanisms through which that process occurs. STM therefore urges NIH to recognize that the Version of Record, robust peer review, and ongoing integrity maintenance are essential to NIH's research priorities and to maximizing the return on NIH's scientific investments and to explicitly affirm that dissemination through trusted scholarly communication channels advances NIH's core mission, and provide appropriate support for those activities.

Several recent NIH statements and RFIs, including this one, appear to distinguish between conducting research and communicating research outcomes. STM believes that distinction should be approached with caution. While publication is a private sector activity that is not directly funded through core research activities, biomedical research does not achieve its full value at the point of discovery alone. It achieves value when findings are evaluated, communicated, preserved, disseminated, made discoverable, corrected when necessary, and used by others to advance science, clinical practice, innovation, and public health.

Publication is therefore not merely an administrative byproduct of research or a narrow productivity measure. It is one of the principal mechanisms through which scientific knowledge enters the research ecosystem. Research findings that are not effectively communicated cannot be scrutinized, replicated, synthesized, incorporated into guidelines, translated into products or practices, used by clinicians, or accessed by the public.

NIH's own policies support this conclusion. NIH's Public Access Policy ensures that the public has access to published results of NIH-funded research in PubMed Central. NIH's public access materials state that access to publications resulting from NIH-funded research benefits the scientific community and the public and supports patients, families, healthcare providers, researchers, students, and members of the public. NIH's Office of Science Policy further states that public access can accelerate research, generate higher quality scientific results, encourage greater scientific integrity, and enable future inquiry, discovery, and translation.

NIH's Replication and Reproducibility Initiative also recognizes the importance of making research findings freely and quickly available to the public as part of NIH's ongoing efforts to strengthen rigorous and reproducible science. This is because replication, validation, reuse, correction, and translation all depend on effective communication and preservation of research findings.

Accordingly, STM urges NIH not to interpret responsible research assessment reform as a reason to diminish the role of publication. Initiatives such as DORA and CoARA rightly caution against overreliance on publication counts, journal impact factors, and citation-based proxies. But these efforts do not imply that publication itself is unimportant. Rather, they support a more nuanced approach that assesses research and researchers based on the substance, reliability, transparency, openness, integrity, and societal value of their contributions. The appropriate conclusion is that assessment should recognize what high-quality publication and scholarly communication actually do: support peer review, editorial assessment, research integrity, discoverability, public access, metadata creation, preservation, correction, reuse, and long-term stewardship of the scientific record.

This point is also consistent with STM's recent submissions on federal research policy. STM has emphasized that research is not complete at the point of discovery; it achieves value only when findings are validated, communicated, and made accessible. In comments on NIH's strategic plan, STM has urged NIH to recognize that the Version of Record, robust peer review, and ongoing integrity maintenance are essential to NIH's research priorities and to maximizing the return on NIH's scientific investments.

STM is concerned that any policy framework suggesting publication does not advance NIH's core mission is inconsistent with NIH's own public access, reproducibility, metadata, and stewardship policies. Publication and dissemination are not separate from NIH's mission of improving health. They are among the principal mechanisms through which NIH-supported research becomes useful, trustworthy, accessible, and actionable.
In addition to recognizing the importance of publication to NIH's core mission, NIH should also ensure that policies related to allowable costs, research assessment, public access, and reproducibility are aligned. Specifically, NIH should continue to ensure that appropriate and sustained funding continues to be available for publishing costs and subscriptions as part of NIH grants. More information on these issues can be found in STM's previous submissions on these topics, which are linked below.

References and resources:
STM submission on NIH Strategic Framework: https://stm-assoc.org/document/stm-response-to-nih-strategic-framework-may-22-2026-docx/
STM's submission to NIH on limiting publication charges: https://stm-assoc.org/document/stm-submission-to-nih-rfi-allowable-publishing-costs-sept-2025/
STM's statement on proposed OMB regulations affecting publication and dissemination of federally funded research: https://stm-assoc.org/stm-statement-on-proposed-omb-revisions/

187. Medical Research Support

I am responding: On behalf of an organization

Name: EMİNE ARZU KANIK

Name of Organization: Medical Research Support

Type of Organization: Professional Organization/Association

Role: Institutional Review Oversight Committee Member

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Rigor and Reproducibility — Proposed Response

Biomedical research needs a system in which rigor and reproducibility are assessed not only through what investigators plan to do, but through verifiable evidence generated across the entire research lifecycle. We therefore recommend that NIH explore a lifecycle-based framework that complements existing grant and publication assessment with structured, transparent, and time-bounded scientific evaluation.

One component of such a framework could be a Time-Bounded Open Scientific Review model. For appropriate studies, before data collection or analysis begins, a registered protocol could enter a defined 30–60 day review period. Qualified reviewers would provide structured assessments of the scientific rationale, study design, causal assumptions, potential sources of bias, sample-size justification, outcome definitions, confounding, missing-data strategies, statistical analysis plans, ethical considerations, feasibility, and plans for data, code, model, and metadata sharing.

Importantly, this process should not become a popularity-based voting system. Reviewers should provide structured scores accompanied by methodological justification, disclose relevant expertise and conflicts of interest, and remain accountable for the quality of their assessments. Investigators should have an opportunity to respond to the reviews, and the protocol should be revised through a transparent, time-stamped version history. The original protocol, reviewer comments, author responses, revisions, and documented reasons for accepting or declining substantive recommendations should become part of the scientific record.

This approach could create a measurable distinction between a research plan that merely states that rigorous methods will be used and a research project for which methodological rigor has been independently examined and improved before the study is conducted.

A second review point could occur when the study is completed and the manuscript is prepared. A time-bounded open review of the resulting research output could evaluate not only the findings but also adherence to the original protocol, deviations and their justification, analytical reproducibility, interpretation of uncertainty, completeness of reporting, and readiness for independent replication.

A particularly useful indicator would be registry-to-publication concordance. Where a study has a registered protocol, NIH could evaluate whether primary and secondary outcomes, sample-size targets, and major elements of the analysis plan correspond to the final publication, while recognizing and documenting legitimate amendments. This would provide a transparent measure of selective reporting and protocol adherence without assuming that every deviation represents misconduct.

NIH could also promote machine-readable implementation of established reporting frameworks such as CONSORT, STROBE, PRISMA, and ARRIVE, adapted to the relevant study design. Such structured checks could identify missing methodological information with relatively low administrative burden while leaving substantive scientific judgment to human experts.

Prioritizing Replication and Reproducibility

Not every published finding requires the same level of replication effort. NIH could develop a risk- and impact-informed framework for prioritizing replication and reproducibility studies.

Priority could be given to findings that:

have potentially substantial implications for clinical practice, public health, or policy;
depend heavily on a single study, laboratory, dataset, software pipeline, model, or population;
have conflicting results or substantial between-study heterogeneity;
have limited statistical power, imprecise effect estimates, or substantial analytical flexibility;
lack accessible data, code, software, models, metadata, or sufficient methodological documentation;
have not undergone independent computational reproduction or external validation;
show substantial unexplained deviations from prospectively specified protocols;
have achieved extensive citation or use despite limited independent validation; or
could cause substantial research waste or public harm if incorrect.

Replication should therefore be viewed not as a universal requirement applied identically to every study, but as a strategic allocation of scientific resources toward findings for which independent confirmation would provide the greatest value.

What Should Be Measured?

Potential indicators of rigor and reproducibility could include:

Documentation of the research protocol and statistical analysis plan before data collection or analysis, where appropriate.
Concordance between registered protocols and published outcomes and analyses.
Methodological completeness according to study-design-appropriate reporting standards.
Number, nature, and justification of substantive protocol deviations.
Quality of investigator responses to methodological review.
Documented improvements resulting from scientific review.
Availability and usability of data, metadata, analysis code, software, and models, subject to ethical and

legal constraints.
Successful independent computational reproduction.
Independent replication or external validation.
Transparent reporting of negative, inconclusive, and conflicting findings.
Correction and updating of the scientific record when errors are identified.
Reuse of research outputs by independent investigators.

These indicators should support scientific judgment rather than automatically determine scientific value. A single numerical score should not replace expert assessment, nor should researchers be incentivized to optimize metrics rather than improve science.

Incentives and Infrastructure

NIH could recognize rigorous research practices through citable records for protocols, methodological reviews, datasets, software, models, and replication studies. Contributions to independent replication, methodological review, research infrastructure, and correction of the scientific record should be visible in investigators' scientific portfolios and considered in progress reports and future funding decisions.

Incentives could also include protected time for replication, dedicated funding mechanisms, and infrastructure supporting version control, persistent identifiers, data and code sharing, computational environments, and transparent scientific review.

At the same time, safeguards are essential. The system should monitor reviewer burden, inter-reviewer agreement, conflicts of interest, retaliation or inappropriate review behavior, disadvantages affecting early-career researchers or smaller institutions, administrative burden, and potential gaming of the metrics.

A Practical Pilot

We recommend that NIH initially test such a framework through a voluntary, independently evaluated pilot involving a defined cohort of researchers across different NIH Institutes, disciplines, career stages, and institution types.

The pilot could compare conventional research workflows with a lifecycle-based framework incorporating protocol review, structured methodological assessment, transparent revision, reproducibility checks, and prioritized replication. Outcomes could include methodological quality, errors detected before publication, reporting completeness, researcher burden, reviewer burden, inter-reviewer agreement, time to scientific output, computational reproducibility, and subsequent independent validation.

MedicReS would be willing to contribute to such a pilot. For nearly two decades, MedicReS has worked on research methodology, biostatistics, research integrity, scientific reporting, and researcher development. Our experience with Good Medical Researcher training and the development of connected research-support environments such as EPICOSAI provides practical experience across

multiple stages of the research lifecycle.

We would be particularly interested in supporting a one-year pilot cohort at no commercial financial obligation to NIH or participating researchers, subject to mutually agreed governance, data protection, interoperability, and independent evaluation requirements. EPICOSAI should not be viewed as a proprietary requirement for such a system; rather, it could serve as one practical environment through which the proposed workflow could be tested.

We would welcome the opportunity to discuss the design of such a pilot with NIH in a scientific meeting or Zoom discussion, including cohort selection, evaluation metrics, governance, interoperability, and independent assessment of outcomes.

The ultimate objective should not be to create another publication metric. It should be to create a research environment in which rigor is visible, methodological improvement is documented, reproducibility is rewarded, replication is strategically supported, and the scientific record can continuously improve.

2. Data, Software, and Model Sharing: Data, Software, and Model Sharing — Proposed Response

NIH should broaden the concept of data, software, and model sharing from making research objects available to demonstrating that they are understandable, verifiable, reproducible, and reusable.

A dataset deposited in a repository, a software package made available on a public platform, or a model accompanied by a download link should not automatically be considered a successful scientific contribution. The more meaningful question is whether another research team can understand the research object, determine how it was generated, verify its intended use, reproduce relevant analyses, and build upon it.

We therefore recommend that NIH develop a Research Object Impact Framework that evaluates data, software, and models throughout their lifecycle.

1. From “Shared” to “Usable”

For appropriate NIH-funded projects, investigators could be encouraged to create a version-controlled research object containing, as applicable:

the dataset or an appropriate controlled-access representation;
data dictionaries and comprehensive metadata;
analysis code;
software and package versions;
computational dependencies;
model versions and model documentation;
information about model development and validation;
performance and uncertainty measures;
known limitations and intended-use conditions;

provenance and version history;
instructions for reproducing key analyses; and
documentation describing how the resource can appropriately be reused.

The objective should not be to impose identical technical requirements on every project. Requirements should be proportionate to the research design, data sensitivity, computational complexity, and scientific purpose.

2. FAIR Should Be Operational, Not Merely Declarative

The FAIR principles — Findable, Accessible, Interoperable, and Reusable — provide a useful foundation. However, compliance should not be reduced to checking whether a file has been deposited in a repository.

NIH could develop measurable indicators for the functional usability of shared research objects.

For example:

Can an independent researcher locate the resource?
Can its contents and structure be understood?
Are the metadata sufficient to interpret the data correctly?
Can the researcher determine which version was used in the original analysis?
Can the published analysis be rerun?
Can the software environment and dependencies be reconstructed?
Can a model's performance and limitations be independently assessed?
Can another research group appropriately reuse the resource?
Has independent reuse actually occurred?

This distinction is important because availability is an input, whereas successful independent reuse is evidence of scientific impact.

3. Reproducibility of Computational Research

For computationally intensive research, NIH could encourage preservation not only of source code but also of the computational environment required to reproduce key analyses.

Where appropriate, this could include:

software and package versions;
dependency specifications;
version-controlled code;
computational workflows;
containerized environments;
model versions;

randomization or seed information where relevant;
documentation of preprocessing and analytical pipelines; and
machine-readable descriptions of the workflow.

The objective should be to make the computational pathway from data to reported result sufficiently transparent that an independent researcher can evaluate and reproduce it.

Importantly, reproducibility should not require identical hardware or an identical computational environment in every circumstance. The appropriate standard should be scientifically meaningful reproducibility, with deviations or technical limitations documented transparently.

4. Model Sharing Requires More Than Sharing Model Files

The rapid expansion of AI and computational models makes this particularly important.

A model should not be evaluated solely on whether its weights or source code are publicly accessible. For meaningful scientific reuse, researchers need information about:

the purpose and intended use of the model;
development and validation datasets or their relevant characteristics;
inclusion and exclusion criteria;
preprocessing procedures;
model version;
performance measures;
uncertainty and calibration;
external validation;
known limitations;
population and setting in which the model was evaluated;
potential sources of bias; and
appropriate and inappropriate uses.

For biomedical AI, model documentation should therefore become part of the scientific contribution itself.

5. Measure Reuse, Not Download Counts

NIH should avoid relying heavily on easily manipulated indicators such as download counts, repository views, or the number of files deposited.

More meaningful indicators could include:

Independent computational reproduction.
Independent reuse of a dataset or research resource.
New publications based on the resource.

Independent validation studies.
Replication or extension of the original findings.
Reuse across institutions or scientific disciplines.
Development of new methods, tools, or models based on the resource.
Identification and correction of errors through independent reuse.
Documented contribution to clinical, translational, or public-health research.
Long-term maintenance and continued usability of the resource.

A resource that is downloaded thousands of times but cannot be understood or reproduced may have less scientific value than a smaller resource that is successfully reused and independently validated.

6. Recognize the People Who Build Research Infrastructure

Data curation, software engineering, statistical programming, model documentation, computational validation, database development, and long-term maintenance are scientific contributions.

NIH should make these contributions visible in investigators' scientific portfolios and in grant evaluation.

Researchers who create and maintain high-value datasets, software packages, computational pipelines, registries, biobanks, harmonized protocols, or reusable models should not be disadvantaged simply because these contributions produce fewer conventional journal articles.

Persistent identifiers and structured contributor attribution could allow these contributions to become citable and independently verifiable scientific outputs.

This would also support recognition of technical and methodological contributors who are often essential to research but remain invisible in conventional publication-based assessment.

7. Controlled Access Should Be Considered a Successful Form of Sharing

Not all biomedical data can or should be openly downloadable.

Patient-level data, genomic data, sensitive health information, and other protected datasets may require controlled access. NIH should therefore distinguish between open accessibility and responsible accessibility.

A dataset should not be considered scientifically unsuccessful because privacy, consent, security, intellectual-property, or regulatory requirements prevent unrestricted access.

Instead, evaluation should consider whether investigators have provided an appropriate access pathway, sufficient metadata, documentation, governance procedures, and a transparent mechanism through which qualified researchers can request access.

The principle should be:

As open as possible, as restricted as necessary.

8. Reward Reuse and Maintenance

Scientific recognition should continue beyond the initial publication.

NIH could provide explicit recognition for:

independent reuse;
successful replication using shared resources;
software maintenance;
dataset updates;
correction of errors;
model recalibration;
external validation;
development of derived resources; and
sustained infrastructure contributions.

Progress reports and future funding decisions could incorporate evidence that research outputs remained useful beyond the original project.

This would create a powerful incentive to produce durable scientific infrastructure rather than one-time research outputs.

9. Interoperability Should Be a Core Principle

NIH-supported resources should, where feasible, use interoperable standards and machine-readable formats that allow integration across research environments.

This is particularly important for multidisciplinary biomedical research, where datasets, registries, imaging resources, electronic health records, statistical software, AI models, and clinical research platforms increasingly need to interact.

Interoperability should therefore be evaluated as a scientific infrastructure contribution rather than treated solely as a technical specification.

10. A Practical Pilot

We recommend that NIH test these principles through a voluntary pilot involving a diverse cohort of funded research projects.

The pilot could evaluate shared datasets, software, and models using a common but adaptable framework and compare:

Availability → Documentation → Verification → Reproducibility → Independent Reuse → Scientific Impact

The pilot should measure not only whether resources were shared, but whether independent researchers could successfully understand, reproduce, validate, and reuse them.

Outcomes could include:

time required for independent reuse;
number and type of reproducibility barriers identified;
successful computational reproduction;
independent validation;
documented reuse;
new scientific outputs generated from the resource;
maintenance burden;
researcher and infrastructure costs; and
unintended consequences.

The framework should be independently evaluated and remain open, provider-neutral, interoperable, and adaptable to different research domains.

MedicReS Perspective

MedicReS has long approached medical research as a connected lifecycle rather than as a sequence ending with publication. Our experience in research methodology, biostatistics, scientific reporting, and research-support environments such as EPICOSAI has reinforced an important practical principle: the value of a research output depends not only on whether it exists, but on whether its methodological pathway and underlying research objects can be understood and appropriately reused.

EPICOSAI can provide practical experience relevant to this concept, particularly in connecting research questions, study design, statistical planning, analysis, reporting, and reproducibility-oriented workflows. However, we do not propose EPICOSAI as a proprietary NIH requirement. Any NIH framework should be based on open standards and should allow multiple interoperable platforms to participate.

We would be willing to contribute to an NIH pilot and, if useful, provide access to relevant research-support capabilities for a defined cohort during a one-year evaluation period without a commercial financial objective, subject to appropriate governance, privacy, security, interoperability, and independent evaluation requirements.

Conclusion

NIH has an opportunity to move the scientific community from a culture of “share because the policy requires it” toward a culture of “build research objects that others can trust, reproduce, and reuse.”

The most meaningful measure of data, software, and model sharing is therefore not simply:

Was it deposited?

but:

Could another researcher understand it, verify it, reproduce it, reuse it, and create new scientific value from it?

That is where sharing becomes measurable scientific impact.

3. Training and Mentorship: NIH should distinguish between training activity and training impact.

Courses completed, training hours, certificates awarded, number of trainees, or number of mentoring meetings are useful measures of participation, but they do not necessarily demonstrate that training has strengthened the biomedical research workforce.

The more meaningful question is:

What can a researcher demonstrably do after training or mentorship that they could not do independently before?

We therefore recommend that NIH complement existing training indicators with a competency- and independence-based framework that follows researchers across the research lifecycle.

From Participation to Demonstrated Research Competence

For biomedical researchers, measurable competencies could include the ability to:

formulate a meaningful and answerable research question;
distinguish exploratory from confirmatory research;
select an appropriate study design;
develop a scientifically and ethically sound protocol;
define appropriate outcomes and estimands;
justify sample size and statistical power;
develop a prespecified statistical analysis plan;
recognize major sources of bias, confounding, and missing-data problems;
manage research data appropriately;
conduct or supervise reproducible analyses;
interpret effect sizes and confidence intervals rather than relying primarily on statistical significance;
use artificial intelligence responsibly and verify AI-generated outputs;
select and apply appropriate reporting guidelines;
conduct systematic reviews and evidence synthesis when appropriate;
write and critically revise a scientific manuscript;

critically evaluate the work of others through structured peer review;
respond scientifically to peer-review comments; and
communicate scientific evidence accurately to clinicians, patients, policymakers, and the public.

These competencies could be documented through real research outputs rather than course attendance alone.

For example, a researcher could demonstrate competence by producing a protocol, statistical analysis plan, data-management plan, reproducible analysis, systematic review, structured peer-review report, conference abstract, or manuscript.

Measure Progress Toward Scientific Independence

The ultimate objective of research mentorship should not be indefinite dependence on a mentor.

It should be progressive scientific independence.

NIH could therefore evaluate whether trainees become increasingly capable of making and defending their own methodological and scientific decisions.

Potential indicators could include:

Verified change in research competencies before and after training.
Ability to develop a research question and protocol with decreasing levels of supervision.
Completion of a scientifically appropriate statistical analysis plan.
Ability to identify methodological weaknesses in one's own research.
Ability to interpret results in terms of magnitude, uncertainty, and scientific or clinical relevance.
Production of reproducible research outputs.
Successful completion and submission of manuscripts or other meaningful scientific outputs.
Ability to respond independently and appropriately to peer review.
Continued research activity after formal training or mentorship ends.
Transition from trainee to independent investigator, collaborator, or mentor.

This would allow NIH to assess not simply whether training was delivered, but whether research capacity was actually created.

Mentorship Should Be Evaluated by What Happens to the Mentee

Mentorship itself should also be evaluated differently.

The number of individuals mentored is an incomplete indicator of mentorship quality. A mentor supervising 30 researchers should not automatically be considered more effective than one who provides intensive mentorship to 5 researchers who subsequently become scientifically independent.

More meaningful mentorship indicators could include:

development of mentee scientific independence;
improvement in methodological competence;
responsible research and research-integrity practices;
successful progression through research milestones;
transition to independent scientific leadership;
retention in biomedical research;
development of interdisciplinary collaboration skills;
equitable access to opportunities, authorship, and scientific networks;
subsequent mentorship of other researchers; and
sustained scientific activity after the formal mentoring relationship ends.

This creates an important longitudinal concept:

The impact of mentorship is not the number of people a scientist has mentored, but the scientific capacity that remains after the mentorship ends.

Recognize the Mentorship Chain

One particularly informative long-term indicator could be the development of a mentorship chain.

If an investigator trains a researcher who subsequently becomes scientifically independent and then successfully mentors another generation of researchers, the original training investment has generated a multiplier effect.

NIH could therefore explore indicators capturing:

Mentor → Trainee → Independent Researcher → Mentor

Such longitudinal measures would be particularly valuable for assessing programs intended to strengthen the biomedical research workforce over many years.

They could also help identify training environments that create sustainable research capacity rather than short-term academic output.

Research Quality Should Matter More Than Publication Counts

Training programs should not be rewarded primarily because their participants produce more publications.

Publication productivity may be one outcome, but it can create unintended incentives if used without measures of quality.

A training program producing fewer but methodologically rigorous, reproducible, transparent, and clinically meaningful studies may have greater scientific impact than one producing many low-value publications.

NIH could therefore examine the quality of research outputs generated during and after training, including:

protocol quality;
adherence to appropriate reporting standards;
reproducibility practices;
appropriate statistical methodology;
transparent reporting of uncertainty;
responsible data and code sharing;
publication of negative or inconclusive findings when scientifically important;
corrections when errors are discovered; and
subsequent reuse or replication of the research.

The objective should be to reward better researchers, not simply more papers.

Training in Responsible Use of Artificial Intelligence

The rapid adoption of generative AI creates an additional workforce competency that should now be considered.

Researchers increasingly use AI for literature exploration, coding, statistical support, scientific writing, evidence synthesis, and methodological assistance. However, the ability to use AI is not equivalent to the ability to use it scientifically and responsibly.

NIH could therefore recognize competencies in:

verification of AI-generated information;
protection of confidential and sensitive research information;
recognition of hallucination and algorithmic bias;
appropriate disclosure of AI use;
preservation of human scientific accountability;
validation of AI-assisted analyses;
provenance of AI-assisted scientific content; and
understanding when AI should not be used.

Responsible AI literacy is increasingly becoming part of research literacy itself.

Recognize Mentorship as Scientific Work

High-quality mentorship requires substantial intellectual effort and time but is often less visible than

publications and grants.

NIH could recognize mentorship explicitly within scientific contribution portfolios and consider mechanisms such as:

protected mentorship time;
supplemental training support;
recognition of documented mentoring contributions in progress reports and renewals;
support for mentor training;
recognition of methodological and interdisciplinary mentorship; and
credit for developing researchers who subsequently achieve scientific independence.

This is particularly important for methodological disciplines such as biostatistics, epidemiology, bioinformatics, data science, research ethics, and research methodology, where a scientist's contribution may substantially improve dozens of other research projects without necessarily being visible through first- or senior-author publications.

Avoid Creating Another Certificate Economy

Any competency-based framework also requires safeguards.

NIH should avoid replacing publication-counting with certificate-counting.

A proliferation of credentials without demonstrated competence would reproduce the same problem in a different form.

Competency assessments should therefore, where possible, be:

performance-based;
independently verifiable;
linked to authentic research activities;
portable across institutions;
proportionate to career stage;
accessible to researchers from institutions with different levels of resources; and
based on open standards rather than dependence on any single commercial provider.
MedicReS Experience

This recommendation reflects an approach MedicReS has pursued since its early development: good research requires development of the researcher across the entire research lifecycle, not isolated instruction in individual techniques.

Our Good Medical Researcher educational approach has therefore emphasized the interconnected competencies required to move from a research question through methodology, study design, biostatistics, research integrity, interpretation, scientific writing, peer review, and communication.

More recently, experience with structured learning environments such as BeGMRAI has reinforced the potential to move beyond attendance-based education toward demonstrated research competencies and outputs.

We offer this experience not as a proposed proprietary NIH standard, but as practical evidence supporting a broader principle:

Training can be designed and evaluated around demonstrated research capability rather than educational participation alone.

Any NIH implementation should use open competency standards that can be adopted, assessed, and independently validated across institutions.

A Practical NIH Pilot

NIH could test this approach in a longitudinal cohort drawn from existing training, career-development, or early-career investigator programs.

At entry, participants could complete a structured assessment of research competencies.

During the program, competencies could be demonstrated through authentic research outputs rather than examinations alone.

At completion and at subsequent intervals, NIH could assess:

Training → Competence → Independence → Scientific Output → Mentorship of Others

A pilot could compare these measures with traditional indicators such as training hours, certificates, publication counts, and grant success to determine which indicators better predict sustained scientific independence and research quality.

MedicReS would be willing to contribute methodological experience to the design of such a pilot and, if useful, make relevant training and research-support capabilities available to a defined NIH pilot cohort for one year without a commercial financial objective, subject to appropriate governance and independent evaluation.

Conclusion

NIH has already invested substantially in training the biomedical research workforce. The next opportunity is to improve how the impact of that investment is measured.

The central shift we recommend is:

Training hours → Demonstrated competence

Certificates → Authentic research outputs

Number of mentees → Quality of mentorship

Publication counts → Quality of research

Dependence → Scientific independence

One generation trained → The next generation capable of training others

Ultimately, the most meaningful indicator of successful training and mentorship may be simple:

Can the researcher independently ask an important question, design a rigorous study, analyze and interpret the evidence responsibly, communicate it transparently, and eventually enable another researcher to do the same?

If the answer is yes, training has produced more than a credential.

It has created sustainable scientific capacity.

4. Collaboration: NIH should assess collaboration not primarily by the number of coauthors, institutions, disciplines, or countries represented in a project, but by the verifiable scientific value that different contributors add to the research.

As biomedical research becomes increasingly interdisciplinary, the conventional publication model does not adequately capture how knowledge is produced. A large collaborative study may depend on investigators who formulate the research question, clinicians who define its relevance, biostatisticians who shape the design and analysis, data scientists who build computational infrastructure, software developers who create analytical tools, methodologists who identify sources of bias, patient representatives who help define meaningful outcomes, and researchers who validate or replicate the findings.

All of these contributions can be scientifically important, yet conventional authorship and bibliometric indicators often make them difficult to distinguish.

We therefore recommend that NIH move from measuring collaboration size toward measuring contribution quality, complementarity, and scientific added value.

1. Make Individual Contributions Visible

At the beginning of an NIH-funded collaborative project, major contributor roles could be defined and subsequently updated as the research evolves.

Relevant contributions may include:

conceptualization;
clinical or scientific domain expertise;
methodology;
study design;
biostatistics;
formal analysis;
software development;
data management and curation;
model development and validation;
research ethics;
patient and community engagement;
visualization;
replication and external validation;
scientific writing;
critical methodological review;
project administration;
development of shared research infrastructure; and
scientific communication and dissemination.

Frameworks such as the CRediT Contributor Roles Taxonomy could provide a useful starting structure, while allowing NIH to recognize additional roles important to contemporary biomedical research.

The objective should be to answer:

Who contributed what, at what stage, and how did that contribution improve the science?

2. Evaluate Scientific Added Value, Not Team Size

Large teams should not automatically be considered more collaborative or more impactful than small teams.

NIH could assess whether collaboration resulted in demonstrable improvements such as:

stronger study design;
better control of bias and confounding;
more appropriate statistical methodology;
improved data quality;
increased population diversity or representativeness;
stronger external validation;
improved reproducibility;
development of reusable research resources;
integration of previously disconnected disciplines;

more clinically meaningful research questions or outcomes; and
increased potential for translation or public benefit.

This would distinguish meaningful interdisciplinarity from nominal interdisciplinarity.

A project involving many disciplines should be rewarded because those disciplines changed or strengthened the research—not simply because their names appear on the application or publication.

3. Recognize Methodological and Statistical Collaboration

NIH should pay particular attention to scientific contributions that are essential but can become invisible under conventional authorship-based assessment.

Biostatisticians, epidemiologists, research methodologists, bioinformaticians, data scientists, software engineers, research-data specialists, and other methodological collaborators may contribute across numerous projects without consistently occupying first- or senior-author positions.

Their impact may be reflected instead in:

preventing inappropriate study designs;
improving sample-size planning;
defining appropriate estimands and outcomes;
developing statistical analysis plans;
identifying bias;
improving reproducibility;
validating analytical methods;
preventing overinterpretation;
developing reusable analytical tools; and
improving the methodological quality of multiple research programs.

These contributions should be explicitly recognized as scientific contributions, not merely as technical support.

This distinction is important for interdisciplinary biomedical science.

4. Recognize Those Who Build the Infrastructure That Makes Collaboration Possible

Some of the most important collaborative contributions do not result immediately in a conventional publication.

Researchers may create:

biobanks;
registries;

shared data platforms;
harmonized protocols;
common data models;
research networks;
multicenter infrastructures;
statistical or computational platforms;
standardized outcome sets; or
shared methodological resources.

Such infrastructure may enable hundreds of subsequent studies.

NIH should therefore recognize the creation, validation, governance, maintenance, and scientific reuse of research infrastructure as collaborative scientific impact.

The contribution of the scientist who creates the system enabling research should not disappear behind the publications subsequently produced through that system.

5. Peer Review and Methodological Review Are Forms of Scientific Collaboration

Scientific collaboration should not necessarily end when a manuscript is written.

Researchers who provide substantive methodological criticism, identify errors, improve protocols, evaluate analytical approaches, or contribute to replication are also contributing to the scientific process.

Under appropriate transparent review models, high-quality scientific reviews could become citable scholarly contributions.

This could include structured methodological assessments of protocols and manuscripts, particularly when the review produces documented improvements in the research.

Recognition should depend on the quality and substance of the review—not merely on the number of manuscripts reviewed.

This would also help address an important imbalance in the current scientific ecosystem: peer review is essential to science but remains largely invisible in conventional measures of scientific contribution.

6. Measure Collaboration Across the Research Lifecycle

Collaboration should be evaluated longitudinally rather than only through the final publication.

NIH could examine contributions at multiple stages:

Research question → Protocol → Study design → Data generation → Analysis → Interpretation → Reporting → Review → Replication → Translation

Different scientists may make their most important contributions at different points in this lifecycle.

A lifecycle-based contribution record would provide a more accurate representation of team science than authorship order alone.

7. Consider Contribution Networks Rather Than Publication Networks Alone

Traditional scientific-network analyses often identify collaboration through coauthorship.

However, coauthorship captures only part of the scientific network.

NIH could explore a broader contribution network in which relationships are based on verified scientific activities such as:

shared protocols;
shared datasets;
methodological contributions;
software development;
validation;
replication;
mentorship;
infrastructure development;
structured scientific review; and
translational implementation.

Such a model could reveal researchers whose influence on biomedical science is substantial but poorly captured by publication counts or authorship position.

8. Recognize Sustainable Collaboration

One important indicator of successful collaboration is whether scientific relationships and infrastructure continue after the original grant ends.

Potential longitudinal indicators could include:

continuation of interdisciplinary research partnerships;
new research questions generated by the collaboration;
subsequent joint projects;
continued reuse of shared infrastructure;
expansion to new institutions or populations;
development of independent investigators within the network;
successful replication or external validation; and
continued sharing of research resources.

The objective should not be to reward networking itself.

It should be to recognize collaboration that creates durable scientific capacity.

9. Patient and Community Contributions Should Be Visible

Contemporary biomedical research increasingly involves patients, caregivers, community representatives, and other stakeholders.

Their contribution should not be treated as symbolic participation.

Where appropriate, NIH could document whether these collaborators meaningfully influenced:

selection of research questions;
outcome selection;
study feasibility;
recruitment strategies;
interpretation of results;
dissemination;
implementation; or
identification of unintended consequences.

Meaningful stakeholder involvement should be distinguishable from nominal inclusion.

10. Safeguards Against Gaming

Any contribution-based system creates potential unintended consequences.

NIH should guard against:

honorary or gift authorship;
inflated contribution claims;
excessive fragmentation of contributor roles;
reciprocal credit arrangements;
artificial expansion of research teams;
dominance of prestigious institutions;
disadvantage to investigators from smaller or less-resourced institutions;
undervaluing individual or small-team science; and
converting contribution metrics into another single ranking system.

Contribution statements should therefore be verifiable where feasible, and NIH should avoid creating a single numerical “collaboration score.”

Instead, collaboration should form one dimension of a broader scientific contribution profile.

11. Potential Indicators

NIH could consider indicators such as:

Clearly defined and verified contributor roles.
Demonstrable influence of interdisciplinary expertise on study design or analysis.
Jointly developed protocols, datasets, methods, software, or models.
Creation and sustained use of shared research infrastructure.
Meaningful methodological and statistical contributions.
Contributions to independent validation and replication.
Substantive peer-review or methodological-review contributions.
Meaningful patient or community participation.
Transparent management of conflicts of interest and contribution changes.
Continued collaboration or infrastructure use after project completion.
Reuse of collaborative outputs by researchers outside the original team.
Development of new investigators or research capacity through collaboration.
Conclusion

The central principle is simple:

Collaboration should be measured by the scientific value created between people, not merely by the number of people connected to a project.

A publication with 50 authors is not necessarily evidence of better collaboration than one with five. Likewise, a scientist who appears infrequently as first or senior author may nevertheless have transformed the quality of dozens of studies through methodology, biostatistics, software, data infrastructure, validation, or scientific review.

NIH has an opportunity to make these previously less-visible contributions visible.

The progression we recommend is:

Number of collaborators → Verified contributions

Authorship position → Scientific role

Institution count → Complementary expertise

Coauthorship network → Contribution network

Project collaboration → Sustainable scientific capacity

Ultimately, team science should be rewarded when collaboration enables biomedical research to become more rigorous, more reproducible, more innovative, more inclusive, and more useful than any individual contributor could have achieved alone.

5. Entrepreneurship and Translation: NIH should distinguish between entrepreneurial activity and translational impact.

Patents filed, licenses executed, companies formed, venture capital raised, or products launched are useful indicators of activity, but they do not necessarily demonstrate that a biomedical discovery has generated meaningful scientific, clinical, or public value.

For the NIH mission, the more important question is:

Did the discovery move from scientific promise to independently validated, safe, effective, accessible, and useful application for people?

We therefore recommend that NIH assess entrepreneurship and translation as a stage-specific pathway from discovery to real-world value, rather than as a count of commercial outputs.

1. Measure the Translational Pathway, Not Only Its Commercial Milestones

A biomedical discovery may progress through several stages:

Discovery → Validation → Prototype → Independent Validation → Regulatory Readiness → Clinical Evaluation → Implementation → Real-World Evidence → Patient/Public Benefit

Different innovations will appropriately reach different points along this pathway.

NIH should therefore recognize progress according to the scientific and developmental stage of the innovation rather than expecting every funded discovery to rapidly generate a patent, company, licensed product, or commercial return.

Potential indicators could include:

successful technical validation;
independent scientific validation;
external clinical validation;
progression from proof-of-concept to prototype;
demonstration of safety and performance;
regulatory readiness or regulatory milestones;
integration into clinical or health-system workflows;
evidence of usability by clinicians, patients, or other intended users;
real-world adoption;
generation of pragmatic or real-world evidence;
contribution to clinical guidelines or health policy;

accessibility and affordability; and
measurable benefit to patients or populations.

The central shift should be:

Commercialization metrics → Translational evidence

2. Independent Validation Should Be a Major Indicator

One of the most important transitions in biomedical innovation occurs when a discovery moves beyond the laboratory, institution, dataset, or team that originally developed it.

NIH should therefore recognize independent validation as a major translational milestone.

For diagnostics, predictive models, digital health tools, AI systems, biomarkers, devices, therapeutic approaches, or research technologies, validation could include evaluation:

by an independent research group;
in an external dataset;
in another clinical institution;
in a different patient population;
under routine clinical conditions; or
through prospective evaluation where appropriate.

An innovation that performs exceptionally in its development environment but fails when independently evaluated should not be considered successfully translated.

Conversely, independent confirmation—even before commercialization—represents substantial translational progress.

3. Distinguish Efficacy From Effectiveness

A particularly important translational question is whether an intervention that works under controlled research conditions also works in routine practice.

NIH could therefore explicitly recognize the transition from:

Efficacy → Effectiveness

as a measurable component of scientific impact.

Pragmatic trials, implementation studies, comparative-effectiveness research, registries, and appropriately designed real-world evidence studies can help determine whether benefits observed in controlled settings persist across:

routine clinical practice;
broader patient populations;
different institutions;
different clinicians;
different socioeconomic settings; and
populations historically underrepresented in research.

The efficacy–effectiveness gap itself can be scientifically informative and should be considered part of translational evaluation.

4. Adoption Is More Meaningful Than Availability

A product becoming technically available does not mean that it has achieved meaningful translation.

NIH should distinguish:

Product exists → Product is available → Product is adopted → Product improves outcomes

Potential indicators of meaningful adoption could include:

integration into routine clinical workflows;
sustained use rather than temporary pilot use;
clinician acceptance;
patient acceptance;
implementation across multiple institutions;
incorporation into guidelines or care pathways;
measurable improvement in clinical processes;
measurable improvement in patient-important outcomes; and
evidence that the innovation reaches its intended population.

This would prevent the existence of a commercial product from being treated automatically as evidence of public benefit.

5. Patient Value Should Remain the Translational Endpoint

For biomedical research, the ultimate translational objective should remain benefit to people.

Where appropriate, NIH could therefore assess:

survival;
morbidity;
symptoms;
functional status;

quality of life;
validated patient-reported outcomes;
diagnostic accuracy;
prevention;
safety;
treatment burden;
access to care;
time to diagnosis or treatment;
health-system efficiency; and
patient experience.

Surrogate or intermediate outcomes may be appropriate at earlier translational stages, but the evaluation framework should distinguish them from demonstrated patient benefit.

6. Accessibility and Health Equity Are Components of Translation

An innovation that works but cannot reach the population that needs it has achieved only partial translation.

NIH should therefore consider:

affordability;
geographic accessibility;
feasibility in lower-resource settings;
usability across different health systems;
representation of relevant populations during validation;
accessibility for underserved populations;
digital-access requirements;
language and cultural barriers; and
potential to reduce or exacerbate existing health disparities.

This is especially important for AI and digital health technologies, where innovation may inadvertently widen disparities if development and validation are concentrated in populations and institutions with greater resources.

Thus:

Translation should include not only whether an innovation works, but also for whom, where, and under what conditions it works.

7. Cost and Value Should Be Considered

A scientifically effective innovation may still have limited translational value if its cost makes implementation unrealistic.

Depending on the stage and type of innovation, NIH could recognize evidence concerning:

cost-effectiveness;
budget impact;
resource utilization;
implementation costs;
scalability;
sustainability; and
opportunity costs.

Economic evaluation should not replace clinical value, but it can help determine whether an innovation can realistically move from research into sustainable practice.

8. Failed Translation Can Still Produce Scientific Value

Not every scientifically promising discovery will become a successful product.

Failure is inherent to innovation.

NIH should avoid creating incentives that encourage researchers or entrepreneurs to hide unsuccessful translational efforts.

A well-designed translational project that demonstrates that an intervention:

does not replicate;
fails external validation;
produces unacceptable adverse effects;
lacks effectiveness in routine practice;
is not cost-effective;
cannot be implemented feasibly; or
performs differently across populations

may generate extremely valuable scientific information.

Such findings can prevent further research waste, unnecessary investment, ineffective clinical implementation, or patient harm.

NIH should therefore recognize transparent, methodologically rigorous translational failure as scientific learning.

9. Entrepreneurship Should Not Move Faster Than the Evidence

Commercial incentives can create pressure to move rapidly from scientific discovery to claims of clinical

benefit.

NIH should maintain a clear distinction between:

commercial success and scientific validation.

Appropriate safeguards should include:

transparent conflict-of-interest disclosure;
independent validation;
clear distinction between exploratory and validated claims;
evidence appropriate to the level of clinical claim;
transparent reporting of negative results;
monitoring of safety and performance after implementation; and
correction of claims when subsequent evidence changes the conclusion.

Commercial success should never substitute for scientific validity.

10. Recognize Different Forms of Innovation

Entrepreneurship should not be interpreted narrowly as company formation.

Biomedical innovation can include:

new therapeutics;
diagnostics;
medical devices;
biomarkers;
AI and predictive models;
digital health technologies;
statistical and methodological tools;
research software;
clinical decision-support systems;
data infrastructures;
new models of care;
public-health interventions;
implementation strategies; and
educational or research-capacity technologies.

Some highly impactful innovations may be open-source, nonprofit, academically maintained, or integrated directly into public infrastructure.

NIH should therefore reward scientific and public value regardless of the business model through which it is delivered.

11. Recognize the People Who Enable Translation

Translation is inherently multidisciplinary.

The pathway from discovery to implementation may involve basic scientists, clinicians, biostatisticians, epidemiologists, engineers, data scientists, regulatory experts, implementation scientists, health economists, patient representatives, entrepreneurs, and health-system leaders.

NIH should recognize these contributions individually rather than attributing translational success only to the inventor, principal investigator, patent holder, or company founder.

Structured contributor frameworks could document who contributed to:

discovery;
validation;
methodology;
product development;
regulatory strategy;
clinical evaluation;
implementation;
economic evaluation;
real-world evidence; and
patient/community engagement.
12. Potential Indicators

NIH could consider a multidimensional translational profile including:

Independent technical validation.
Independent clinical validation.
External validation across populations and settings.
Progression from prototype to real-world use.
Regulatory readiness and appropriate regulatory milestones.
Safety and performance evidence.
Integration into clinical or health-system workflows.
Real-world adoption and sustained use.
Pragmatic research and real-world evidence.
Evidence addressing the efficacy–effectiveness gap.
Patient-important outcomes.
Guideline or policy influence.
Accessibility and affordability.
Health-equity implications.
Cost-effectiveness or implementation value where appropriate.
Transparent reporting of unsuccessful translation.

Sustained benefit after implementation.

These indicators should be interpreted according to the type and maturity of the innovation rather than combined into a universal numerical entrepreneurship score.

13. A Stage-Appropriate NIH Framework

NIH could evaluate translational projects according to their position on the research-to-impact pathway.

For example:

Early stage:
Scientific rationale, reproducibility, technical validation.

Development stage:
Prototype performance, independent validation, safety, usability.

Clinical translation:
Clinical validity, clinical utility, regulatory readiness.

Implementation:
Workflow integration, adoption, effectiveness, accessibility.

Real-world impact:
Patient outcomes, population impact, health equity, sustainability, and value.

This would reduce pressure for premature commercialization while allowing NIH to recognize genuine translational progress before a product reaches the market.

Conclusion

NIH has an opportunity to redefine entrepreneurship in biomedical research around evidence rather than commercial activity alone.

The progression we recommend is:

Patent → Independent validation

Prototype → Demonstrated utility

Company formation → Sustainable implementation

Investment → Evidence generation

Efficacy → Effectiveness

Product launch → Real-world adoption

Commercial success → Patient and public value

The most important question should therefore not be:

“Did this research generate a patent, company, or product?”

but:

“Did this discovery survive independent validation, move responsibly toward real-world use, reach the people who need it, and create demonstrable scientific, clinical, or public value?”

That is the translational impact most closely aligned with the NIH mission.

6. Foundational Scientific Exploration: Foundational science and high-risk, high-reward research require an impact framework fundamentally different from that used for mature translational research.

By definition, exploratory research operates under substantial uncertainty. Some projects will not confirm their hypotheses, generate immediately publishable positive findings, produce patents, or lead rapidly to clinical applications. This should not automatically be interpreted as scientific failure.

The central question should instead be:

Did the research reduce important uncertainty, rigorously test a consequential idea, create new knowledge or capability, or open a scientifically valuable path for others to pursue?

NIH should therefore protect foundational research from evaluation systems that unintentionally reward short-term productivity, predictable positive results, or rapid translation at the expense of scientific exploration.

1. Distinguish Scientific Failure From Research Failure

A negative result is not necessarily a failed research project.

There is an important distinction between:

A hypothesis that fails after rigorous testing

and

Research that fails because it was poorly designed, inadequately conducted, or insufficiently documented.

The first can be scientifically valuable. The second should not be rewarded merely because it is described as high-risk research.

For foundational exploration, NIH should therefore evaluate whether:

the scientific question was important;
the hypothesis or proposition was meaningfully testable;
the study was appropriately designed;
methods were sufficiently rigorous;
uncertainty was honestly characterized;
alternative explanations were considered;
negative or unexpected findings were transparently reported; and
the resulting knowledge can inform subsequent research.

A rigorously demonstrated negative result can eliminate an incorrect hypothesis, prevent duplication of unproductive work, redirect resources, or reveal previously unrecognized biological complexity.

That is scientific progress.

2. Reward Reduction of Scientific Uncertainty

Foundational research often creates value by changing what the scientific community knows not to be true, narrowing the range of plausible explanations, or identifying what must be investigated next.

NIH could therefore recognize reduction of important scientific uncertainty as a legitimate form of impact.

Potential evidence could include:

rejection or refinement of an important hypothesis;
identification of previously unknown mechanisms;
discovery of unexpected relationships;
clarification of conflicting evidence;
identification of boundary conditions;
development of new conceptual models;
generation of new research questions; and
redirection of a scientific field toward more promising hypotheses.

This would allow NIH to recognize knowledge generation without requiring every project to produce a positive or immediately actionable result.

3. Use Longer Time Horizons

Foundational discoveries frequently require years—or decades—before their importance becomes apparent.

A scientific observation that initially has little clinical relevance may later become essential to a therapeutic technology, diagnostic method, computational approach, or entirely new field.

NIH should therefore avoid assessing foundational research primarily through short-term indicators such as:

publications within the grant period;
immediate citation counts;
journal prestige;
patents;
company formation;
licensing activity; or
rapid clinical translation.

These measures may become relevant later, but they are often poor indicators of the value of early scientific exploration.

NIH could instead use longitudinal and portfolio-based evaluation, allowing the influence of foundational discoveries to emerge over appropriate time periods.

4. Measure What Foundational Research Creates

Foundational research can produce many forms of scientific value beyond conventional publications.

Potential indicators could include:

Creation of a new hypothesis or conceptual framework.
Development of a new experimental or analytical method.
Creation of a new measurement instrument or assay.
Generation of a valuable dataset or biological resource.
Development of a model system.
Creation of software or computational methods.
Establishment of shared research infrastructure.
Identification of previously unknown mechanisms.
Resolution of important scientific uncertainty.
Generation of new questions subsequently pursued by independent researchers.
Independent testing or replication of the finding.
Adoption of methods or resources by other laboratories.
Emergence of new interdisciplinary research directions.
Long-term contribution to translational or clinical research.

The emphasis should be on what scientific capability or knowledge became possible because the research was conducted.

5. Protect Negative and Null Results

The current research ecosystem can unintentionally reward positive findings and discourage publication of negative, null, or inconclusive results.

This creates scientific waste and may distort the evidence base.

NIH could explicitly recognize high-quality negative findings when:

the research question was scientifically important;
the study had adequate methodological rigor;
the analysis was appropriate;
uncertainty was reported transparently;
the result meaningfully informs the hypothesis; and
the findings and methods are accessible to other researchers.

This recognition could reduce unnecessary duplication and make failed scientific paths informative to the broader research community.

However, NIH should not create incentives for low-quality negative studies. Methodological rigor must remain the prerequisite for scientific credit.

6. Distinguish Exploratory From Confirmatory Research

Foundational science often requires exploratory analyses.

Exploration should not be discouraged. It should be identified honestly.

NIH could encourage researchers to distinguish clearly between:

Exploratory findings: hypothesis-generating observations requiring further testing.

and

Confirmatory findings: evidence generated through appropriately prespecified testing.

For high-dimensional, omics, imaging, AI, or other data-intensive research, NIH could encourage transparent documentation of:

which analyses were prespecified;
which analyses emerged during exploration;

multiple-comparison procedures where appropriate;
internal validation;
external validation;
sensitivity analyses; and
plans for subsequent confirmatory investigation.

The problem is not exploration.

The problem is presenting exploration as confirmation.

7. Recognize Scientific Risk Properly

NIH should also distinguish genuine scientific risk from poor planning.

A high-risk project should be high-risk because:

the hypothesis is bold;
the scientific uncertainty is substantial;
the mechanism is poorly understood;
the methodology is novel;
existing evidence is limited;
or the potential discovery challenges established assumptions.

It should not be considered high-risk simply because the protocol is underdeveloped or the methods are weak.

A useful principle would therefore be:

High scientific uncertainty should be encouraged; low methodological rigor should not.

This distinction is essential if NIH wants to reward genuinely transformative research.

8. Evaluate Learning Milestones, Not Only Success Milestones

For high-risk projects, investigators could define learning milestones at the beginning of the project.

Instead of asking only:

“Did the intervention work?”

or

“Was the original hypothesis confirmed?”

NIH could also ask:

What uncertainty was resolved?
Which assumptions were supported or rejected?
What unexpected findings emerged?
Which methodological barriers were identified?
What new hypotheses resulted?
What should the next experiment test?
Should this research direction continue, change, or stop?

This would make rigorous learning visible even when the original hypothesis is not confirmed.

9. Reward the Decision to Stop When Evidence Justifies It

Scientific culture strongly rewards continuation: another experiment, another grant, another publication.

But sometimes the scientifically responsible decision is to stop.

If rigorous evidence demonstrates that a research direction is unlikely to succeed, documenting that conclusion may prevent substantial duplication of effort and public expenditure.

NIH could therefore recognize a well-supported “stop decision” as a legitimate scientific outcome when it is based on transparent evidence.

Stopping an unproductive research direction can itself represent responsible stewardship of research resources.

10. Evaluate Research Programs as Portfolios

High-risk/high-reward science is particularly poorly suited to evaluation one project at a time.

If every individual project is expected to succeed, investigators will rationally reduce risk.

NIH could instead consider portfolio-level assessment.

Within a portfolio of well-designed exploratory projects:

some may confirm hypotheses;
some may refute them;
some may create tools;
some may generate unexpected findings;
some may fail technically;
and a small number may transform a field.

The appropriate question is therefore not whether every project succeeded.

It is whether the portfolio created enough new knowledge, capability, and opportunity to justify the scientific risk undertaken.

This is particularly important for genuinely high-risk, high-reward funding programs.

11. Protect Early-Career Investigators

A system that encourages risk but penalizes unsuccessful outcomes creates a contradiction.

This is especially consequential for early-career investigators, who may reasonably avoid ambitious research if a negative result threatens their next grant, publication record, or academic advancement.

NIH could consider mechanisms that recognize rigorous high-risk research independently of whether the initial hypothesis was confirmed.

Evaluation should emphasize:

quality of reasoning;
methodological rigor;
transparency;
learning generated;
resources created;
and subsequent scientific directions.

This would allow investigators to take scientifically justified risks without being rewarded for recklessness or punished for honest negative results.

12. Avoid a Single Foundational-Impact Score

The diversity of foundational science makes a single numerical score inappropriate.

A theoretical contribution, new biological mechanism, assay, dataset, computational method, animal model, or unexpected negative finding creates value in different ways.

NIH should therefore use a multidimensional profile rather than attempting to collapse foundational impact into one ranking.

Relevant dimensions could include:

Novelty

Rigor

Uncertainty reduced

Knowledge generated

Methods/resources created

Independent reuse

New research enabled

Long-term influence

This profile could complement rather than replace expert scientific judgment.

13. Follow the Scientific Genealogy of Discoveries

For long-term evaluation, NIH could also examine the scientific genealogy of foundational discoveries.

Rather than relying solely on citations, NIH could ask whether an early discovery subsequently:

generated new hypotheses;
enabled new methods;
was incorporated into independent research programs;
contributed to new therapeutic or diagnostic approaches;
supported a new field;
enabled interdisciplinary research;
or became part of a later translational pathway.

This would help make visible something traditional metrics often miss:

The scientific impact of foundational research may appear far downstream from the original investigator and many years after the original grant.

14. Potential Indicators

NIH could consider:

Importance and originality of the scientific question.
Testability of the central hypothesis or proposition.
Methodological rigor.
Transparent characterization of uncertainty.
Transparent reporting of negative and unexpected findings.
Creation of new hypotheses or conceptual frameworks.

Creation of methods, tools, datasets, models, or infrastructure.
Independent reuse of these resources.
Independent testing, validation, or replication.
Generation of subsequent research questions.
Emergence of interdisciplinary research directions.
Evidence that negative findings prevented unnecessary duplication.
Progression of exploratory findings to confirmatory studies.
Long-term scientific or translational influence.

These indicators should be interpreted according to scientific context and over an appropriate time horizon.

Conclusion

NIH's support for foundational exploration requires an assessment system that accepts a fundamental property of science:

A scientifically valuable experiment does not have to confirm the hypothesis.

The progression we recommend is:

Positive results → Informative results

Short-term productivity → Long-term scientific value

Publication count → Knowledge created

Citation count → Research enabled

Failure avoidance → Responsible scientific risk

Unexpected result → New scientific direction

Individual project success → Portfolio learning

The most important distinction is therefore not between success and failure.

It is between:

Research that rigorously teaches us something important and research that does not.

A bold hypothesis that is carefully tested and ultimately rejected may advance science further than a predictable hypothesis that produces another positive publication.

If NIH wants to sustain genuinely transformative research, researchers must be able to pursue important uncertainty without fearing that an honest negative answer will be interpreted as scientific failure.

7. Public Impact: NIH should distinguish between public visibility and public impact.

Media coverage, website visits, social-media engagement, news mentions, or public attention can demonstrate dissemination, but they do not necessarily demonstrate that biomedical research has improved health, clinical practice, public policy, accessibility, public trust, or societal well-being.

The more meaningful question is:

What verifiable difference did the research make to people, health systems, communities, policy, or society—and what evidence supports that contribution?

We recommend that NIH assess public impact through a multidimensional and longitudinal framework that recognizes different pathways from scientific discovery to societal benefit.

1. Measure Contribution to Health Outcomes

For biomedical research, one of the most important forms of public impact is improvement in health.

Where appropriate, NIH could recognize evidence of contribution to:

reduced mortality;
reduced morbidity;
improved prevention;
earlier or more accurate diagnosis;
improved treatment effectiveness;
improved treatment safety;
improved functional status;
improved quality of life;
reduced disease burden;
improved access to effective care; and
reduction of avoidable health disparities.

Importantly, impact should not be restricted to statistically significant changes.

The magnitude of benefit, confidence intervals, clinical importance, population reach, durability, and potential harms should also be considered.

This would align public-impact assessment with the fundamental principle that statistical significance is not equivalent to clinical or societal significance.

2. Recognize Patient-Important Outcomes

Research impact should be evaluated partly through outcomes that matter directly to patients.

Depending on the research question, these may include:

survival;
symptoms;
functional ability;
quality of life;
treatment burden;
adverse effects;
ability to work or participate in daily life;
patient experience; and
validated patient-reported outcome measures.

Intermediate biomarkers and surrogate outcomes may be scientifically important, particularly during earlier stages of research, but NIH should distinguish between evidence of biological effect and demonstrated benefit meaningful to patients.

3. Measure Influence on Clinical Practice

A research finding may have substantial public impact when it changes how health care is delivered.

Potential indicators could include:

incorporation into clinical practice guidelines;
changes in standards of care;
adoption of diagnostic or treatment approaches;
discontinuation of ineffective or harmful practices;
incorporation into clinical decision-support systems;
changes in screening or prevention strategies;
adoption across health systems; and
sustained implementation in routine clinical practice.

However, guideline citation alone should not automatically constitute high public impact.

NIH could distinguish:

Evidence cited → Recommendation changed → Practice changed → Patient outcomes changed

These represent progressively stronger evidence of translation into public benefit.

4. Recognize Contributions to Public Health and Policy

Biomedical research can also create impact beyond individual clinical encounters.

NIH could recognize contributions to:

public-health recommendations;
vaccination strategies;
screening programs;
disease-surveillance systems;
prevention policies;
environmental or occupational-health policies;
health-system planning;
emergency preparedness;
regulatory decisions; and
allocation of health resources.

Where feasible, evaluation should document the pathway through which research informed a policy or program rather than assuming that citation in a policy document demonstrates causation.

5. Evaluate Accessibility and Reach

An effective intervention cannot generate broad public benefit if people cannot access it.

NIH should therefore consider:

affordability;
geographic accessibility;
availability across different health systems;
implementation in underserved communities;
language and cultural accessibility;
digital accessibility;
barriers affecting people with disabilities;
scalability;
population reach; and
sustainability.

A scientific innovation that benefits only a narrow, highly resourced population may have important scientific value, but its broader public impact should be distinguished from an intervention that reaches populations with substantial unmet need.

6. Include Health Equity and Representativeness

Public impact assessment should ask:

Who benefited—and who did not?

NIH could evaluate whether study populations appropriately represent the populations expected to use or benefit from the resulting evidence.

Where scientifically appropriate, this could include assessment by:

age;
sex;
race and ethnicity;
socioeconomic circumstances;
geographic setting;
relevant comorbidities; and
other clinically meaningful population characteristics.

Stratified analyses and interaction assessments should be used when justified by the research question and design, rather than applied mechanically to every study.

Researchers should also communicate clearly the populations to which their conclusions can and cannot reasonably be generalized.

This is particularly important for AI, precision medicine, genomic research, and predictive models, where performance may differ substantially across populations.

7. Public Trust Should Be Considered a Scientific Asset

Public trust is essential to biomedical research.

Trust should not be measured merely through favorable public opinion. It is better supported through observable research practices.

Potential indicators could include:

transparent protocols;
appropriate public access to research findings;
transparent disclosure of conflicts of interest;
understandable communication of benefits and risks;
appropriate communication of uncertainty;
accessible explanations of research findings;
transparent corrections;
acknowledgment of errors;
publication of negative or conflicting evidence;
disclosure of important protocol changes;
responsible communication of AI use; and
availability of replication or validation evidence.

Trust is strengthened not by presenting science as infallible, but by demonstrating that science has mechanisms to identify uncertainty, detect errors, correct itself, and update conclusions when evidence changes.

8. Recognize Meaningful Patient and Community Engagement

Patients, caregivers, and communities should not be viewed only as recipients of scientific impact.

They can contribute to defining it.

Where appropriate, NIH could recognize meaningful involvement in:

identifying research priorities;
defining relevant research questions;
selecting patient-important outcomes;
improving recruitment and retention;
assessing study feasibility;
interpreting results;
designing dissemination strategies; and
supporting implementation.

The key word is meaningful.

Participation should be assessed according to whether stakeholder input actually influenced research decisions, rather than simply whether a patient or community representative attended meetings.

9. Economic Impact Should Be Broader Than Commercial Revenue

Biomedical research can generate economic value without necessarily creating a commercial product.

NIH could recognize contributions to:

reduced hospitalization;
prevention of avoidable disease;
reduced treatment burden;
improved health-system efficiency;
improved diagnostic efficiency;
reduced use of ineffective interventions;
improved workforce participation;
reduced caregiver burden;
development of research infrastructure;
creation of skilled scientific employment; and
development of technologies or industries based on publicly funded science.

Where appropriate, cost-effectiveness, budget impact, and resource-utilization analyses can provide useful evidence.

However, economic return should remain one dimension of public impact, not its dominant definition.

10. Recognize the Public Value of Preventing Harm and Research Waste

Research can create societal benefit not only by introducing something new but also by showing that something should not be done.

A rigorous study demonstrating that an intervention is ineffective, harmful, unnecessarily expensive, or unsupported by evidence may create substantial public value.

Similarly, replication or reproducibility work that prevents widespread reliance on an unreliable finding may save research resources and protect patients.

NIH should therefore recognize:

Evidence supporting adoption

and

Evidence supporting de-adoption

as legitimate forms of public impact.

This is especially important because current incentive structures often make positive innovation more visible than the scientific work that prevents ineffective or harmful practices.

11. Distinguish Contribution From Attribution

A major methodological challenge in measuring public impact is attribution.

Changes in population health, clinical practice, or policy rarely result from a single paper, grant, laboratory, or investigator.

They often emerge through cumulative evidence produced by many researchers over years.

NIH should therefore be cautious about asking:

“Did this researcher cause this societal outcome?”

A more scientifically defensible question is:

“What documented contribution did this research make to the pathway that produced the outcome?”

Evidence could include:

Research finding → Evidence synthesis → Guideline → Implementation → Practice change → Health outcome

An individual project may contribute importantly at one or more points along this pathway without being solely responsible for the final outcome.

This contribution-based approach would reduce exaggerated impact claims while still recognizing meaningful scientific influence.

12. Public Impact Requires Time

Some research influences practice within months.

Other discoveries require decades.

NIH should therefore avoid imposing a single time horizon for public impact.

A multidimensional scientific-impact record could distinguish:

Immediate impact — dissemination, public access, stakeholder engagement.

Intermediate impact — replication, evidence synthesis, guideline consideration, implementation research.

Long-term impact — clinical adoption, population-health effects, policy influence, economic value, or establishment of new standards of care.

This would prevent foundational research from being disadvantaged simply because its societal impact has not yet matured.

13. Potential Indicators

NIH could consider indicators including:

Contribution to clinical guidelines or standards of care.
Use in public-health programs or policy.
Adoption of diagnostic, therapeutic, preventive, or care approaches.
Demonstrable contribution to patient-important outcomes.
Use of validated patient-reported outcomes where appropriate.
Contribution to accessibility and health equity.

Meaningful patient, caregiver, or community involvement.
Population reach and representativeness.
Transparent communication of uncertainty.
Appropriate public access to findings.
Transparent correction of scientific errors.
Replication or validation supporting public use.
Evidence supporting de-adoption of ineffective or harmful practices.
Economic or health-system value.
Reuse of publicly funded research resources.
Documented contribution to a broader pathway from evidence to societal outcome.

These indicators should not be collapsed into a universal “public impact score.”

Different forms of biomedical research produce public value through different pathways and over different periods.

14. Avoid Rewarding Visibility Instead of Value

Any public-impact framework should include safeguards against metric gaming.

NIH should avoid creating incentives for researchers to:

exaggerate the implications of preliminary findings;
equate media coverage with scientific importance;
pursue social-media attention;
overstate causality;
overstate clinical relevance;
selectively communicate positive findings;
minimize uncertainty; or
choose highly visible research questions over scientifically important but less publicly attractive ones.

Altmetrics and media indicators may document dissemination, but they should remain contextual indicators rather than primary evidence of societal impact.

15. A Public Impact Pathway

NIH could organize public impact around a transparent pathway:

Discovery

↓

Validation

↓

Evidence Synthesis

↓

Guideline / Policy Consideration

↓

Implementation

↓

Adoption

↓

Patient / Population Outcome

↓

Societal Value

At each stage, researchers could document their contribution and the evidence supporting it.

This would create a far more defensible measure of public impact than attempting to infer societal benefit directly from publication or citation metrics.

MedicReS Perspective

For nearly two decades, MedicReS has approached good medical research as a lifecycle extending from the research question through methodology, biostatistics, interpretation, scientific communication, and ultimately the responsible use of evidence.

This experience has reinforced a central principle relevant to the NIH RFI:

Research is not valuable merely because it is published. Its ultimate value depends on whether trustworthy evidence can contribute to better decisions.

Transparency is essential to this process. Protocols, methodological evaluation, author responses, corrections, replication findings, and appropriately accessible research outputs can strengthen not only reproducibility but also public confidence in how scientific knowledge is produced and corrected.

Our broader recommendation is therefore that public impact be connected to the same lifecycle-based framework proposed across the other NIH domains.

Conclusion

NIH has an opportunity to move public-impact assessment beyond visibility metrics toward verifiable societal contribution.

The progression we recommend is:

Media attention → Evidence of use

Citation → Contribution to decisions

Statistical significance → Clinically meaningful benefit

Guideline citation → Practice change

Availability → Accessibility

Innovation → Population benefit

Publicity → Public trust

Claimed impact → Documented contribution

And perhaps most importantly:

“Who saw the research?” → “What changed because trustworthy evidence became available?”

The goal should not be to reward scientists for making the largest claims about their impact.

It should be to recognize research that can demonstrate a credible pathway from scientific evidence to better health, better decisions, greater accessibility, stronger public trust, and meaningful societal benefit.

8. Other Areas: Building Better Researchers Before Measuring Their Impact

In addition to the areas identified in this RFI, we recommend that NIH consider a fundamental upstream question:

Before measuring the scientific impact of researchers, are we systematically ensuring that researchers have the competencies required to produce rigorous, reproducible, ethical, and impactful research?

Many problems that later emerge as failures of reproducibility, inappropriate study design or statistical

analysis, selective reporting, inefficient use of research resources, or inappropriate interpretation originate much earlier in the research lifecycle.

NIH already invests substantially in research projects, research infrastructure, training, and protected time. We suggest complementing these investments with a structured researcher competency pathway that can be evaluated through demonstrated research performance rather than course attendance or certificates alone.

A Good Medical Researcher Pilot

MedicReS has supported the development of the Good Medical Researcher (GMR) Online Training Program, developed under the editorship of Prof. E. Arzu Kanık.

The program can be completed online and is designed around the complete biomedical research lifecycle, integrating competencies in research ethics and integrity, research question formulation, study design, protocol development, population and sampling, sample-size planning, biostatistical reasoning, interpretation of effect sizes and confidence intervals, evidence synthesis, scientific reporting, peer review, and responsible use of artificial intelligence.

We propose that NIH consider testing this existing educational approach in a defined cohort of NIH-supported investigators.

MedicReS would be willing to provide NIH-selected participants with free access to the online Good Medical Researcher training program for a one-year pilot period, without a commercial financial objective.

The purpose would not be to promote or evaluate a training product. It would be to test a scientific hypothesis:

Does structured, lifecycle-based researcher education measurably improve the rigor, reproducibility, methodological competence, and scientific independence of NIH-supported investigators?

Baseline and follow-up indicators could include research-methodology competence, protocol quality, appropriateness of study design, sample-size justification, statistical analysis planning, interpretation of effect sizes and uncertainty, recognition of bias and confounding, reproducibility practices, research integrity, responsible AI use, reporting quality, manuscript quality, and progression toward independent research capability.

Where feasible, outcomes could be compared with an appropriate comparison group or baseline, and the evaluation methodology could be defined prospectively with NIH.

Measure the Researcher, Not the Certificate

Completion of the program itself should not be considered evidence of impact.

The relevant outcome is whether the researcher subsequently demonstrates improved scientific competence and produces better research.

In other words:

Course completion → Activity

Certificate → Credential

Demonstrated research competence → Outcome

Improved research quality → Impact

This distinction is essential. We do not propose replacing publication counting with certificate counting.

From Pilot to Evidence-Based Policy

If independent evaluation demonstrates meaningful improvements in research quality, reproducibility, or researcher independence, NIH could then consider whether demonstrated lifecycle-based research competency should become a recommended component of researcher development for appropriate categories of NIH-supported investigators.

Any future NIH framework should remain open and provider-neutral, allowing equivalent programs developed by universities, scientific organizations, or other qualified providers to participate if they satisfy transparent and independently assessable competency standards.

MedicReS offers its program as a pilot environment in which the concept can be tested, not as a proprietary requirement.

Why This May Have Long-Term Impact

The return on researcher development may extend far beyond an individual grant.

A research project has a defined duration. A scientifically well-trained independent investigator may conduct rigorous research for decades, develop research teams, mentor future investigators, build infrastructure, and transmit good scientific practices to subsequent generations.

The potential pathway is:

NIH investment

→ Researcher competency

→ Better-designed research

→ More rigorous and reproducible evidence

→ Greater scientific impact

→ Independent research leadership

→ Future mentorship

→ Sustainable scientific capacity

This also connects the seven areas described in this RFI. Rigor, reproducibility, responsible data sharing, collaboration, translation, foundational exploration, and public impact ultimately depend, at least in part, on the competencies of the researchers responsible for producing and interpreting the science.

Our Proposal

Rather than assuming that researcher education improves research quality, we propose testing it as a scientific intervention.

NIH could identify the cohort. MedicReS could provide the existing online Good Medical Researcher educational environment without charge during the pilot. Outcomes could be prospectively defined and independently evaluated.

If it works, the evidence would support expansion.

If it does not, the results would still provide valuable information about how researcher development should be redesigned.

This approach reflects the principle underlying the broader recommendations in our response:

Do not only measure better research. Build better researchers—and measure whether better research follows.

Description: Supplementary report presenting a lifecycle-based framework for measuring scientific impact, the rationale and proposed indicators across all NIH RFI domains, and proposed NIH pilots for time-bounded open scientific review and lifecycle-based researcher competency development.

188. Xiaoxin Wang

I am responding: On behalf of myself

Name: Xiaoxin Wang

Name of Organization: Georgetown University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Nature published a paper a decade ago demonstrating that nearly 90% of tested landmark studies were not reproducible. By extension, a vast portion of grant funding could be supporting non-reproducible research. The frustrating reality is that the scientific community seems to have grown comfortable with this baseline.

With new administration, this issue must be brought back to the forefront. Our current peer-review system and study sections fail to prevent non-reproducibility—and may actively reward it. Without a dramatic overhaul of how grants are evaluated, I propose the following targeted changes:

Timing & PI Justification: Every grant application should explicitly answer two questions: What recent scientific or technological advance makes this proposed experiment viable right now rather than years ago? and Why is this specific PI, rather than their peers, uniquely qualified to execute it?

Verifiable & Standardized Data: Key data in grant and progress reports should prioritize publicly verifiable approaches (e.g., RNA-seq) deposited in public repositories where data manipulation is difficult. Alternatively, core samples should be submitted to NIH-designated facilities to generate independent validation. Easily manipulated, qualitative figures from traditional assays should no longer serve as primary proof.

Panel Anonymity & Annual Rotation: To eliminate cronyism and implicit bias, study section panels should rotate fully on an annual basis, and reviewer identities must remain undisclosed prior to review.

Fraud Flags & Accountability: Unreproducible findings should be treated as an explicit fraud flag rather than a harmless error, triggering closer administrative scrutiny. Furthermore, a paper's stated limitations must directly downgrade its claimed impact, preventing authors from making overblown claims while burying critical caveats in the fine print.

2. Data, Software, and Model Sharing: To accelerate scientific discovery through crowdsourced effort, I propose a fast-reaction funding mechanism dedicated to reproducible, novel observations. Because these early discoveries are initially unfunded, applicants should not be expected to provide complex preliminary experiments. Instead, funding should focus on validating the core observation, and the

originating Principal Investigator (PI) should have the option to delegate or assign specific follow-up studies to other qualified research groups.

3. Training and Mentorship: When the most experienced research workforce relies almost entirely on soft money, scientists face immediate displacement the moment a supporting grant ends. Without a viable solution to retain this veteran workforce, the rationale and necessity for continually training a young workforce become nearly impossible to justify.

4. Collaboration: Productive scientific collaboration relies on a fair system for awarding credit. To encourage this, I propose a funding mechanism where a PI with a novel finding is awarded a grant designed to bring in external collaborators to address secondary or complementary questions. Under this framework, the PI who generated the original finding receives a controlling 51% majority credit, while the remaining 49% is distributed among collaborating partners who address the necessary secondary questions to collectively advance the science.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

189. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Direct reproducibility studies are typically only good at finding that labs do things differently or use slightly different materials. Under close scrutiny the 'replication crisis' does not hold up in many fields. If a finding or idea is significant and impactful, then it is used as a foundation to build upon as others try to advance the finding. It is at this stage that reproducibility matters. If the idea can't be confirmed, then science moves on. I would oppose using funds to support experiments with the goal of simply testing if a finding can be reproduced in another lab. Funding impactful science as determined by peer-review that is free of political influence is the best way to address rigor and reproducibility.

2. Data, Software, and Model Sharing: The current sharing of data through publications and open-source databases and data warehouses is an important way that has advanced research. It is working well, as demonstrated by a study from Holst et al. published in Nature 2026 that disruptive science and technology have remained steady since the 1960s despite a significant reduction in government investment in science and technology when adjusted for inflation. If the NIH reduces or limits the support for paying publication charges occurs, this will significantly hamper the goal of sharing. Another important way of sharing data, software, and models is through conferences and collaborations (many of which start at a conference). Scientific communities have meetings that support real-time sharing that help advance research and technological development. Conferences allow scientists from small and large private companies, non-profits, governments, academia, and more to engage in sharing. Moreover, students, trainees, researchers, scientists, and lead investigators are involved, allowing the best ideas to flow from several places. Removing support for attending conferences and meetings will further harm the sharing necessary for breakthroughs.

3. Training and Mentorship: Mentoring and training the next generation of creative researchers must happen to build healthier communities and strong economies within the US. The NIH should be proud of its strong history in funding F, T, and K awards that have contributed to innovative breakthroughs and supplied a critical workforce that has strengthened the US economy. Funding research and training at universities through direct and indirect costs has made the NIH an amazing partner in mentoring, training, and supporting the clinicians who address our pressing medical needs and the scientific workforce that stimulates our economy. Peer review of these applications should remain. The NIH should create additional paths for F or T awardees to apply for transitional awards to help retain the initial investment in people. The timelines on K awards should be re-examined and extended to be more inclusive since some research takes longer than others (i.e., developing a truly innovative project related

to aging is likely to take more time than a project on a rapidly growing cancer). Importantly, NIH should focus on a decentralized approach in partnerships with states and universities to best fit the state and local communities' needs. NIH should provide a supporting role, not a regulatory role. Moreover, NIH investment in training and mentorship should help recruit the best and brightest in the world to come to the USA for training. We should also retain them to make the US's biomedical research and health systems the envy of the world. Restricting access would be counterproductive.

4. Collaboration: I agree that collaborations are essential for pushing our boundaries of knowledge and technology and am encouraged that the NIH may start to improve its support for collaborative research. Collaborations have become a bedrock of modern science. However, some NIH changes threaten to harm collaborations. For example, capping grant submissions to 6 has discouraged MPI projects. Some potential NIH changes further threaten collaborations. For example, capping a PI at 2, 3, or 4 grants. NIH should encourage and grow collaborations. One simple way to stimulate collaboration is to exclude collaborative grants from the aforementioned caps. Funding new collaborative projects would also help. The NIH could do this immediately by reducing the number of forward-funded and multiyear-funded NIH projects. Additionally, the NIH could create new initiatives to support interdisciplinary research without drawing resources away from institutes that have already faced multiple cuts.

5. Entrepreneurship and Translation: Entrepreneurship and translation should be an important aspect of NIH's responsibility to taxpayers. Support for startup entrepreneurship projects should be continued and should be done in partnership with non-profits, academic institutions, and private startups. However, it should not come at the cost of foundational (i.e., basic) research that all translational, innovation, and entrepreneurial effort is based on. Support for foundational research must remain a priority.

6. Foundational Scientific Exploration: I agree that we need to support innovative ideas. It is important to remember that paradigm-shifting breakthroughs can come from high-reward, well-justified, and feasible projects that are currently funded through peer review of investigator-initiated R01 projects. Recognizing the importance of high-risk, high-reward projects is easy, but supporting them is challenging. One easy step the NIH could take today is to exclude applications to high-risk funding mechanisms from the 6-application-per-PI cap and the potential 2-4 funded-projects-per-PI cap the NIH is considering. Additionally, the NIH could create new initiatives to support a high-risk, high-reward project on existing R01s (i.e., a new 3rd or 4th aim). This could be a separate, funded 5th and 6th year for a project that has led to an innovative, novel idea that is too high-risk for an R01 renewal.

7. Public Impact: It is important that the NIH helps educate taxpayers, politicians, and the public at large that research is a long game. Today's discoveries do not translate to tomorrow's cures. The process takes years, and it is important that the public understands the process and the importance of this timeline to public health and safety. Moreover, the NIH must continue to recognize that the impact of a project is best assessed by a strengthened peer review process. Politically appointed personnel that lack the true expertise that is provided by peer review should not override identification of the most meritorious research when making funding decisions. AI cannot assess impact and should not be used to make funding decisions. If the NIH uses politically appointed individuals or AI to override peer review, trust from taxpayers and stakeholders in the NIH will be further eroded. Finally, the NIH needs to invest in research that will impact policy making. The public needs to feel the impact as they have enjoyed with clean water, safe food, and clear air to breathe that we enjoyed prior to 2025.

8. Other Areas: The NIH must restore trust from taxpayers and stakeholders, not be leading the charge to further erode it. Trust in the NIH is falling more rapidly now than at any point in my career. Currently, policy decisions are made without scientific foundations or unbiased data. Attacks on scientists, including previous government officials, are further harming progress that the NIH can make at this time. Changes outlined in “Strengthening Replication and Reproducibility of NIH-funded Research” and “Leading in Gold Standard Science” will be with skepticism because of the attacks on the CDC and NIH. The NIH should not move forward with the recently proposed changes. Instead of focusing on changing the way science is done, it would be better if we had an NIH administration that was focused on supporting research.

190. Maryam Al-Mujtaba

I am responding: On behalf of myself

Name: Maryam Al-Mujtaba

Name of Organization: Avarynth Group

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Research assessment should explicitly distinguish the quality of research from its reception. Publication counts, journal prestige, and citation metrics may capture aspects of visibility or reach, but they do not directly measure whether research was conducted rigorously or whether its findings can be reproduced. Future assessment systems should therefore explicitly reward methodological rigor, transparency, scientific integrity, and reproducibility.

I recommend that research assessment treat quality and reach as two distinct, separately measured dimensions rather than combining them into a single score. Quality—including contribution to knowledge, methodological rigor, and reproducibility, should be evaluated through structured expert judgment, similar to the way grant reviewers evaluate proposals: a reviewer examines the work in depth and translates that judgment into a structured assessment. Reach, by contrast, should be measured separately and quantitatively, with appropriate adjustments for factors such as accessibility and career stage so that these measures are not systematically distorted by them.

This mixed-methods approach is reflected in my Ten Triple R Principles for Guiding the Development of Future Research Assessment Metrics, which call for expert judgment for dimensions that cannot be meaningfully reduced to counts, alongside appropriately adjusted quantitative measures for dimensions that can be measured quantitatively, within a transparent and multidimensional assessment system. I am currently developing a working framework built from these principles and expect to share it as it matures.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: Research assessment should recognize that scientific contribution extends beyond publications and citations. Training and mentorship are essential mechanisms through which knowledge, skills, and scientific standards are transmitted across generations of researchers.

Specific measurable indicators could include: mentoring activities; trainee completion and progression; development of trainees into independent researchers; contributions to methodological or professional development; and evidence of sustained mentorship relationships. Quantitative indicators should be interpreted alongside qualitative evidence because trainee outcomes are influenced by discipline, career stage, institutional resources, and other contextual factors.

Implementation approach: NIH could allow investigators to document mentorship contributions and outcomes as a distinct category of scientific contribution, supported by structured narrative evidence and selected indicators.

Potential benefit: This would reward researchers for investing in the scientific workforce rather than creating incentives that prioritize individual productivity at the expense of mentorship.

Feasibility consideration: Assessment should avoid comparing early-career and senior investigators using identical cumulative indicators. Career-stage adjustment is essential because opportunities and accumulated outcomes differ substantially across stages.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Scientific impact extends beyond citations and academic attention. Research can influence clinical practice, policy, education, public understanding, innovation, accessibility, and societal well-being. Future assessment systems should therefore distinguish attention from actual contribution. A paper may receive substantial media or social-media attention without producing meaningful impact, while research with limited academic visibility may substantially influence policy or practice.

I recommend assessment approaches that explicitly separate volume of attention from demonstrated translation into clinical practice, policy, education, industry, innovation, or other forms of meaningful use. Public impact should be credited where genuine application or use of research can be demonstrated, rather than inferred solely from mentions, media coverage, social-media attention, or other indicators of visibility.

My The Ten Triple R Principles for Guiding the Development of Research Assessment Metrics identify societal impact and knowledge translation as core dimensions of research assessment, alongside the production of knowledge itself. The principles also call for measures of research reach to account for factors such as accessibility and career stage, so that equally rigorous and valuable work is not systematically undercounted because of unequal access to the infrastructure through which visibility and citations accumulate.

I am currently developing a working framework built from these principles and expect to share it as it matures.

8. Other Areas: A principles-based approach to the development of future research assessment metrics

I propose that NIH consider not only which individual indicators should be added to existing systems, but also the principles that should guide the development, governance, and implementation of future research assessment metrics.

I have developed the Ten Triple R Principles for Guiding the Development of Research Assessment

Metrics, organized into three domains: foundational principles, assessment principles, and system design principles.

Foundational Principles

1. Understand the Philosophical Divide: Research assessment is not purely a technical exercise. The metrics used to evaluate science operate within scientific, economic, and political systems. Those developing future assessment systems should consider how incentives, ownership, and governance shape what research is rewarded.

2. Researchers Must Author and Ratify the Standards: The research community should play a central role in developing and collectively validating the standards by which scientific contributions are assessed.

Assessment Principles

3. Measure Contribution, Not Prestige: Research should be assessed according to what it contributes to knowledge rather than the prestige of the journal in which it appears.

4. Reward Methodological Rigor: Assessment should explicitly recognize methodological quality, transparency, scientific integrity, and trustworthy evidence.

5. Value Reproducibility: Research should be rewarded for being sufficiently transparent and robust to permit testing, reproduction, replication, and cumulative scientific progress.

6. Adjust for Accessibility: Measures of reach should account for differences in access to research because accessibility affects visibility and the opportunity for knowledge to be used.

7. Adjust for Career Stage: Cumulative indicators such as citation counts should not systematically advantage researchers simply because they have had more years to accumulate them.

8. Recognize Societal Impact and Knowledge Translation: Research assessment should recognize meaningful contributions to policy, clinical practice, education, innovation, industry, and public engagement.

System Design Principles

9. Assess Through Mixed Methods: Research assessment should combine structured quantitative indicators with expert human judgment rather than assuming that either approach alone can adequately capture scientific contribution.

10. Transparency in Assessment Methods: Researchers should be able to understand how assessment systems operate, how indicators are calculated, who governs them, and how they evolve.

Together, these principles suggest that future research assessment should be multidimensional. No

single metric can adequately capture the quality, rigor, reproducibility, reach, accessibility, translation, and broader contribution of scientific research. NIH could support the development and empirical testing of assessment approaches that combine multiple dimensions rather than seeking a single replacement for publication counts or citation metrics.

Finally, I recommend that any future assessment system developed in response to this initiative be created through meaningful engagement with researchers across career stages, disciplines, institution types, and sectors. The people who produce and validate scientific knowledge should play a central role in defining how scientific contribution is measured and rewarded.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/M.Al-Mujtaba_NIH_RFI_Research-Assessment.pdf

Description: The Ten Triple R Principles for Guiding the Development of Research Assessment Metrics," developed by Maryam Al-Mujtaba. The document outlines a principles-based approach to the development, governance, and implementation of fair, comprehensive, and multidimensional research assessment systems.

I propose the Ten Triple R Principles for Guiding the Development of Research Assessment Metrics, organized in three parts: the foundations on which assessment should rest, what assessment should measure, and how the assessment system itself should be designed (see Figure 1).

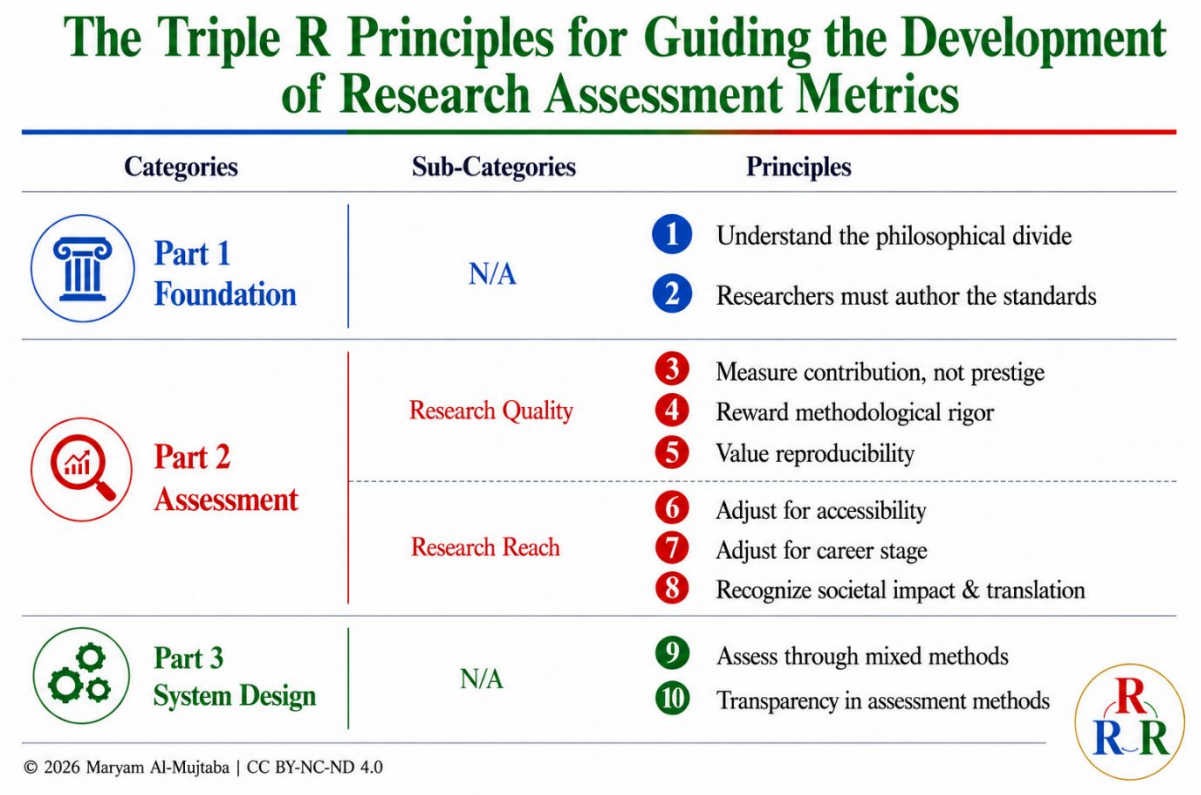


Figure 1: The Triple R Principles for Guiding the Development of Research Assessment Metrics

Part I: Foundational Principles *Why we assess, and who should hold the authority to define it.*

Principle 1, Understand the Philosophical Divide: Academic publishing has three dimensions: the scientific, the economic, and the political. For decades, our philosophical perspective as researchers has confined us to the first. We have seen science as our contribution to society and peer review as our service to it, and in doing so we left the economic and political dimensions uncontested. Publishers saw the same things differently: not as contribution and service, but as free content and free labor to be monetized. That is *The Philosophical Divide*, and it is how an equitable bargain, in which researchers created knowledge and the system served them, was quietly broken. Anyone building a new system of research assessment, and it should be researchers who build it, must understand this divide, because it determines what they build. A researcher who does not see the economic and

© 2026 Maryam Al-Mujtaba | Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0)

political forces at work could, however unintentionally, produce another metric that feeds the current inequitable business model.

Principle 2, Researchers Must Author and Ratify the Standards: The authority to define good science should rest with those who produce and validate it. Researchers should not only author the criteria by which research is assessed but also collectively ratify them. This is precisely what never happened with the impact factor. It was devised by the information scientist Eugene Garfield as a commercial tool to help libraries choose which journals to subscribe to, not as a measure of research quality. No consensus was reached, no policy was agreed; it simply drifted into use, repurposed until the journal you published in came to define how good a scientist you were. A standard that governs hiring, promotion, and funding should never be adopted by default.

Part II: Assessment Principles *What good research assessment should measure.* These principles fall into two groups. The first concerns the quality of the research itself, what it contributes and how soundly it was done. The second concerns its reach, how far, and how fairly, that knowledge travels into the world. *Today's metrics largely measure reach (citation count) and call it quality.* A fair assessment must measure both and keep them distinct.

Research Quality

Principle 3, Measure Contribution, Not Prestige: The reputation of a journal is not the same as the value of a piece of research. Neither journal prestige nor citation counts alone should determine scientific quality. Research should be evaluated according to what it contributes to knowledge.

Principle 4, Reward Methodological Rigor: Scientific progress depends not only on novel ideas but on trustworthy evidence. Research assessment should explicitly value methodological rigor, transparency, and scientific integrity, the qualities that make evidence trustworthy.

Principle 5, Value Reproducibility: A finding that cannot be reproduced cannot be relied upon. Research assessment should reward work transparent enough to be tested, replicated, and built upon by others, because reproducibility is the foundation of cumulative science, not a bureaucratic checkbox.

Research Reach

Principle 6, Adjust for Accessibility: Knowledge cannot create impact if it cannot be accessed, and citation-based measures quietly punish work that is harder to reach. Open-access research is systematically more visible, and therefore more cited, than work behind a paywall, independent of its quality. Any assessment that uses citations as a measure of reach must correct for this, so that a paper is not judged less influential simply because fewer people could access it.

Principle 7, Adjust for Career Stage: Citation counts accumulate over time, so they inevitably favor established researchers over early-career ones. An early-career researcher has had fewer years for their work to be read and cited, not because it is less valuable, but because it is more recent. Where citation-based measures are used, they must be calibrated

© 2026 Maryam Al-Mujtaba | Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0)

for career stage, so that two equally strong researchers are not ranked differently simply because one has had twenty years and the other three.

Principle 8, Recognize Societal Impact and Knowledge Translation: Research exists to improve the world beyond academia. Helping knowledge reach and be used by others, through translation into policy, clinical practice, education, industry, innovation, and public engagement, matters just as much as creating it in the first place. Scientific influence extends well beyond citations.

Part III: System Design Principles *How the assessment system itself should operate.*

Principle 9, Assess Through Mixed Methods: The two halves of assessment call for two different methods. The dimensions of quality, contribution, rigor, and reproducibility cannot be reduced to citation counts; they require expert human judgment. However, that judgment is not a vague verdict. Just as grant reviewers read a proposal in depth and then assign it a score, expert assessors evaluate a work's quality and express that evaluation as a structured number. The dimensions of reach, by contrast, are quantitative from the start, captured through numbers corrected for the conditions that distort them. Research assessment should therefore be mixed-methods, reflecting the methodological pluralism that already characterizes good research. Both halves are expressed as scores and combined into a single, structured assessment, so that expert judgment and measured reach carry transparent, comparable weight. Neither the reviewer's judgment nor the raw number is sufficient by itself.

Principle 10, Transparency in Assessment Methods: A metric that cannot be understood cannot be trusted. If researchers do not know how an assessment is calculated, who controls it, or how it changes over time, the door reopens to the very capture these principles resist. Future research assessment must be transparent in its methods, its governance, and its evolution, open to scrutiny by the community it serves.

How to Apply These Principles: A single system or metric may not comprehensively address every principle across all three categories, and that is to be expected. But all three categories matter; assessment must be multidimensional, and a legitimate system must reflect that:

1. **Foundational Minimum:** The two foundational principles (*Understand the Philosophical Divide* and *Researchers Must Author and Ratify the Standards*) are non-negotiable and must be fulfilled.
2. **Assessment Minimum:** Within the assessment category, a system must address at least one of the three principles in the research quality subcategory and at least one of the three in the research reach subcategory, so that it captures different dimensions of a work's value and produces a more rounded assessment, rather than mistaking reception for contribution.
3. **System Design Minimum:** It must satisfy at least one of the two system design principles, so that the way it operates is itself sound.

Why Principles Matter: For decades, we have debated which metric should replace the impact factor. Perhaps we have been asking the wrong question. Instead of asking which metric, we should first ask which principles should guide the development of those metrics.

© 2026 Maryam Al-Mujtaba | Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0)

Consider where we are now. As one recent review notes, the scientific community has developed more than sixty indices to evaluate researchers (Bihari et al., 2023). Beyond the impact factor, the h-index, SCImago Journal Rank, Field-Weighted Citation Impact, and Altmetrics that I discussed in Issue 9, there is the g-index, the w-index, the Ψ-index, and dozens more. But look closely, and nearly all of them are measuring the same thing. They are citation-based, differing mainly in the formula applied to the same underlying data: some count an author's citations, some a journal's, some weight highly-cited papers more heavily. The few that reach beyond citations do not go far. Altmetrics, for instance, tries to capture societal reach by tracking how often a work is mentioned in the news, on social media, in policy documents, or online. But it counts attention, not impact. It records that a paper was mentioned, not why. A work might be mentioned because it informed a policy or improved practice, or because it is controversial and people are citing it to say it is wrong. The metric cannot tell the difference. Attention is not the same as contribution.

I am not the first to propose principles for research assessment: Over the past decade, these are the four documented that I came across:

1. *The DORA* (2013) urged us to stop treating the Journal Impact Factor as a measure of individual quality.
2. *The Leiden Manifesto* (2015) offered ten principles for the responsible use of metrics, insisting that quantitative indicators support rather than replace expert judgment.
3. *The Metric Tide* (2015) set out five dimensions of responsible metrics: robustness, humility, transparency, diversity, and reflexivity.
4. *The Hong Kong Principles* (2019) focused on rewarding research integrity.

These four principles, for all their value of telling evaluators how to use metrics responsibly; they share three limits:

1. They do not tell us how to design the metrics in the first place. They are principles for behavior, not blueprints for building. Second, none confronts the economic and political capture of research assessment.
2. The Philosophical Divide is absent from all of them.
3. None of them was ever operationalized into a working assessment instrument. They are declarations and guidelines.

A decade on, we still evaluate research with the very metrics they criticize, because principles alone change nothing until someone builds the tool that embodies them.

I developed the Triple R Principles independently, from my own analysis of the economic and political dimensions of academic publishing. Yet several of my principles converge with theirs: mixed-methods assessment, transparency, and fairness appear in their work as well as in mine. I take that convergence as a strength. When people reasoning from different starting points arrive at the same conclusions, those conclusions are more likely to be sound. However, the Triple R Principles go further. Where the others guide how to *use* metrics responsibly, mine are designed to guide the *development* of research assessment metrics that are fair, comprehensive, and multidimensional, and that can be operationalized into a working framework (See Figure 2).

© 2026 Maryam Al-Mujtaba | Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0)

The Ten Triple R Principles in Relation to Existing Principles

Category	Triple R Principle	Annotation (What This Principle Does)	Explanation (Why It Matters)	Relation to Existing Principles
Foundation Why we assess, and who should hold the authority to define it.	1 **Understand the Philosophical Divide**	• Clarifies the philosophical differences in research values that shape assessment.	• Makes explicit the value divide behind assessment systems.	**New** — no existing principle explicitly states either of these two principles.
	2 **Researchers Must Author & Ratify the Standards**	• Ensures researchers are the authors and approvers of the standards that assess them.	• Returns authority and accountability to the research community.	
Assessment: Quality The quality of the research itself — what it contributes and how soundly it was done.	3 **Measure Contribution, Not Prestige**	• Evaluates what researchers achieve, not where they publish.	• Shifts focus from journal prestige to real research contribution.	**Convergent** — central to DORA; echoed in the Leiden Manifesto.
	4 **Reward Methodological Rigor**	• Values robust, transparent, and reproducible methods.	• Rewards the rigor of how research is designed and conducted.	**Convergent** — central to the Hong Kong Principles.
	5 **Value Reproducibility**	• Promotes openness and the ability to verify and build on research.	• Encourages transparency and verifiability as core values.	**Convergent** — reflected in the Hong Kong Principles.
Assessment: Reach How far, and how fairly, knowledge travels into the world.	6 **Adjust for Accessibility**	• Ensures research is accessible to all, broadening reach and impact.	• Corrects for barriers that limit dissemination and visibility.	**New** — no existing principle explicitly states this.
	7 **Adjust for Career Stage**	• Addresses structural disadvantages faced by early-career researchers.	• Creates a fairer playing field across career stages.	**Partial** — existing principles recognize this as an issue but do not provide an explicit principle or requirement to address it.
	8 **Recognize Societal Impact & Knowledge Translation**	• Elevates contributions that advance society and translate knowledge beyond academia.	• Values impact in the real world, not just within academia.	
System Design How the assessment system itself should be designed.	9 **Use Quantitative Assessment Measures**	• Balances metrics with expert judgment for a holistic assessment.	• Prevents overreliance on metrics and preserves expert judgment.	**Convergent** — Leiden Manifesto Principle 1 (supporting quantitative data with qualitative, expert judgment).
	10 **Transparency in Assessment Methods**	• Ensures openness, clarity, and accountability in assessment processes.	• Builds trust and enables sccutiny, improvement, and accountability.	**Convergent** — Leiden Manifesto Principle 4 ("Keep data collection and analytical processes open, transparent and simple"); Metric Tide (Transparency).

Figure 2: The Ten Triple R Principles in Relation to Existing Principles

© 2026 Maryam Al-Mujtaba | Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0)

© 2026 Maryam Al-Mujtaba | Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0)

191. The Massive Analysis and Quality Control (MAQC) Society

I am responding: On behalf of an organization

Name: Wendell Jones

Name of Organization: The Massive Analysis and Quality Control (MAQC) Society

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: The MAQC Society (maqcsociety.org) appreciates this opportunity to respond to the NIH Request for Information on Measuring and Rewarding Scientific Impact. As an organization dedicated to advancing quality control and reproducibility in biomedical research (Shi L et al., Nature Biotechnology 2017, PMID: 29220036), we strongly support NIH's efforts to align its measures of scientific success with its core mission of generating knowledge that enhances health, lengthens life, and reduces illness and disability. We focus our comments on the category of Rigor and Reproducibility, drawing upon our recently accepted position paper, "Ensuring multiomics data reproducibility for artificial intelligence with reference materials as a common calibrator" (Shi L et al., Nature Biotechnology, in press); a proof copy is attached for reference.

The Challenge: A Systemic Root Cause of Irreproducibility
For three decades, the omics field has struggled with batch effects and irreproducibility that fundamentally undermine the rigor of biomedical research. A primary root cause is a deeply flawed practice: treating raw instrument responses (fluorescent intensities, read counts, or peak areas) as direct surrogates for analyte concentration. Because instrument sensitivity varies substantially across batches, laboratories, and platforms, the same underlying concentration can generate dramatically different numerical values under different conditions. This technical variation is often larger than true biological signals of interest and can become amplified by AI/ML models as spurious biological features. Hundreds of batch-effect correction algorithms have been developed. However, without proper calibration or stable reference material, these methods do not directly measure or address the root cause and thus, can induce technical effects as well as remove them.

A Solution: Reference Materials as a Common Calibrator
The MAQC Society proposes a practical, measurement-based solution: the routine co-profiling of community-adopted reference materials alongside study samples, with results reported as sample-to-reference ratios (SRR). When a common reference sample is analyzed in the same batch as study samples, the ratio of the study sample response to the reference response removes the batch-specific sensitivity factor on a feature-by-feature basis. This converts unstable, batch-specific instrument responses into measurements anchored to a common calibrator, making data far more comparable across batches, laboratories, platforms, and time. The Quartet multiomics project has demonstrated the feasibility of this approach across epigenomics, transcriptomics, proteomics, metabolomics, and other platforms (www.nature.com/collections/ibadeahigd). We note that SRR is platform-agnostic and can be

implemented with existing reference materials from NIST, the Quartet project, and other community resources.

Specific Recommendations for NIH
We respectfully urge NIH to consider the following specific, measurable indicators and implementation approaches:

1. Require Reference Material Co-Profiling as a Default Expectation for Funded Omics Studies. The cost of co-profiling with an accessible reference sample is modest compared to the cost of irreproducible research. NIH should establish grant review criteria that evaluate whether applicants have included appropriate reference materials in their experimental design and budget. Pilot programs could support the development of matrix-specific references for underserved areas (e.g., tissues, stool, single-cell, or spatial omics). This requirement would serve as a measurable indicator of rigor that is objective, feasible across institution types, and scalable.

2. Fund the Metrological Characterization of Community-Adopted Reference Materials. NIH should support international consortia in developing additional reference material optimized for different biological specimens (such as specific tissues or blood) and fully characterizing the absolute concentrations of analytes in these additional and currently existing reference materials. This one-time investment would create a shared resource, bypassing the need for individual laboratories to perform costly SI-traceable measurements for every analyte in every study sample. It would also enable a future upgrade from SRR to absolute quantification, maximizing the value of NIH's research investments.

3. Incentivize Data Sharing That Includes Reference Material Raw Data. NIH should require that data deposition for funded omics studies include raw data from reference materials co-profiled with study samples, enabling independent SRR calculation, reanalysis, and quality evaluation. This provides an objective, measurable indicator of data quality and reusability that goes beyond mere publication counts.

4. Support Training in Measurement Science Fundamentals. NIH training programs should ensure that the next generation of biomedical researchers are trained in measurement science fundamentals, including the principles of calibration, reference materials, and quality control. This would create consistency and standards across career stages and institution types.

Potential Benefits
The SRR framework offers transformative benefits: improved reproducibility, enhanced data interoperability for AI/ML applications, reduced false discoveries, and more reliable translation of omics findings to clinical and regulatory settings. It shifts reproducibility from a retrospective validation to a proactive design principle.

The SRR approach is feasible across diverse research settings. For early-career investigators, co-profiling reference materials provides an objective quality control metric that can strengthen grant applications and publications. For resource-limited institutions, reference materials can be shared across laboratories, and the cost of co-profiling a small number of replicates is minimal compared to the total cost of a

typical omics experiment. For large-scale cohort studies, reference material co-profiling should be incorporated from the outset. For emerging technologies such as single-cell and spatial omics, analogous strategies using matrix-matched reference samples or synthetic spike-ins should be actively developed.

We acknowledge potential challenges. No single reference material can perfectly mimic every biological matrix, and matrix mismatch or commutability issues may persist. Analytes absent from a reference cannot be ratio-normalized. However, these are surmountable limitations that should be addressed through continued community investment rather than used as justification for inaction. The MAQC Society is actively working to expand reference materials for additional matrices.

Conclusion
Reproducibility is not a luxury; it is fundamental to the integrity of science. The MAQC Society urges NIH to embrace reference-material co-profiling using established community reference material and ratio-based quantification as a new default expectation for funded omics research. By requiring this practice, funding the characterization of community reference materials, and incentivizing training in measurement science, NIH can lead the biomedical community toward a truly reproducible foundation for discovery in the AI-omics era.

We thank NIH for this opportunity to provide input and stand ready to assist in implementing these recommendations.

Respectfully submitted,

The MAQC Society
(Correspondence: Wendell Jones, Wendell.Jones@iqvia.com)

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Multiomics-data-reproducibility-for-AI_eProof_NBT202607.pdf

Description: Latest MAQC position paper on Ensuring multiomics data reproducibility for artificial intelligence with reference materials as a common calibrator

https://doi.org/10.1038/s41587-026-03267-1

Ensuring multiomics data reproducibility for artificial intelligence with reference materials as a common calibrator

Leming Shi, Wenjun Bao, Aedín C. Culhane, Xiang Fang, Cesare Furlanello, Ruixin Hao, Wendell D. Jones, Rebecca Kusko, Samir Lababidi, Christopher E. Mason, Andreas Scherer, Shraddha Thakkar, Weida Tong, Russell D. Wolfinger, Joshua Xu, Ying Yu, Rui Zhang & Yuanting Zheng

Check for updates

Reference materials should be adopted as a common calibrator for multiomics measurement and co-profiled with study samples. Multiomics results should be reported as sample-to-reference ratios so that they are reproducible and suitable for artificial intelligence tools.

The convergence of multiomics, systems biology and artificial intelligence/machine learning (AI/ML) delivers tools such as virtual cells[1] and personalized predictive models[2], but these applications are limited by the quality of their training data[3]. Batch effects, measurement platform differences and other technical variations can become amplified features that AI/ML models learn as 'biological signals', which can then become downstream sources of spurious biological inference. Recent benchmarks show that several state-of-the-art deep-learning perturbation models do not yet consistently outperform simple linear baselines in current evaluation settings, underscoring the need for better data standards[4]. The emerging reproducibility crisis in AI-driven biomedicine will only intensify unless the community redefines multiomics data quality.

Flagship initiatives such as the UK Biobank, All of Us, Fudan Taizhou Longitudinal Cohort and China Kadoorie Biobank are generating vast multiomics datasets across millions of individuals. AI-driven research will use and integrate this data, requiring controls for accuracy and interlaboratory reproducibility. There already exist prespecified data acceptance criteria and well-established guidelines for bioanalytical method validation and study sample analysis for the pharmaceutical industry[5], and multiomics data should also adopt requirements that consider within-laboratory, cross-laboratory and cross-platform reproducibility.

Commercial plasma proteomics platforms (for example, Alamar NULISA, Illumina IPP, Olink Explore HT) rely on their own proprietary references to improve cross-batch reproducibility. However, proprietary references are not community-adopted standards and can limit interoperability across technologies by confining users to vendor-specific ecosystems. A shared, open, physically anchored standard set would support comparable measurements across laboratories, platforms and time, as advocated for blood proteomics[6]. Well-characterized reference-material resources already exist for several important biological and technical uses, including the Quartet multiomics materials[3] for cell-based applications from the Massive Analysis and Quality Control (MAQC) Society[7], materials for plasma metabolomics or proteomics[8] and liver proteomics[9] from the National Institute of Standards and Technology (NIST) of the United States of America, and the Plasmix suite of reference materials[10] developed by MAQC for plasma multiomics. These resources establish reference-material-based harmonization as already practical for parts of the multiomics ecosystem, and it can be expanded through coordinated community efforts.

Here, we propose the use of suitable community-adopted standards so that data generated today remain reproducible, interoperable, interpretable and suitable for future scientific or regulatory decision-making. This requires the development, management and continued use of reference materials[11], harmonized metadata, transparent sample processing and analysis pipelines, and AI-ready data values. Within this framework, reproducibility shifts from a retrospective validation to a proactive design principle. Operationally, this means that high-quality multiomics datasets should include stable reference anchors, harmonized metadata, transparent analytical provenance, and reporting formats that preserve comparability across batches, laboratories and future analytical methods[3].

The root cause of irreproducibility: instrument responses as surrogates of concentration

For single-analyte measurements, the instrument response is related to concentration by establishing a calibration curve using reference standards for that analyte. The concentration of a new sample can then be deduced from its instrument response and the calibration curve[5]. This approach is feasible when the number of analytes is small and standards are available for each analyte. In large 'omics studies, thousands to millions of analytes (genes, transcripts, proteins or metabolites) are measured simultaneously, and establishing a calibration curve for each analyte is practically impossible because of the lack of individual reference standards and prohibitive cost.

The 'omics field has adopted a pragmatic but deeply flawed shortcut since 1996 by treating raw instrument responses as direct surrogates for concentration[12]. The same underlying concentration of an analyte in the same sample can generate substantially different instrument responses when measured under various batch- or platform-specific sensitivity states (Fig. 1a). The consequence is that biological information from sample analytes becomes disguised by unwanted technical variation such as batch effects, directly undermining downstream

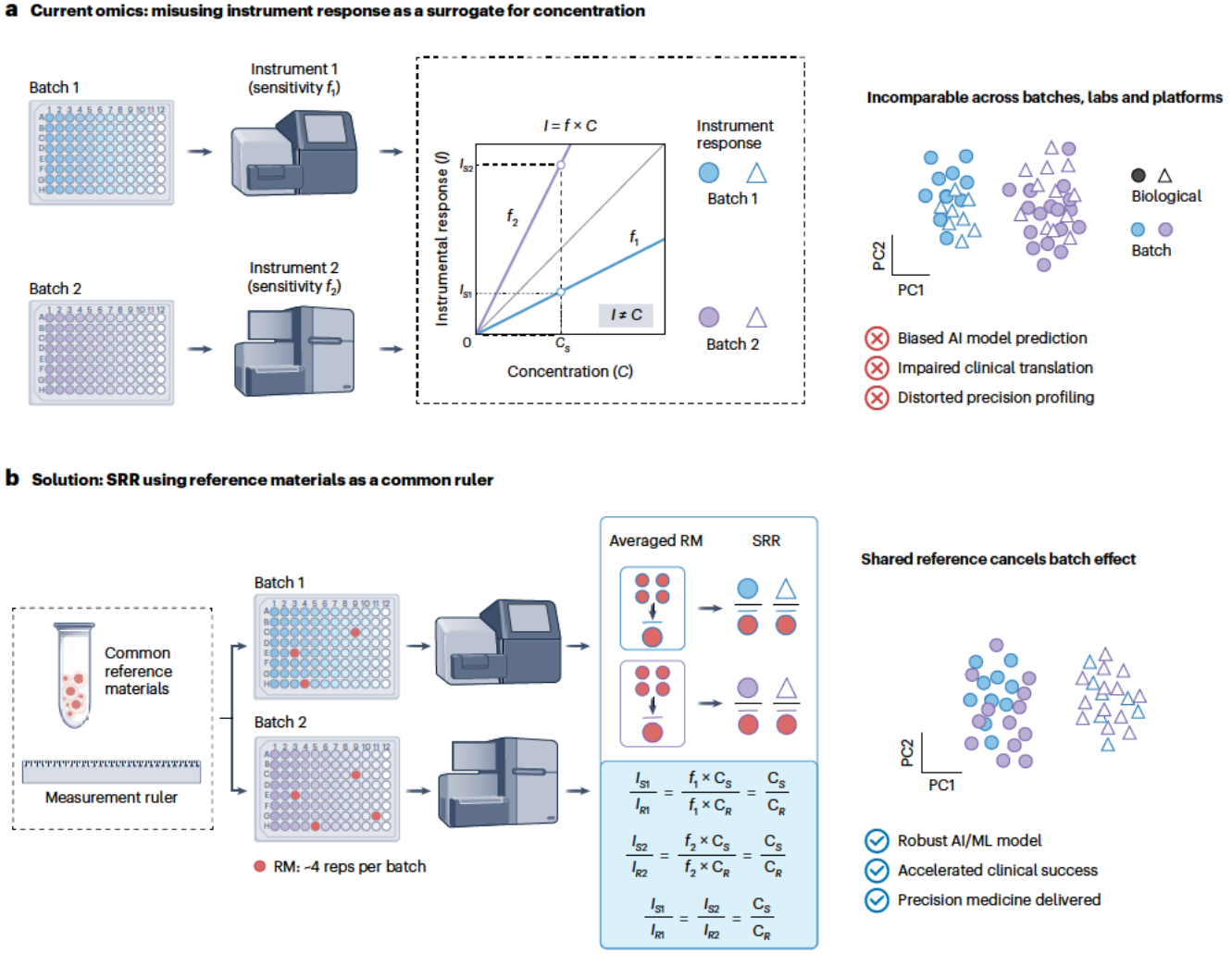


Fig. 1 | Root cause of multiomics data irreproducibility and the role of reference materials as a common calibrator. a, Current practice: raw instrument response is used as a surrogate for concentration. Because instrument sensitivity varies across batches, laboratories and platforms, the same analyte concentration can generate different numeric responses under different conditions (for example, $I = f \times C$ with $f_1 \neq f_2$). As a result, samples may cluster by batch rather than by biology, compromising downstream interpretation, AI/ML modeling, clinical translation and precision-medicine applications. PC, principal component. **b**, Proposed practice: a common reference material is co-profiled with study samples in every batch, typically in randomized technical replicates. Expressing each analyte as a sample-to-reference ratio (SRR; I_S/I_R) reduces the shared batch-specific sensitivity term and anchors measurements to a common calibrator. This makes data far more comparable across batches, laboratories, platforms and time, thereby improving the reliability of multiomics integration and AI-ready analysis.

inference, data integration and future AI/ML model training. The data generated within a single batch may appear 'high quality' when restricted to that batch yet become incomparable across batches. This paradox has misled and mystified the field for three decades.

The magnitude of batch effects is usually much larger than true, often subtle, biological signals of interest, such as the differences between responders and non-responders to a particular stimulus. Furthermore, batch effects are routinely confounded with biological signals, leading to false discoveries when batch differences are misinterpreted as biology, or true signals are attenuated when 'correction' algorithms remove genuine biological signals along with technical noise. Even when not confounded, deep learning models can detect batch effects and interpret them as biologically significant. Hundreds of increasingly complex batch-effect correction algorithms have been published and some of them widely used, yet none can recover information that was never comparable to begin with[13,14] (Fig. 1a).

Unfortunately, the 'omics field has come to accept batch effects as normal and unavoidable, without addressing the underlying and fundamental causes. Instrument responses (fluorescent intensities, read counts or peak areas) are widely used as surrogates of concentration but are often inherently variable as a result of feature-dependent fluctuating sensitivity and background noise. A major driver of

analyte-specific instrument sensitivity drift is platform hybridization dynamics. Microarrays, targeted sequencing capture, antibody-based proteomics and even some single-molecule methods rely on probe–target hybridization or binding. Small variations in temperature, buffer composition or incubation time cause analyte-specific changes in binding efficiency[15]. This is why batch effects are often feature-dependent and cannot be corrected by global scaling alone[13,14].

The solution: reference materials as a common calibrator

Reference materials possess homogeneity, stability, accuracy, traceability, cost-effectiveness and ample supply and are routinely used for measurement system calibration and quality control[3,11]. When a common reference sample R is analyzed alongside study samples S in the same batch, the ratio I_S/I_R eliminates the batch-specific sensitivity factor f on a feature-by-feature basis:

$$\frac{I_{S1}}{I_{R1}} = \frac{f_1 \times C_S}{f_1 \times C_R} = \frac{C_S}{C_R} \quad \frac{I_{S2}}{I_{R2}} = \frac{f_2 \times C_S}{f_2 \times C_R} = \frac{C_S}{C_R} \quad \frac{I_{S1}}{I_{R1}} = \frac{I_{S2}}{I_{R2}} = \frac{C_S}{C_R}$$

where I is the instrument response, C is the true concentration and f is the batch- and analyte-specific sensitivity factor, assuming a linear relationship between I and C[16].

Once a common reference material is measured in each batch, the shared batch effect in sensitivity factor f is captured in both the study sample and reference responses and is therefore reduced through sample-to-reference ratio (SRR) processing (Fig. 1b). This converts unstable, batch-specific instrument-response values into measurements anchored to a common calibrator so that biological structure is more likely to dominate over batch structure in downstream analyses[3]. This shift is important for AI/ML applications, in which technical artifacts can be interpreted as biologically meaningful signals.

The SRR approach produces more harmonized data across diverse sources and can provide an important physical anchor for the large-scale data integration required by AI-based biomedical research. When every dataset is expressed as ratios relative to suitable, common reference materials, data from different laboratories, platforms and time points become far more comparable. This is more than an incremental improvement: it is a practical component of the measurement infrastructure needed for robust AI-era omics. Indeed, recent evidence suggests that deep-learning model performance depends strongly on the calibration of evaluation metrics[17,18].

Physical reference-based harmonization using SRR and computational batch-correction methods are complementary rather than competing strategies: the former can improve measurement comparability at the source whereas the latter can further address residual structure that remains after experimental control[19].

The SRR concept is not limited to affinity-based vendor platforms. The Quartet multi-omics study demonstrated its use in liquid chromatography–tandem mass spectrometry-based proteomics and metabolomics, showing that ratio-based profiling with a concurrently measured common reference sample improves comparability across batches, laboratories, platforms and 'omics types[3,20]. Thus, SRR is a general quantitative strategy for multiomics and is platform agnostic. It can be implemented globally with any suitable reference materials, including those developed by NIST[8,9], the Quartet[3], the Plasmix[10] or future community standards[6,11].

No single reference material can perfectly mimic every biological matrix. The Quartet materials (cell lines) are most appropriate for blood derived or cultured cell studies; Plasmix is suitable for plasma and serum applications. The broad community must create appropriate reference materials for tissues, urine, cerebrospinal fluid, stool and other matrices. In addition, the reference-material-based SRR method is not limited to existing reference materials or technologies. Emerging technologies such as single-cell, spatio-temporal and low-input 'omics should actively pursue analogous strategies by designing matched reference samples or synthetic spike-ins that enable ratio-based calibration within their unique constraints. The underlying principle of co-profiling a common reference to reduce shared batch-specific variation should guide the design of calibration strategies across matrices and assays, even though the specific reference material will need to be adapted to each technology and use case.

The SRR approach has known limitations[3]. Analytes absent from a reference cannot be ratio normalized, matrix mismatch and commutability issues can persist, preanalytical variability (for example, collection or storage) is not corrected unless the reference is processed identically from the start, low-abundance response may be dominated by noise, and non-linear assay responses may violate the multiplicative model. The SRR approach assumes that background signal is low and relatively uniform between study samples and the reference material across analytes. For platforms with high or analyte-specific background, ratio calculation may amplify noise and compress the ratio values. In such cases, additional background correction may be needed, but co-profiling a reference material remains essential for estimating batch-specific parameters.

If the community collectively characterizes the true absolute concentrations of analytes in a commonly used reference material, then, in principle, the absolute concentration of an analyte in a study sample could be estimated by combining its SRR value with the calibrated concentration of the corresponding analyte in the reference material. In practice, this will also require attention to uncertainty budgets, commutability, analyte stability, and transferability across platforms and workflows. This 'one-time' characterization creates a shared resource, bypassing the need for individual laboratories to perform costly International System of Units (SI)-traceable measurements on every analyte in every study sample. We urge international consortia and standards bodies to invest in this collective absolute characterization initiative. By sharing the cost of fully characterizing community-adopted reference materials, practical, low-cost SRR can be achieved today, and a clear upgrade path to absolute concentrations (SI) can be realized in the near future.

Urgent call to action: co-profiling reference materials and reporting SRRs

The omics field has been challenged by batch effects, irreproducibility and data incompatibility for three decades. The SRR framework provides a practical path toward improved harmonization and reproducibility. Certified or well-characterized reference materials are available for several important use cases, additional matrix-specific standards are under development, and the conceptual foundations for broader adoption are in place. What is required is a coordinated implementation across the biomedical community to enable adoption of these practices at scale.

We call upon the following:

Funding agencies should require co-profiling of reference samples as a default expectation for funded 'omics studies. The cost of co-profiling with an accessible reference sample is modest compared to the cost of irreproducible research. Pilot programs could support the development of matrix-specific references for underserved areas

(for example, tissues, stool, single-cell or spatial). Agencies should also fund the development and metrological characterization of community-adopted reference materials (for example, absolute concentration assignment) to enable the future upgrade from SRR to absolute quantification.

Journals should adopt editorial policies that require authors to report whether suitable reference materials were co-profiled with study samples. Manuscripts omitting such controls should provide clear scientific rationale for their absence. Data deposition must include raw data from reference materials co-profiled with study samples, enabling independent SRR calculation, reanalysis and quality evaluation.

Regulators should update guidance documents to recommend, and in appropriate contexts encourage, reference-based calibration for regulatory submissions involving 'omics data, particularly in longitudinal, multi-site, and diagnostic-development settings. The US Food and Drug Administration has played a leading role in MAQC since its inception, and participation should be translated into policy. Regulators should also consider requiring SRR-based reporting for pivotal clinical studies that rely on 'omics endpoints.

Technology vendors should calibrate instruments using shared community reference materials where available and report SRR values as a primary harmonized output alongside raw instrument response and sufficient metadata to enable independent reanalysis. To ensure cross-platform compatibility, vendors should work with third-party reference-material providers to adopt well-characterized common reference materials, rather than perpetuating proprietary silos. Vendors should also adopt the platform- and omics-agnostic signal-to-noise ratio metric as an objective, overall quality measure of their products[3,20]. Vendors should ensure that users are trained to use reference materials appropriately before valuable study samples are analyzed.

Institutions and laboratories should implement SRR reporting as a standard operating procedure for 'omics experiments where appropriate references exist and participate in proficiency testing programs using common reference materials. Laboratory directors should allocate dedicated resources for reference material acquisition and co-profiling. At least three technical replicates of the reference material per batch should be included to estimate intrabatch variability and calculate SRR[20].

For large-scale cohort studies, researchers and consortia should incorporate reference material co-profiling from the outset. Researchers should also demand that vendors demonstrate their products' ability to generate reproducible data with high signal-to-noise ratio across various scenarios.

Training programs should ensure that the next generation of biomedical researchers are trained in measurement science fundamentals. For new 'omics technologies (single-cell, spatial or metagenomics), the community should include matrix-matched reference materials and SRR protocols as an integral part of method development, not an afterthought.

AI-enabled biomedical discovery, large-scale population cohorts and precision medicine initiatives all depend on reliable and interoperable 'omics measurements. The MAQC Society calls upon the entire biomedical community to embrace reference-material co-profiling and ratio-based quantification as a new default practice for 'omics data generation and reporting. Reproducibility is not a luxury: it is fundamental to the integrity of science[7]. Let us move beyond measuring "with rulers of uncertain length" or without rulers at all and build a truly reproducible foundation for the future of biomedical discovery in the AI-omics era with reference materials as a common calibrator.

Leming Shi[1] ✉, Wenjun Bao[2], Aedín C. Culhane[3], Xiang Fang[4], Cesare Furlanello[5], Ruixin Hao[6], Wendell D. Jones[7], Rebecca Kusko[8], Samir Lababidi[9], Christopher E. Mason[10], Andreas Scherer[11], Shraddha Thakkar[12], Weida Tong[13], Russell D. Wolfinger[2], Joshua Xu[13], Ying Yu[1], Rui Zhang[14] & Yuanting Zheng[1] ✉

[1]Fudan University and International Human Phenome Institutes, Shanghai, China. [2]JMP Statistical Discovery, SAS Institute, Cary, NC, USA. [3]University of Limerick, Limerick, Ireland. [4]National Institute of Metrology, Beijing, China. [5]HK3 Lab S.R.L, Orobix Life & LIGHT Center Brescia, Brescia, Italy. [6]opAIda, Inc, Princeton, NJ, USA. [7]IQVIA Laboratories LLC, Durham, NC, USA. [8]Cellino Biotech, Inc, Cambridge, MA, USA. [9]Office of the Commissioner, US Food and Drug Administration, Silver Spring, MD, USA. [10]Weill Cornell Medicine, New York, NY, USA. [11]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [12]Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA. [13]National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, USA. [14]National Center for Clinical Laboratories, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China.

✉e-mail: lemingshi@fudan.edu.cn; zhengyuanting@fudan.edu.cn

References

1. Roohani, Y. H. et al. *Cell* **188**, 3370–3374 (2025).
2. Wang, X. et al. *Nat. Rev. Clin. Oncol.* **23**, 180–200 (2026).
3. Zheng, Y. et al. *Nat. Biotechnol.* **42**, 1133–1149 (2024).
4. Ahlmann-Eltze, C., Huber, W. & Anders, S. *Nat. Methods* **22**, 1657–1661 (2025).
5. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. M10 on bioanalytical method validation and study sample analysis. (2023). Q10
6. Cai, X. et al. *Nat. Genet.* **57**, 2371–2380 (2025).
7. Shi, L. et al. *Nat. Biotechnol.* **35**, 1127–1128 (2017).
8. Mandal, R. et al. *Anal. Chem.* **97**, 667–675 (2025).
9. Davis, W. C., Kilpatrick, L. E., Ellisor, D. L. & Neely, B. A. *Sci. Data* **6**, 324 (2019). Q11
10. Liu, Y. et al. Preprint at *bioRxiv* https://doi.org/xxx (2026). Q12
11. Ren, L., Shi, L. & Zheng, Y. *Phenomics* **4**, 487–521 (2024).
12. Lockhart, D. J. et al. *Nat. Biotechnol.* **14**, 1675–1680 (1996).
13. Yu, Y. et al. *Genome Biol.* **24**, 201 (2023).
14. Yu, Y., Mai, Y., Zheng, Y. & Shi, L. *Genome Biol.* **25**, 254 (2024).
15. Jakubek, Y. A. & Cutler, D. J. *BMC Genomics* **13**, 737 (2012).
16. Shi, L. et al. *BMC Bioinformatics* **6**, S11 (2005). (Suppl. 2).
17. Miller, H. E. et al. Preprint at *bioRxiv* https://doi.org/xxx (2025). Q13
18. Zhang, M. et al. *Nat. Biomed. Eng.* **10**, 413–426 (2026).
19. Molania, R. et al. *Nat. Biotechnol.* **41**, 82–95 (2023).
20. Yu, Y. et al. *Nat. Biotechnol.* **42**, 1118–1132 (2024).

Acknowledgements

We thank F. Goodsaid, Z. Peng and W. Xu for comments on early versions of the manuscript. L.S. and Y.Z. were supported in part by the National Key R&D Project of China (2023YFC3402500 and 2023YFC3402501). The views expressed are those of the authors and do not necessarily represent the views or policies of their respective organizations. Q8

Author contributions

L.S. and Y.Z. drafted the manuscript. All authors contributed to the development of the concepts, reviewed the manuscript and approved its submission.

Competing interests

Certain authors are employees of SAS Institute, IQVIA, Cellino Bio, FDA, and Weill Cornell Medicine. L.S. and Y.Z. are co-leaders of the MAQC Quartet and Plasmix projects. Q9

192. PLOS

I am responding: On behalf of an organization

Name: Iain Hrynaszkiewicz

Name of Organization: PLOS

Type of Organization: Publisher

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: For at least a decade, we have known that reproducibility and replicability in the study of biological and living systems is difficult to achieve in practice. Myriad intrinsic and environmental factors, as well as experimental details and indirect measurements compound to challenge the ability to reproduce.
Several well documented studies have established the challenges one encounters when attempting to reproduce research studies (https://doi.org/10.7554/eLife.67995; https://doi.org/10.1038/s41586-025-09844-9). We propose that multiplying examples of lack of reproducibility by attempting to systematically reproduce more studies would not be particularly productive. Instead, we can learn broadly three things from these published replication initiatives:
1) Incentivize open science practices that facilitate replication attempts
Published replication initiatives have highlighted that a surmountable challenge for replication attempts is the availability of data, code, and methodological details (https://doi.org/10.7554/eLife.67995). Incentivizing the sharing of these outputs alongside study findings, and incentivizing open science practices in general, is a logical step towards facilitating replication attempts when they are deemed necessary. Many resources already exist that the NIH could use and emulate, for example:
The National Academies have published a toolkit to align incentives by raising expectations of open science practice at multiple stages of the research life cycle and researchers' career: grant application, grant report, hiring, tenure and promotion, annual faculty reports, etc (https://doi.org/10.17226/26308).
Publishers and replication experts have developed reporting guidelines to support transparency, rigor and reproducibility of published research (e.g. https://doi.org/10.1073/pnas.2103238118.) Raising an expectation of this kind of level of transparency early in the research process is essential.
Reporting transparency and access to all outputs of research does not guarantee rigor and reproducibility but it facilitates examination of rigor and replication attempts. Recognizing and rewarding responsible, open science practices wherever researchers are evaluated will help foster a culture of transparency.
2) Support infrastructure that enable open science practices
The efforts required to effectively share various research outputs should not be an undue burden and researchers need access to well-maintained, user-friendly repositories, metadata standards, and other resources and support across the research lifecycle to support transparent and reproducible practices across the research lifecycle.
The development and maintenance of open infrastructure supporting these practices is essential to ensure the effectiveness of researchers' efforts. A robust, resilient, well-maintained infrastructure helps

ensure the value of these outputs.
3) Fund studies that probe additional parameters that can be optimized for reproducibility
Experimental design, discipline-specific factors and environmental factors also influence the rigor and replicability of findings. As an example, a report of the NIH Advisory Committee to the Director issued in June 2021 outlines many of these factors in the case of animal research (https://acd.od.nih.gov/working-groups/eprar.html). Many of these recommendations would apply to other types of experimental work. Discipline-specific studies (such as those performed in the social and behavioral sciences – https://doi.org/10.1038/s41586-025-09844-9) and experts committees (https://acd.od.nih.gov/working-groups/eprar.html) exploring the impact of specific factors will help provide better guidance for optimizing rigor and reproducibility by improving study design and reporting of key parameters.

2. Data, Software, and Model Sharing: Sharing data, software and models and other research outputs and practices can help enable reproducible research. Indicators of transparent research practices can provide useful information about whether the methods and other information needed to assess and reproduce research are available. How measures of rigor and reproducibility, and other new metrics, are used in practice is also important. Numerous studies have highlighted the challenges of independently reproducing much (published) research, which suggests that caution is needed in using indicators/metrics of reproducibility as compliance enforcement tools or as a basis for punitive action, not least because current technological approaches are not able to make these judgements at a meaningful scale. Rather, PLOS views new metrics and indicators as a means to understand researcher practices, how they differ in different contexts and why, and then use the insights they provide to create positive incentives to improve practices.

A key example of this approach is PLOS's Open Science Indicators (OSIs), which measure the prevalence of transparent and responsible research practices in articles. PLOS developed OSI according to six principles (https://doi.org/10.6084/m9.figshare.21640889), one of which is to use these metrics responsibly, such as not using these metrics to create rankings or league tables of researchers based on open science (or reproducibility metrics), but using this information to understand practices and monitor for improvement, rather than compliance.
PLOS publishes OSIs for all articles it publishes, as well as for a comparator group of non-PLOS articles available in PubMed Central. The indicators can be explored through the PLOS Open Science Indicators dashboard: https://explore.plos.org/open-science-indicators
For rigor and reproducibility specifically, the availability of complete and detailed research methods and protocols ("protocol sharing"), and the registration or preregistration of research ("study registration"), are particularly relevant.
Protocols can be shared publicly in several ways: through dedicated repositories and platforms such as protocols.io, as standalone methods and protocol articles, or as supplementary information. Research conducted by PLOS found that around one third of researchers share methods and protocols publicly (https://doi.org/10.7717/peerj.16731).
Other potential indicators of good methodological practice and research reporting include the use of community-endorsed reporting guidelines. PLOS supports the Transparency and Openness Promotion (TOP) Guidelines, which cover a range of research practices relevant to rigor, reproducibility and open sharing, and we recommend that NIH consider the TOP Guidelines when identifying potential indicators (updated in 2025: https://doi.org/10.1186/s41073-026-00223-0)
While transparency and availability of research methods are useful indicators, there are gaps in our

collective ability to assess credibility and quality of research methods at scale, and to develop new indicators that better meet the needs of research evaluators. Research by PLOS (https://doi.org/10.7717/peerj.20502) found that, when assessing research outputs as part of hiring, promotion and grant review, evaluators look for indications that:
the research is well designed;
the statistical analyses are valid;
the research was conducted ethically; and
there are no signs of fabrication, falsification or plagiarism.
But there are not yet convenient or widely accepted indicators for all these qualities. This contributes to the use of measures such as Journal Impact Factor and personal or institutional reputation as proxies for research rigor and quality, even though they were not designed for that purpose.
More work is needed to test how new indicators are used in practice, and whether they can provide better alternatives to the inappropriate proxies for research quality and impact that are often used in research assessment. To do this effectively, NIH will need indicators that can distinguish not only whether sharing takes place, but whether shared outputs are usable and contribute to further research. This can help ensure that recognition and incentives are directed towards practices that add value, as well as when sharing is primarily for transparency and accountability purposes.

Concerning data and software sharing, indicators can be considered in two areas: the prevalence of data and code / software sharing, and the quality and impact of the outputs that are shared. Indeed, PLOS's OSIs also include indicators for data and code or software sharing. Similar indicators are used in initiatives such as the French Open Science Monitor. For PLOS, these indicators have been useful in understanding differences in practice between research domains and regions, and in developing policies, tools, partnerships, incentives and training to improve those practices.

Measuring the prevalence of sharing is a useful and pragmatic starting point. More important, but also more difficult, is understanding the quality and impact of the data, software and code that are shared. A number of studies have linked the sharing of research data and code supporting publications with increased citations (academic impact) to the associated research, for example:
https://doi.org/10.1162/QSS.a.487
Reuse of data and code/ software in new research may provide a more direct indication of the impact of shared data and code than citations to the associated article. Traditional bibliometric approaches do not capture most instances of reuse, although new approaches are beginning to emerge. PLOS collaborated with DataSeer through the UK Reproducibility Network Open Research Indicator pilots to develop an experimental indicator for measuring data reuse using agentic artificial intelligence. This work suggests that data reuse in publications may be close to 40% (https://arxiv.org/abs/2604.28061) The work is still experimental, but it points to the potential for developing indicators that measure reuse more directly. Indicators of quality could include the presence of complete documentation and technical information, such as README files, which help researchers assess and reuse data and code. Other recent research by PLOS found that this kind of documentation is important to researchers when considering whether to read or potentially reuse research outputs (https://doi.org/10.6084/m9.figshare.32287383).

3. Training and Mentorship: N/A

4. Collaboration: Collaboration and team science rely on the ability of researchers to build on each other's expertise and outputs. It is essential therefore that different types of contributions corresponding to this expertise and outputs are recognized and that they are visible. Achieving this more broadly necessitates many stakeholders including funding agencies, research institutions and publishers acting at multiple levels: We would therefore like to bring three major, connected initiatives to NIH's attention:
Create a culture of inclusive credit that considers substantial contributions to team science as bona fide research contributions. Authorship of scientific articles remains a primary currency of credit and the culture of authorship needs to evolve to reflect the evolution of research conducted. A working group convened by the National Academies has recently published a set of recommendations to create a more inclusive, adaptable and responsible authorship culture (https://doi.org/10.1073/pnas.2531268123). PLOS is implementing these recommendations across our journal portfolio. The white paper also contains specific recommendations for research institutions and research funders that apply to the NIH.
Using standards and approaches that enable different contributions to be captured and monitored. We recommend wider use of Contributor Roles Taxonomy (CRediT) to capture the diverse contributions of researchers . The CRediT taxonomy provides a controlled list of contributor roles that provides greater nuance to researchers' contributions to a research study beyond their inclusion in an article's author list, and can also be applied to datasets and other research outputs where the supporting infrastructure allows. It provides a more detailed account of contributions such as data curation, software development, supervision, methodology and writing, and can help make visible the different ways individuals contribute to multidisciplinary and distributed research teams. A number of publishers, including PLOS, use CRediT across all articles they publish. Wider adoption could help NIH recognize a broader range of contributions to team science and understand more about the composition of collaborative research teams.
Rethink the future of scientific communication: PLOS is continuing to explore and develop ways of increasing the visibility of diverse contributions through the systems of scholarly communication, in addition to research articles – to provide a basis on which recognition mechanisms such as those sought by the NIH can be built. This is represented in our efforts to develop a new publishing model (a "knowledge stack") that connects articles and preprints with associated research outputs — data, code, methods, and materials — into a structured, open, machine-readable record that reflects the research process and credits everyone who contributed. The model is designed to work with the open science infrastructure and standards researchers already use, such as persistent identifiers, repositories, and the CRediT taxonomy. The knowledge stack concept is designed to provide the structured relationships, attributions and context that make the distributed outputs visible as a coherent whole to help credit, understand, verify and reuse the numerous different contributions to collaborative science.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: In February 2025, PLOS convened predominantly US-based funders and institutional leaders in Washington, DC, to discuss researcher incentives and research assessment in the context of open science publishing. Greater recognition for measured risk-taking, and for research that does not produce the expected outcome, was an important theme (https://doi.org/10.6084/m9.figshare.32287383)
One potential indicator is whether methodologically sound research and research outputs are shared regardless of the outcome. Making negative, positive or no-difference results available can help reduce unnecessary duplication and address bias in a research literature that tends to favor positive findings.

This is particularly important in fields such as healthcare research, where an incomplete evidence base can have wider consequences.
Despite the range of venues now available (including but not limited to journals such as PLOS One) for sharing research outputs, negative or unsuccessful results remain much less visible. This suggests that infrastructure alone is not enough: incentives, resources and research culture also influence whether these findings are shared.
NIH could help address this by recognizing rigorous research regardless of outcome within its normal assessment and funding processes. This could include recognizing a wider range of outputs, such as preprints and shared data, and preregistered research articles alongside journal articles when assessing researchers and funded work.

7. Public Impact: PLOS supports NIH's interest in better understanding and recognizing the broader public impact of biomedical research. Some impacts, particularly economic and societal impacts, may take many years to emerge and can be difficult to attribute directly to an individual research output or project.
There is, however, evidence that open science practices can contribute to wider economic and societal value, particularly when research and research resources are made available for others to access and reuse.
PLOS recently commissioned an independent economic analysis of open science. The analysis found evidence that open science creates economic value, including through the reuse of research outputs and resources (https://zenodo.org/records/20669372).
This suggests there may be value in considering carefully selected leading indicators of longer-term public impact, including open access to research and the sharing and reuse of data and code.
Other recent research also found an association between open access and wider use of research in innovation, measured through citations to research in patents (https://link.springer.com/article/10.1007/s11192-026-05725-8).

8. Other Areas: PLOS strongly supports NIH's effort to develop a broader and more meaningful approach to measuring scientific impact, including research practices, contributions and outputs beyond the journal article.
We recommend that NIH take a values- and principles-led approach when deciding which indicators to develop and use. Established frameworks, including the INORMS SCOPE Framework (https://inorms.net/scope-framework-for-research-evaluation/) and the Principles of Open Science Monitoring (https://doi.org/10.5281/zenodo.15807480) provide useful guidance on how to approach metrics and indicator selection and development.
A central principle of these approaches is to start by identifying what is valuable and important to understand or assess, rather than beginning with the data or metrics that are already available. This helps avoid measurement being shaped by what is easiest to count, rather than what is most relevant in a specific context.
PLOS took a similar approach when developing its Open Science Indicators, beginning with a set of solution-agnostic principles (https://doi.org/10.6084/m9.figshare.21640889). Responsible use was an important part of that work. New indicators should provide additional information to support qualitative and contextual assessment, rather than be used to create uncontextualized comparisons, rankings or league tables.
PLOS research (https://doi.org/10.7717/peerj.20502) has also shown that, where better indicators are

not available, evaluators may rely on proxies such as Journal Impact Factor or personal and institutional reputation when assessing research quality and rigor.

For this reason, new indicators should be tested in the settings in which they are intended to be used, and developed with the researchers, funders, institutions and others who will rely on them. This will help establish whether they provide better information for assessment and identify unintended consequences before they become embedded in practice.

NIH may also wish to draw on the wider body of work already underway in research assessment and open science measurement, including CoARA, HELIOS, DORA, the Open Science Impact Indicators Handbook (https://handbook.pathos-project.eu/) the Open and Universal Science Project (https://opusproject.eu/) and the Science and Technology Innovation Indicators community. Indeed, while the focus of our responses have been on work PLOS has led, PLOS supports and has collaborated with numerous other initiatives to which are relevant to this RFI. These include but are not limited to: TIER2 (https://tier2-project.eu/), iRISE (https://irise-project.eu/), and OSIRIS (https://osiris4r.eu/) which have all produced outputs that will help multiple stakeholders further their thinking about measuring and improving reproducibility of published research.

193. Washington University in St. Louis

I am responding: On behalf of an organization

Name:

Name of Organization: Washington University in St. Louis

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Recommendations:
• Recognize and reward independent external replication of findings across datasets, sites, scanners, and populations, not simply publication of novel findings.
• Create dedicated funding mechanisms for replication studies so they can be evaluated on their own terms rather than competing directly with novelty-driven research. NIH should also consider targeted mechanisms to support independent replication of contested findings or coordinated studies involving investigators with competing interpretations to rigorously test areas of scientific disagreement.
• Create durable publication and recognition pathways for rigorous replication studies, including studies that confirm, fail to confirm, or refine prior findings. Funding replication without ensuring that high-quality replication work can be disseminated and credited risks creating a second-tier research pathway that investigators, particularly early-career investigators, cannot afford to pursue.
• Promote publication of study protocols and registered reports.
• Invest in shared reproducibility infrastructure, including validated reference materials, benchmark datasets, standardized protocols, and training in rigorous study design, statistical methods, bias mitigation, and data stewardship. NIH should fund this infrastructure directly rather than shifting the cost of new reproducibility expectations to individual investigators or institutions.
• Refine the existing Contributions to Science section of the Biosketch: keep the cap of 4-5 “Contributions to Science” but explicitly allow and provide a clear rubric for narrative examples of open-source tool adoption, replication study leadership or protocol adoption, or start-up/tech transfer milestones.
• Suggested measurable indicators: % of studies with openly registered protocols; number of independent validation studies; number of studies adopting methodology established by NIH-funded investigators; compliance with ClinicalTrials.gov reporting deadlines. For measures such as adoption or reuse, NIH should first establish reliable methods for attribution and validation so that citations, downloads, or other easily captured activity are not automatically treated as evidence of meaningful scientific impact.

Potential unintended consequences:
• Replication studies require additional time, funding, and access to independent datasets, which may disadvantage early-career investigators or research areas where external validation cohorts are limited. For early-career investigators in particular, sustained investment in replication work may also create career risk under promotion and tenure systems that continue to emphasize novel findings and

traditional measures of productivity.
• Overemphasis on reproducibility metrics could discourage trainees and investigators from pursuing innovative, high-risk research in favor of safer projects that are more easily replicated.
• Journals are typically uninterested in a study that simply confirms a prior, well-conducted finding, so any incentive system NIH builds will need to address where this work gets published, not just how it gets funded.

Washington University has previously developed a more detailed set of recommendations for strengthening rigor and reproducibility in federally funded biomedical research. Our attached white paper, Reproducibility in Practice: Washington University's Perspective on Strengthening Federally Funded Biomedical Research, provides additional recommendations related to replication studies, reproducibility infrastructure, training, standardized research tools and resources, data sharing, and academic-industry collaboration.

2. Data, Software, and Model Sharing: Recommendations:
• Formally credit non-traditional research products such as datasets, software, standard operating procedures so that they are citable scholarly outputs.
• Reward open-source hardware/software, custom instrumentation, protocols, and curated baseline datasets that enable other labs' discoveries.
• Integration with Existing Data Repositories (ORCID & Crossref): expand the capability of SciENcv and the eRA Commons to pull verified digital object identifiers (DOIs) for non-traditional outputs automatically. If an investigator uploads a validated baseline dataset to a public repository, open sources a software package on GitHub, or registers a patent, these repositories issue distinct DOIs. Allowing these verified DOIs to auto-populate into standard NIH biosketches and progress reports (RPPRs) treats open assets with the exact same administrative simplicity as a traditional PubMed ID.
• Suggested indicators of sharing and reuse: repository downloads, material transfer requests, documented reuse of data or methods by independent labs, adoption of a tool or protocol by other institutions or journals. This is one area that is best suited to a fully automated measurement using repository/registry analytics or expanding iCite-type analytics rather than requiring new ones. Where feasible, NIH should favor verifiable indicators of sharing, such as deposition, functional availability, and documented reuse, rather than attempting to assign a subjective quality score to shared products.
• BioLINCC at NHLBI already shows what this can look like: a shared, freely accessible repository that lowers the barrier to depositing and reusing data. Building out more infrastructure like this, rather than adding new reporting requirements, would likely do more to encourage sharing.

Potential unintended consequences:
• Measures such as downloads, citations, or repository activity can be easy to capture but may not reliably reflect meaningful scientific reuse or impact and could be susceptible to gaming.
• Expectations for open sharing must account for legitimate limitations involving privacy, consent, intellectual property, proprietary information, data-use agreements, and the costs of long-term data or software maintenance.
• NIH should avoid creating incentives that disproportionately favor fields where digital outputs are easier to share and quantify.

3. Training and Mentorship: Recommendations:
· Track trainee career trajectories across sectors (academia, industry, government) as one contextual indicator of mentoring impact, rather than relying on co-authored paper counts. Career placement or advancement should not be used as a standalone measure of mentorship quality.
· Multi-year tracking of mentee career outcomes, not just a single snapshot at first job placement, built by reusing existing NIH data systems rather than creating new survey burden for training grant administrators.
· Fund mechanisms that encourage and support co-mentorship across disciplines or with non-academic partners to enhance innovation and translation.
· Consider structured, confidential trainee feedback as one qualitative input to mentorship assessment. Such feedback may provide a more direct measure of mentoring quality than downstream career outcomes alone.
· Track mentorship as a protected, distinct effort line, credited separately from research percent-effort, so faculty who invest heavily in mentoring aren't penalized against peers who don't in productivity-based comparisons. Allow mentees to identify who actually mentored them rather than assuming the PI-of-record captures it. Outside lab-based fields especially, people are often mentored by several individuals or informal networks, and any tracking system should be built to reflect that.

Potential unintended consequences:
• Mentorship outcomes often emerge over many years and are influenced by factors beyond an individual mentor's control. Career placement or advancement should therefore not be treated as a simple measure of mentorship quality. Outcome-based measures could also create disincentives to mentor trainees whose career paths are less predictable or less readily captured by conventional placement metrics.
• Any mentorship metric should recognize team and informal mentoring relationships and avoid rewarding nominal mentorship arrangements over substantive contributions.
• New mentorship reporting requirements should rely on existing or automated data wherever possible rather than creating additional tracking responsibilities for investigators, trainees, or institutions.

4. Collaboration: Recommendations:
· Early career scientists can benefit greatly from being on interdisciplinary teams to kick start their NIH careers. But, they are discouraged from being MPIs and gaining credit on these teams because they lose their ESI status. A new category for recognizing team members without losing ESI status would encourage more leadership from early career scientists in team proposals.
· Recognize a role and give credit for people whose technical or methodological expertise makes a project possible, but who are neither PI nor first author.
· Suggested indicators: Use contributor-role data to identify "hub" team members whose expertise is broadly relied upon and formally recognize this role in review and award criteria. Contributor-role statements can provide useful context for expert review, but NIH should avoid converting individual contribution categories into a rigid point system. For multi-site trials, use site-level trial performance metrics.
· The state of the art for identifying who actually contributed to a team, and how much, is still weak. NIH should fund research into team science measurement itself, especially for large, multidisciplinary teams outside basic science, rather than assuming existing tools transfer cleanly. NIH should also add a section in NOFO that requires the role for each contributor to be defined.

· Team science should be measured at more than one level. A team's output is more than the sum of individual contributors and must be measured at multiple nested levels (individual, team, institution) simultaneously.
· New measures should also distinguish individual contribution from team-, institutional-, and infrastructure-level impact. NIH should avoid assigning outcomes generated collectively by teams, cores, networks, or institutions to a single investigator simply because that individual is the PI of record.

Potential unintended consequences:
• Rewarding "hub" status also creates a gaming risk: contributor lists could become inflated, with credit flowing to well-known names rather than the people who actually did the work. Any new collaboration metric needs a safeguard against this before it's tied to review or funding decisions.
• Contributor roles may change over the life of a project, and formal titles may not capture the relative significance of individual contributions. Any system used to assess team-science contributions should allow roles to evolve and should avoid creating rigid hierarchies among different types of scientific contribution.
• New approaches to recognizing team science should not inadvertently disadvantage investigators who work in fields or research settings where smaller teams or individual-led research remain appropriate. Collaboration should be recognized when it adds scientific value, but should not become an expectation or proxy for impact in every field.

5. Entrepreneurship and Translation: Recommendations:
· Early-stage translational work (patents, licensing, regulatory-compliant SOPs, startup formation) is under-recognized because investigators can appear unproductive by traditional metrics during these long, high-effort phases.
· Recognize early-stage tech transfer milestones such as invention disclosures, licensing agreements, SBIR/STTR transitions.
· Track implementation/scale-up work and community partnerships as translational impact in public-health and clinical settings.
· Support more funding mechanisms that bridge basic discovery to clinical trial in one program.

Potential unintended consequences:
• For basic biomedical research, entrepreneurship and commercialization should be recognized when relevant but should not become a required dimension of individual scientific impact. Making patents, licensing, startup formation, or other commercial outcomes an expected marker of success could redirect investigators toward patentable outputs rather than the most important scientific questions.
• Entrepreneurship and commercialization opportunities vary substantially across disciplines and types of research. Metrics such as patents, licenses, startup formation, or industry partnerships should therefore supplement rather than replace other measures of scientific contribution.
• Overweighting commercial outcomes could disadvantage foundational, population-based, implementation, or other research whose primary value is scientific or public rather than commercial.
• Translation often occurs over long-time horizons and involves contributions from multiple investigators and organizations, making attribution to an individual investigator difficult.

6. Foundational Scientific Exploration: Recommendations:
· Recognize the development of new methods and research tools, and the adaptation of concepts from

other fields (e.g., applying physics, mathematics, or engineering approaches to new biomedical problems), as foundational contributions. These tools and approaches often enable many downstream discoveries but are rarely rewarded as heavily as the discoveries they make possible.
· Explicitly define “success” to include informative negative or null results, since the value of high-risk work lies partly in ruling out unproductive directions.
· Expand R21-type funding for high-risk ideas ahead of R01 submission and consider renewal pathways for demonstrated foundational progress even without traditional productivity metrics, creating an R21-to-R01 pipeline.
· Consider longer-duration, flexible funding mechanisms for foundational and high-risk research that allow investigators to change direction as evidence evolves. Renewal evaluation should emphasize the significance of the scientific question and the rigor with which it was pursued rather than milestone completion alone.
· Support more R35 and DP2 activity code funding within each institute to promote creative thinking and exploration.

Potential unintended consequences:
• Any new reward system should be explicitly evaluated for whether it disincentivizes foundational, high-risk research. It is critical that NIH not overweight near-term or traditionally measured impact, as this could starve the unglamorous validation, optimization, and replication work needed to refine foundational advances into reproducible tools and protocols.
• If “high-risk” or “paradigm-shifting” becomes a checkbox metric, investigators may manufacture synthetic risk or overhype standard, incremental work to satisfy the criteria.
• Well-funded, established labs can absorb the failure of a high-risk project through a portfolio of safer grants, but for an early-stage investigator a project that yields no publishable data over three to five years can be career-ending under current tenure and promotion frameworks.

7. Public Impact: Recommendations:
• Automated Cross-Disciplinary Citation and Funding Analytics: Instead of requiring investigators to manually track and justify their reach, the NIH should utilize automated bibliographic tools (akin to the existing NIH iCite platform) to provide contextual indicators of a grant’s reach and downstream influence. The platform could automatically flag cross-disciplinary breadth (e.g., the diversity of journal subject categories citing the work) and downstream proposal catalysis (e.g., tracking when an NIH grant number or its resulting publication is cited within the Significance or Innovation sections of subsequent, independent applications).
• Recognize completion of clinical trials and downstream use of results to inform clinical practice, including incorporation into clinical guidelines, as a core public-impact outcome. Reward site-level contributions in multicenter trials. Site PIs provide essential local leadership but are rarely top-listed authors and go unrecognized. NIH should reward participating sites and investigators commensurate with their contributions and track site-level performance (accrual and data-submission commitments) across studies.
• Washington University can point to an existing example of this kind of framework in the Translational Science Benefits Model (https://www.translationalsciencebenefits.wustl.edu/), which already assesses public impact across a research portfolio.

Potential unintended consequences:

• “Public impact” is ill-defined. Important research may never directly change clinical practice, and null findings can themselves represent meaningful scientific contributions. Public impact also differs substantially across basic science, clinical research, and public health and may emerge years or even decades after the original discovery. NIH should therefore avoid measurement approaches that systematically favor outcomes that are visible within a single grant cycle or that privilege near-term translational impact over foundational research.
• Attribution of public impact is inherently difficult because changes in clinical practice, health outcomes, policy, or economic activity typically result from cumulative contributions across many studies, investigators, institutions, and sectors. NIH should be cautious about assigning downstream societal outcomes to individual investigators or awards.
• Measures of public impact should not depend primarily on visibility, media attention, public engagement, or other indicators that may favor certain topics or investigators without demonstrating greater scientific or health impact. Measures should emphasize demonstrable contribution rather than prominence or reach alone.

8. Other Areas: · Washington University supports broadening the ways NIH recognizes scientific contribution but strongly cautions against replacing imperfect traditional metrics with a new layer of unvalidated metrics. No single score, automated index, or checklist can capture scientific impact across disciplines, career stages, research types, and time horizons. New indicators should supplement, not substitute for, expert scientific judgment and should be used only for purposes for which they have been validated.
· Substitution instead of addition: replacing traditional metrics rather than supplementing them would penalize excellent small, focused labs that do not do team science or that lack the capacity to pursue innovative discovery research and replication activities simultaneously. New measures should broaden the ways scientific contribution can be recognized without creating an expectation that every investigator demonstrates impact across every dimension.
· Avoid a single composite measure of scientific impact: NIH should resist collapsing diverse dimensions of scientific impact into a single score or ranking. Different forms of impact are appropriate to different disciplines, career stages, research objectives, and time horizons. A multidimensional framework would preserve this context and reduce incentives for investigators to optimize research activity toward a narrow set of measurable outcomes.
· Prioritize scientific quality over publication quantity or venue prestige: NIH should avoid relying on publication counts, journal prestige, or raw citation measures as proxies for scientific impact. Assessment should place greater emphasis on the rigor and reproducibility of the work, methodological contributions, the significance of the scientific questions addressed, and the extent to which the work advances the field.
· Distinguish measurement from decision-making: NIH should distinguish between indicators used for portfolio-level evaluation, information provided as context during peer review, and measures used directly in funding decisions. Evidence that an indicator is informative for understanding NIH’s overall research portfolio does not establish that it is valid or appropriate for evaluating an individual investigator or application.
· Recognize meaningful scientific peer review: Sustained, rigorous peer review is an important contribution to the quality and integrity of the scientific enterprise. NIH should consider approaches for recognizing substantive peer-review service without incentivizing review volume over review quality.
· Protect expert review and reviewer capacity: Broader measures of scientific impact should not simply

add new scoring categories to an already complex peer-review process. Meaningful qualitative assessment requires reviewers with appropriate expertise and sufficient time to engage deeply with an investigator's contributions. NIH should consider the effect of any new assessment framework on reviewer workload and the quality of peer review.
· Validate before implementation: Before incorporating new impact measures into peer review or funding decisions, NIH should first determine that new tracking is necessary and pilot proposed measures. The ability to automate a measure does not establish its validity or usefulness. New measures should be evaluated for reliability, validity, administrative burden, and differential effects across disciplines, career stages, institution types, and research approaches. Measures that do not provide meaningful information beyond existing indicators should be revised or abandoned.
· Impact on junior investigators: New approaches may unfairly affect junior researchers given that they have less time to demonstrate their impact. They may also alter their research to over-conform or anticipate the new approaches since they have more concern regarding job stability. University promotion and tenure rubrics lag NIH incentives, and NIH should coordinate with institutions on this.
· Chilling effect on high-risk work: putting too much weight on reproducibility or near-term translation could push researchers toward safer, more easily replicated projects.
· Adopting impact measures that are difficult to document in a consistent and verifiable manner could lead to "storytelling" and inflated narratives. Broadening impact measures must not result in "metric inflation" that forces investigators and staff to act as public relations agents or administrative data entry clerks.
· Adding administrative burden: new categories of impact could add significant administrative burden and detract investigators from time conducting research. New measures must be automated as opposed to self-reported.
· Consider institutional resource differences: New reporting or infrastructure requirements that are not accompanied by NIH-supported resources may disproportionately burden smaller and less-resourced institutions. NIH should provide shared infrastructure and, where appropriate, financial support for new requirements so that changes intended to broaden the recognition of scientific impact do not instead widen disparities among research institutions.
· Move deliberately: Consult experts across disciplines and research settings, assess unintended consequences, and phase in any new approaches only after testing and evaluation.
· Normalize metrics across disciplines and subfields. Publication and citation rates vary widely (e.g., hardware-intensive versus computational fields), and only the "hottest" subfields attract high citations. NIH should normalize to granular subfields (e.g., by expanding Relative Citation Ratio) and exclude review and survey articles from citation counts, recognizing them separately as community service.
· Preserve scientific and apolitical review. Impact must be assessed impartially using scientifically grounded criteria, appropriate evidence, and expert judgment, free of political agendas. Scientific review should be conducted by scientists rather than driven by political appointees.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/WashU_Perspective_Reproducibility-in-Research_FINAL.docx

Description: Washington University has previously developed a more detailed set of recommendations for strengthening rigor and reproducibility in federally funded biomedical research. Our attached white paper, Reproducibility in Practice: Washington University's Perspective on Strengthening Federally Funded Biomedical Research, provides additional recommendations related to replication studies,

reproducibility infrastructure, training, standardized research tools and resources, data sharing, and academic-industry collaboration.

Title:
"Reproducibility in Practice: Washington University's Perspective on Strengthening Federally Funded Biomedical Research"

Prepared by: Mark Lowe, Bill Powderly, Sandro Galea, Charles Goss, Nichole Moore, and Doug E. Frantz
Washington University in St. Louis
June 2025

Executive Summary

Reproducibility is a foundational principle of scientific integrity, yet it remains a persistent challenge across all scientific disciplines. Despite major advances in technology and discovery, studies estimate that a significant portion of federally funded biomedical research is difficult or impossible to replicate—undermining public trust and slowing the translation of findings into clinical impact.

Washington University in St. Louis (WashU) has a strong track-record and national reputation of conducting NIH-funded research that meets the highest standards of rigor and reproducibility. This white paper outlines our perspective of the top contributors to irreproducibility and presents actionable recommendations for how federal funding agencies—especially the NIH—can strengthen the reproducibility of U.S. biomedical research through targeted investments, cross-agency collaboration, and engagement with industry partners.

Key Challenges to Reproducibility of Scientific Findings

1. Systemic Challenges in the Scientific Ecosystem

Reproducibility in science, as emphasized by the Office of Science and Technology Policy (OSTP), reflects the ability of independent researchers to test a hypothesis using diverse methods and consistently reach results that validate or challenge the original findings—ensuring scientific conclusions are transparent, trustworthy, and broadly generalizable. However, incentive structures across academia and funding agencies often prioritize novelty over validation, speed over rigor, and novel approaches over reproducible methods. These pressures discourage replication studies and shortchange robust experimental design, statistical power, and transparent reporting. Further, tight timelines and limited budgets for early-stage research often lead to underpowered studies and insufficient preclinical validation. Additionally, there may be substantial cognitive bias where certain hypotheses become accepted wisdom and studies challenging these 'agreed' conclusions are less likely to be supported.

2. Lack of Formal Training in Rigor and Reproducibility

Despite widespread acknowledgment of these challenges, structured training in scientific rigor and reproducibility remains inconsistent across institutions. Many trainees—and even faculty—lack foundational instruction in areas such as experimental design, bias mitigation, data stewardship, and statistical reproducibility. This educational gap perpetuates poor practices and limits the adoption of emerging standards.

3. Limited Access to Standards, Tools, and Sharable Resources

Many investigators lack access to well-characterized research tools, standardized protocols, and high-quality reagents. The absence of universally accepted benchmarks for materials (e.g., cell lines,

chemical probes), analytical methods, and computational workflows increases variability and undermines reproducibility. Similarly, the absence of standardization of datasets that can be used to replicate results hampers reproducibility. Legal, logistical, and cultural barriers often prevent the timely sharing of raw data, code, and research samples—especially across institutional or national boundaries.

Recommendations for Federal Action

To meaningfully address the reproducibility crisis, federal agencies should invest in a coordinated, multi-faceted response. Washington University recommends the following actions aligned with the three challenges noted above:

1. Overcoming challenges in the scientific ecosystem:

- **Create reproducibility pilot programs** within NIH translational initiatives (e.g., NCATS, RADx, BRAIN Initiative, HEAL Initiative, NSF-GOALI program) to support independent replication studies and develop standard operating procedures (SOPs) before downstream translation and commercialization.
- **Align grant review criteria and funding decisions more closely with translational readiness**, requiring detailed reproducibility plans and encouraging applicants to integrate cross-sector collaborators early in the research lifecycle to validate data generated.
- Develop mechanisms to allow investigators to propose novel hypotheses that run counter to accepted thought.
- **Fund infrastructure for standardization and benchmarking**, including common reagents, reference and tool compounds, extrinsic factors and uncertainty quantification studies for biomedical research (similar to PAR-24-167: Utilizing Equipment to Study Environmental Extrinsic Factors and Enhance Rigor and Reproducibility of Animal Research (R24, Clinical Trials Not-Allowed) FAIR-compliant data repositories, and instruments to facilitate data sharing, that serve both academic and private-sector users.
- **Establish institutional GLP laboratories/core facilities through strategic investments** to lower the barrier for academic labs to facilitate the reliability, reproducibility, and traceability of biomedical research under GLP-compliant protocols.
- **Seed industry-academic partnerships**, possibly through new RFAs. cost-share mechanisms, and faculty sabbatical programs where academic findings are validated in industrial settings under GLP/GMP-like rigor especially those involving assay development, biomarker validation, or therapeutic discovery.

2. Training in rigor and reproducibility

- **Develop training resources through** NIH, NIST, FDA, DOE, and other agencies that can create opportunities for training programs at graduate and post-graduate level that change the scientific culture and convey the nine core tenets of Gold Standard Science as outlined by the OSTP. Resources should include training in statistical methods, study design, and the importance of casual inference frameworks in the analysis of biomedical datasets.
- **Establish and expand reproducibility-focused consortia** that unite federally funded researchers with industry scientists to cocreate educational and developmental opportunities centered around shared protocols, reference standards, and validation datasets.

3. Robust sharing of data, materials, software and other research tools

- **Advance Interagency Collaboration through the co-development and dissemination of standardized reference materials** and protocols (e.g., authenticated cell lines, benchmark datasets, small molecule probe compounds) through centralized national repositories that undergo routine QA/QC validation throughout their lifecycle.
- **Invest in FAIR-compliant data repositories, sample-sharing platforms, and adoption of electronic notebook documentation** that incentivize and lower the barrier to open access to raw data, code, thorough descriptions of methods or protocols, and research samples.
- **Encourage the development of application protocol interfaces (APIs)** and cloud-based tools that allow for real-time reproducibility checks, automated protocol validation, and collaborative data analysis.

Conclusion: A Call to Action

Ensuring the reproducibility of biomedical research is essential to maintaining the integrity of the scientific enterprise and the impact of federal investment. Addressing this issue requires not only institutional reforms, but bold, coordinated action by federal agencies and their partners. Washington University in St. Louis stands ready to work with OSTP, OMB, the NIH and other federal stakeholders, as well as private-sector collaborators, to lead this cultural and structural transformation.

194. Society for Academic Emergency Medicine (SAEM)

I am responding: On behalf of an organization

Name: Melissa McMillian

Name of Organization: Society for Academic Emergency Medicine (SAEM)

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Senior Director

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Emergency medicine research operates in a uniquely challenging environment for reproducibility. Our studies are conducted in high-acuity, time-constrained settings with heterogeneous patient populations, variable institutional protocols, and outcomes that are inherently difficult to standardize across sites. NIH should recognize that reproducibility in emergency medicine often requires multi-site pragmatic trials rather than single-site replication studies, and that these studies require dedicated multicenter infrastructure that conventional measures of scientific productivity may not capture.. Incentives for reproducibility should specifically support emergency care networks and collaborative research consortia such as PECARN and SIREN, which are specifically designed to enable reproducible multi-site emergency medicine research at scale. Metrics of rigor should reward pre-registration of clinical trials, prospective data collection, and transparent reporting of null results. These are practices that emergency medicine investigators already value, although they are not currently recognized in grant scoring or career advancement decisions.

2. Data, Software, and Model Sharing: Emergency department data, including electronic health records, triage timestamps, medication administration records, and disposition data, represents an underutilized national resource for health services research, comparative effectiveness research, and health equity science. NIH should develop incentive structures that reward emergency medicine investigators for sharing de-identified ED data, developing and releasing open-source clinical decision support tools, and contributing to national registries. Currently, the administrative and legal burden of data sharing falls disproportionately on individual investigators at community-based academic medical centers who lack the dedicated informatics and legal infrastructure of large research universities. This barrier disproportionately disadvantages early-career EM investigators and those at smaller or community institutions. These smaller and community institutions represent the majority of emergency care that is delivered in the US. Ideally, the NIH should fund the infrastructure necessary to make data sharing feasible, not merely require it as a policy mandate without allocating resource support.

3. Training and Mentorship: Emergency medicine research training operates through a narrow and fragile pipeline. While the specialty has produced a generation of physician-scientists through SAEM Foundation grants, emergency care K to R award pathways, and ACEP research funding mechanisms, these are limited relative to the clinical and public-health needs that emergency medicine research is positioned to address. The NIH should develop metrics that explicitly reward the mentorship of

emergency medicine trainees into research careers, including recognition of successful K award sponsorship, research fellowship direction, and the production of independently funded physician-scientists in clinical departments. Currently, faculty who invest substantial effort in research mentorship at community-based academic hospitals receive no credit in grant scoring, promotion criteria, or NIH impact metrics, contributing to a structural disincentive that undermines the long-term sustainability of the emergency medicine research workforce.

4. Collaboration: Emergency medicine is fundamentally an interprofessional and collaborative specialty. Without an Institute of our own, emergency medicine research increasingly spans multiple NIH Institutes and disciplines like pharmacology, toxicology, simulation science, health systems, global health, multiple body systems, and implementation science. Nonetheless, current NIH metrics heavily reward the single-PI model and do not adequately recognize contributions to multi-PI and team-based science. For example, an emergency medicine investigator who serves as site PI for a 40-site national trial, leads patient enrollment, and generates a substantial proportion of the study's data may receive limited recognition because conventional metrics emphasize lead authorship and overall PI status rather than the scientific and operational contributions required to execute a rigorous multisite study. The NIH should develop metrics that recognize the value of clinical trial execution, site PI leadership, and collaborative contributions to large-scale team science. These are the contributions that emergency medicine investigators make routinely and that are essential to the field's evidence base.

5. Entrepreneurship and Translation: Emergency medicine is uniquely positioned to translate scientific discovery into practical innovations because investigators work at the interface of acute clinical need, health systems, technology, and implementation. Emergency physicians routinely identify unmet needs that lead to new clinical decision-support tools, diagnostic technologies, medical devices, digital health platforms, and novel models of care. However, traditional measures of scientific impact often undervalue contributions across the translational lifecycle, including validation, regulatory development, commercialization, workflow integration, and implementation widely across health systems. The NIH should recognize entrepreneurship and successful translation as legitimate forms of scientific impact and apply stage-appropriate measures, including technical validation, regulatory and clinical milestones, adoption, reach, affordability, sustainability, and patient outcomes. Patents, licenses, startups, and private investment can be useful indicators but should not be treated as ends in themselves. The NIH should also recognize translation through clinical protocols, open tools, nonproprietary interventions, and implementation in community, rural, and resource-constrained settings. Such metrics would encourage investigators not only to produce new knowledge, but also to ensure that discoveries are successfully translated into scalable solutions with meaningful public health impact.

6. Foundational Scientific Exploration: Emergency medicine plays a critical role in identifying clinical phenomena that generate basic science hypotheses, ranging from the pathophysiology of sepsis-induced immunosuppression to the cardiovascular toxicology of emerging drug adulterants. The NIH should preserve robust support for foundational discovery while also recognizing the role of hypothesis-generating observational research and pilot investigations in inspiring questions for subsequent basic, translational, and clinical investigation. Stage-appropriate metrics should recognize bedside observation and characterization as early contributions within this continuum, while evaluating later foundational and translational work according to the knowledge, mechanisms, and capabilities it creates.

7. Public Impact: Emergency medicine generates scientific value that is immediate, population-level, and directly translatable to patient care, but this value is poorly captured by citation metrics and journal impact factors. A clinical decision rule that reduces unnecessary CT scanning in children presenting with head trauma, a protocol that improves sepsis recognition in triage nurses, or a simulation curriculum that reduces procedural complications in residents represents enormous public health impact that may not generate correlative high citation counts. To best improve the health of the entire population, NIH should develop metrics that explicitly value clinical practice change, guideline adoption, reduction in healthcare utilization, patient safety improvement, and health equity outcomes as measures of scientific impact.

Recognizing these outcomes should complement, not displace, support for foundational research whose eventual applications may be uncertain or decades away. NIH needs both sustained discovery and effective pathways for mature evidence to reach patients.

8. Other Areas: Recommendations:

NIH should sustain a balanced portfolio encompassing foundational, translational, clinical, and implementation research. The goal is not to privilege one stage over another, but to apply stage-appropriate measures that recognize the distinct contributions each makes to scientific progress and public health.
NIH should develop a diversified portfolio of impact metrics that includes: clinical practice change and guideline adoption; patient-centered outcomes beyond mortality; contributions to open-access data and tools; mentorship productivity measured by trainee career outcomes; and equitable recognition of collaborative and site-level scientific contributions. These metrics should be applied with sensitivity to career stage, institution type, specialty and research stage, without misattributing institutional resources or a particular type of scientific output for investigator impact.
Emergency medicine's contribution to the NIH mission is substantial, immediate, and population-level. The metrics that NIH adopts to measure and reward scientific impact will ultimately determine whether the next generation of emergency medicine investigators can build sustainable research careers. We hope that the NIH will use this opportunity to design impact metrics that finally recognize the full breadth of scientific value that clinical investigators in underrepresented specialties generate.

195. Washington State University

I am responding: On behalf of an organization

Name: Kim Christen, PhD

Name of Organization: Washington State University

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Washington State University believes rigor and reproducibility should be recognized as fundamental scientific contributions rather than solely as compliance requirements. For biomedical research assessments, NIH could consider adapting industry principles such as the methodologies of Gage Repeatability and Reproducibility, which evaluate measurement consistency and identify sources of variation. Potential biomedical science indicators could include the percentage of publications with publicly available protocols, analysis plans, data, and code; use of preregistration, blinded analyses, validation cohorts, and replication studies; documentation of assay performance and measurement variation; and dissemination of well-designed studies that yield positive, negative, or null findings. Importantly, null findings are a valuable scientific outcome that can refine hypotheses, prevent duplication of effort, and strengthen the overall evidence base and should therefore be recognized as meaningful contributions to scientific progress.

NIH should also recognize contributions to causal research, defined as research that improves understanding of whether and how specific mechanisms, interventions, exposures, programs, practices, or policies influence health-related outcomes. Because causal questions arise across the biomedical research continuum, measures of causal research should be appropriate to disciplinary context and study design. Indicators could include the use of rigorous causal inference approaches, such as randomized controlled trials, natural experiments, quasi-experimental designs, instrumental variable analyses, difference-in-differences methods, regression discontinuity designs, propensity score methods, mediation analyses, longitudinal studies, mechanistic investigations, and the use of causal frameworks. Additional measures could include replication of causal findings across populations or settings and demonstrated influence on clinical practice, public health interventions, or policy decisions.

NIH could support dedicated replication funding opportunities, establish reproducibility networks, and incorporate rigor, transparency, and causal inference considerations into progress reporting and review criteria. While quantitative indicators can provide useful evidence of rigor and impact, excessive reliance on metrics may increase administrative burden for investigators and institutions, incentivize reporting activities over scientific quality, and may overlook meaningful scientific contributions that are not easily quantified. To minimize burden, NIH should prioritize measures that can be derived from existing grant reports, publications, and public data repositories rather than creating new reporting requirements. Quantitative indicators should be complemented by qualitative assessments, including investigator narratives describing methodological innovations, contributions to community standards, advances in

causal understanding, efforts to improve research transparency, and the translation of research findings into practice. Such an approach would better capture meaningful scientific advances while preserving flexibility across disciplines. Potential benefits include greater confidence in scientific findings, stronger evidence for decision-making, more efficient use of research resources, and enhanced public trust. However, excessive emphasis on replication, causal attribution, or resource-intensive validation requirements could discourage innovative or exploratory work and disproportionately burden early-career investigators and smaller institutions with limited administrative personnel and support infrastructure. Expectations should therefore be adapted to disciplinary norms and emphasize transparency, methodological rigor, sound causal reasoning, and meaningful scientific contribution rather than uniform replication requirements or metric-driven evaluation.

2. Data, Software, and Model Sharing: Washington State University supports expanded recognition of data, software, and model sharing as valuable scientific outputs that accelerate discovery, enhance transparency, improve reproducibility, and increase the return on public research investments. Broad access to high-quality datasets, software tools, and computational models enables researchers to validate findings, conduct secondary analyses, develop new hypotheses, leverage existing data for novel applications, and avoid unnecessary duplication of effort. Data sharing can also facilitate interdisciplinary collaboration, support workforce development, improve research efficiency, and accelerate the translation of scientific discoveries into clinical, public health, and community benefits. To realize these outcomes, NIH should continue investing in infrastructure and tools that make data collection, curation, storage, preservation, and sharing easier for researchers. Effective data-sharing policies depend on systems that reduce administrative burden while ensuring research outputs remain accessible, reusable, interoperable, and discoverable by the broader scientific community. Researchers are more likely to participate in data-sharing efforts when repositories, standards, and submission processes are user-friendly, sustainable, and supported through appropriate training, technical assistance, and dedicated funding for data stewardship activities.

The value of data sharing, however, depends not simply on whether data are deposited in a repository, but on the quality, usability, and reusability of the shared resources. Ensuring data quality is therefore essential to realizing the full benefits of data sharing. NIH should encourage the adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, community-developed data and metadata standards, comprehensive data dictionaries, and clear documentation that enable other researchers to understand and reuse shared datasets. Quality assurance could be further supported through repository-based validation checks, quality control procedures, version control systems, provenance tracking, and the use of certified repositories that provide quality review and long-term preservation. NIH could also encourage independent review of datasets during peer review or repository submission processes to ensure completeness, accuracy, and usability. Because high-quality data sharing requires specialized expertise, continued investment in professional data managers, curators, informatics support, and researcher training in data management and quality assurance practices is essential.

Measurable indicators could include deposition of FAIR-compliant datasets, compliance with NIH Data Management and Sharing requirements, use of persistent identifiers and interoperable metadata, dataset citations and downloads, software adoption metrics, evidence of successful dataset reuse, and demonstrations that shared resources have enabled new collaborations, publications, discoveries, or applications. Assessments should also consider indicators of data completeness, documentation quality,

adherence to community standards, reproducibility, and successful reuse by other researchers, rather than relying solely on measures of dataset volume or repository deposits. NIH could further encourage these activities by recognizing datasets, software tools, workflows, computational models, and curated data resources as distinct research products in grant applications, progress reports, and promotion processes; providing supplemental funding for data curation, maintenance, and stewardship; and recognizing investigators who develop widely used community resources. Community-governed data-sharing trust models may provide additional mechanisms for balancing collaboration, stewardship, accountability, and appropriate protections for sensitive data.

The benefits of these approaches include increased data reuse, accelerated validation of findings, more efficient use of research resources, stronger collaboration across institutions, improved reproducibility, and broader dissemination of scientific knowledge. Potential unintended consequences include incentives for quantity over quality, increased burdens associated with maintaining shared resources, and challenges related to privacy, consent, and the sharing of sensitive human participant data. Flexibility will be required to accommodate disciplinary differences and ensure that early-career researchers receive meaningful credit for creating and maintaining high-quality shared resources regardless of traditional publication volume. By emphasizing data quality, stewardship, and meaningful reuse alongside accessibility, NIH can encourage the sharing of well-curated, reusable datasets while minimizing incentives that prioritize quantity over scientific value.

3. Training and Mentorship: Washington State University recommends that training and mentorship be recognized as critical scientific contributions that support the long-term strength, sustainability, and diversity of the biomedical research workforce. Effective evaluation should include both leading indicators, such as mentoring engagement, completion of individual development plans, participation in professional development activities, and mentor training, and lagging indicators, including trainee retention, graduation rates, career advancement, workforce placement, research productivity, success in obtaining independent funding, contributions to publications and shared research resources, development of interdisciplinary collaborations, participation in team science, and translation of research into clinical, public health, community, or policy outcomes. Because many important outcomes of training and mentorship are not readily captured through quantitative metrics alone, NIH should also incorporate qualitative assessments. Particularly valuable measures could include confidential trainee and alumni feedback, mentoring effectiveness assessments, evidence of leadership development, case studies demonstrating career progression, examples of successful sponsorship and networking, contributions to creating inclusive and supportive research environments, and narratives describing how mentoring has influenced scientific development, career decisions, collaboration, and professional growth. Standardized and confidential mentoring assessments, including exit surveys, could provide additional evidence of mentoring quality while minimizing reporting burden. Benefits of recognizing training and mentorship contributions include stronger workforce development, improved trainee retention, enhanced research culture, and more sustainable scientific careers. However, overreliance on quantitative outcomes may disadvantage mentors working with populations facing structural barriers or in disciplines with longer training and career timelines. NIH should therefore account for career stage, institutional context, and disciplinary variation when evaluating mentoring impact and adopt a balanced approach that integrates both quantitative and qualitative evidence.

4. Collaboration: Washington State University encourages NIH to develop mechanisms that recognize meaningful contributions to collaborative, team-based, interdisciplinary, and community-engaged biomedical research rather than relying primarily on conventional authorship metrics. As many of today's complex biomedical challenges require expertise spanning multiple disciplines, sectors, and stakeholder groups, assessment frameworks should reflect the increasingly collaborative nature of scientific discovery. At the same time, the quality, depth, and integration of collaboration should be more highly valued than the number of collaborators involved, because increased team size does not necessarily equate to increased scientific impact.

Potential quantitative indicators include documented individual contributions through standardized contributor role reporting systems; participation in multi-investigator, multi-institutional, and cross-sector collaborations; the number and diversity of disciplines represented on research teams; cross-disciplinary publications; sustained partnerships; development of shared datasets, tools, methods, software, and research infrastructure; and evidence that interdisciplinary approaches have generated novel research directions, discoveries, or applications. Additional indicators could include achievement of team-specific project goals, establishment of new collaborations, and community participation in research activities.

Because the value of collaborative and interdisciplinary research is not fully captured by quantitative measures alone, NIH should also incorporate qualitative assessments. Useful indicators could include investigator and partner narratives describing how diverse expertise contributed to scientific advances, examples of successful integration of different methodologies or perspectives, evidence of collaboration across traditionally separate disciplines, stakeholder and community feedback, and case studies demonstrating how interdisciplinary approaches improved research relevance, quality, translation, or societal benefit. NIH should also recognize contributions that may not be adequately reflected through authorship or publication metrics, including the development of community partnerships, stakeholder engagement activities, clinical trial recruitment and retention efforts, shared research infrastructure, and other activities that enable successful team science. NIH could support these efforts by incorporating contributor statements into reporting processes and encouraging the use of collaborative impact portfolios that capture diverse forms of participation and scientific contribution.

Such approaches would better reflect the realities of modern biomedical research, promote interdisciplinary problem solving, strengthen community engagement, and recognize expertise that may otherwise be overlooked in traditional assessment systems. Potential benefits include increased innovation, stronger scientific rigor, accelerated translation of discoveries, and improved capacity to address complex biomedical and public health challenges. To minimize administrative burden, NIH should rely whenever possible on information already available through grant applications, publications, progress reports, and publicly available research outputs rather than creating new reporting requirements. Evaluation frameworks should also provide sufficient flexibility to account for disciplinary and institutional differences while maintaining transparency and accountability in the assessment of individual and team contributions.

5. Entrepreneurship and Translation: Washington State University encourages NIH to recognize innovation and entrepreneurship as important pathways through which biomedical research generates public benefit and advances the agency's mission of improving health. While publications and citations

remain important indicators of scholarly activity, they do not fully capture the translational, clinical, economic, and societal impacts of scientific discovery. NIH should therefore recognize a broad range of innovation-related outputs, including patents, licenses, startup companies, software tools, diagnostics, therapeutics, medical devices, implementation strategies, and other technologies that advance research or improve health outcomes. Potential indicators could include invention disclosures, patent and licensing activity, startup formation, participation in and outcomes from Small Business Innovation Research and Small Business Technology Transfer programs, follow-on investment, public-private partnerships, technology adoption, and demonstrated impacts on healthcare practice, public health, or economic development.

Assessment frameworks should also recognize key translational and regulatory milestones that mark progress along the pathway from discovery to application, including Investigational New Drug applications, Investigational Device Exemption filings, Federal Drug Administration submissions and approvals, advancement through clinical trial phases, development of clinical guidelines, and implementation of evidence-based interventions. Because innovation often follows a lengthy and non-linear trajectory, NIH should consider measures of long-term impact, such as sustained adoption of technologies or interventions, downstream improvements in health outcomes, continued commercial development, market penetration, follow-on scientific advances enabled by the innovation, and enduring contributions to research infrastructure, public health practice, and economic growth.

Economic indicators can provide valuable evidence of the broader return on public research investments and may include follow-on federal, nonprofit, or private-sector funding; licensing activity and revenue; startup company sustainability; job creation and workforce development; healthcare cost savings or efficiency gains; technology adoption; manufacturing activity; and economic activity generated within local or regional innovation ecosystems. However, economic measures should complement, rather than replace, assessments of scientific quality, innovation, and public health benefits.

To ensure fairness, NIH should evaluate innovation and entrepreneurial impact within the context of institutional resources, disciplinary norms, and career stage rather than applying uniform expectations across the biomedical research enterprise. Smaller and rural institutions may have more limited access to technology transfer infrastructure, commercialization support, industry partnerships, venture capital networks, and regulatory expertise, which can affect their ability to translate discoveries into traditional entrepreneurial outcomes despite producing highly impactful research. Similarly, pathways to innovation vary substantially across disciplines, with some fields generating patents, licenses, and commercial products, while others create impact through advances in scientific knowledge, clinical practice, public health, policy, or community outcomes. Early-career investigators may also have fewer opportunities to demonstrate long-term innovation impacts because translational and commercialization outcomes often require many years to mature. NIH should therefore assess innovation relative to opportunity and stage of development and recognize a broad range of translational, scientific, societal, and economic contributions. Such an approach would preserve incentives for innovation across diverse institutional settings, disciplines, and career stages.

Given the diverse pathways through which innovation creates value, quantitative indicators should be complemented by qualitative assessments, including case studies, investigator narratives, peer review,

and evidence of contributions to innovation ecosystems, entrepreneurial mentorship, workforce development, and cross-sector partnerships. A balanced approach would encourage translation and entrepreneurship while recognizing the many ways federally funded research produces lasting scientific, health, societal, and economic benefits.

6. Foundational Scientific Exploration: Washington State University strongly supports NIH's continued investment in foundational scientific exploration, including high-risk, high-reward research that may not generate immediate results but has the potential to produce transformative advances. NIH has long demonstrated its commitment to this area through programs such as the High-Risk, High-Reward Research Program, Pioneer Awards, New Innovator Awards, Transformative Research Awards, and other funding mechanisms designed to support bold, creative, and unconventional ideas. Because paradigm-shifting discoveries often emerge from ambitious scientific questions, novel approaches, and investigations whose outcomes are inherently uncertain, NIH should ensure that foundational and exploratory research remains a visible and valued component of the broader biomedical research enterprise.

Indicators of impact should extend beyond traditional measures of productivity and continue to emphasize novelty, originality, and scientific creativity. Additional indicators could include the development of new theories, methodologies, or enabling technologies; the establishment of entirely new areas of inquiry; follow-on research stimulated by exploratory projects; and expert assessments of scientific risk, innovation potential, and intellectual contribution. NIH should continue to employ review criteria specifically designed for exploratory research and maintain funding mechanisms that provide investigators with the flexibility to pursue ambitious and unconventional ideas, including milestone-based awards, pilot and proof-of-concept programs, and longer-term funding opportunities that recognize the extended timelines often required for transformative discoveries.

Because the impacts of foundational research frequently emerge over years or decades, evaluation frameworks should emphasize long-term influence rather than short-term productivity alone. Quantitative and qualitative measures could include the development of new scientific fields, theories, methods, datasets, software tools, or research platforms; follow-on funding and research activity enabled by an initial discovery; downstream clinical, public health, or technological applications; and expert assessments or case studies documenting how exploratory research influenced subsequent advances. NIH could further embed and sustain high-risk, high-reward research within the academic enterprise by encouraging institutional recognition of such contributions in promotion, tenure, and annual evaluation processes; supporting interdisciplinary collaborations that foster innovative thinking; and providing career-stage protections that allow investigators, particularly early-career researchers, to pursue ambitious research directions without undue concern over short-term publication or funding outcomes. NIH should also promote a research culture that recognizes well-executed studies with negative or inconclusive results as valuable scientific contributions when they advance understanding, refine hypotheses, or open new avenues for investigation.

The principal benefit of this approach is the promotion of creativity, innovation, and breakthrough discoveries that expand the boundaries of knowledge and create opportunities for future scientific, clinical, and societal advances. Potential challenges include the difficulty of measuring impact over shorter timescales and distinguishing thoughtful scientific risk-taking from poorly conceived projects.

Evaluation frameworks should therefore account for the extended timeline often associated with foundational discovery and ensure that investigators and institutions are appropriately recognized and rewarded for pursuing innovative research with the potential to generate substantial long-term scientific and public benefit.

7. Public Impact: Washington State University believes assessments of the public impact of biomedical research should focus on demonstrable contributions to health, society, and public well-being. While publications and citations provide evidence of scholarly productivity, they do not fully capture the many ways research benefits society. Biomedical research creates public value through improvements in health outcomes, clinical practice, healthcare delivery, economic growth, workforce development, public trust in science, accessibility of healthcare and scientific knowledge, and other societal benefits. Researchers also make important contributions through community-engaged research, clinical trial recruitment and retention, stakeholder engagement, mentorship, development of shared research resources, and dissemination of evidence to practitioners, policymakers, patients, and the public.

Evaluation frameworks should therefore incorporate a broad portfolio of indicators that reflect these diverse pathways to impact. Potential measures include contributions to clinical practice guidelines, public health recommendations, healthcare quality improvements, disease prevention efforts, policy development, patient and community engagement, adoption of research-informed practices, workforce development, economic impacts, and evidence of increased public trust and confidence in science. Assessments may also consider longer-term outcomes such as improvements in population health, reductions in disease burden, healthcare cost savings, job creation, technology adoption, and economic activity resulting from research translation and innovation.

Because many public benefits emerge over extended timeframes and cannot easily be attributed to a single project, NIH should employ mixed-method assessment approaches that combine quantitative indicators with qualitative evidence. Implementation could include standardized impact narratives within existing reporting processes, periodic case studies documenting pathways from discovery to application, stakeholder and community feedback, expert review, and the use of evidence already captured through publications, clinical guidelines, policy documents, and public data sources. NIH may also consider periodic retrospective assessments to evaluate the long-term influence of research investments and document how discoveries contribute to health, policy, economic, and societal outcomes over time.

Because no single metric can adequately capture the full public impact of biomedical research, NIH should adopt assessment approaches that integrate multiple forms of evidence. A mixed-method approach offers several benefits, including a more comprehensive assessment of research value, recognition of multiple pathways to impact, reduced reliance on any single metric, and stronger alignment with NIH's mission to improve health and reduce disease burden. However, NIH should also consider potential unintended consequences. Excessive reporting requirements could increase administrative burden, while overly prescriptive impact frameworks could disadvantage foundational research, early-career investigators, and projects whose benefits emerge only over longer time horizons. Attribution may also be challenging when societal outcomes result from the cumulative contributions of multiple investigators, institutions, and funding sources. Flexible impact assessment frameworks that rely primarily on existing data sources can help balance accountability, feasibility, and fairness across disciplines, institutions, and career stages.

8. Other Areas: Finally, as NIH considers new approaches for assessing and incentivizing scientific impact, it is essential that evaluation frameworks reflect the full breadth of contributions that advance biomedical research and improve human health. The considerations outlined above underscore the need for a balanced, flexible, and context-sensitive approach to impact assessment that recognizes the diverse pathways through which research generates scientific, clinical, economic, and societal value. Washington State University (WSU) further emphasizes that expert peer review should remain a central component of any assessment framework and should be integrated with both quantitative and qualitative indicators. Reviewers should possess the scientific expertise and contextual understanding necessary to evaluate the significance of research contributions, including their methodologies, hypotheses, and intended outcomes. This is particularly important for interdisciplinary, foundational, community-engaged, and emerging areas of research, where conventional metrics may not adequately capture scientific significance, long-term influence, or future potential. No single metric can fully reflect the value of scientific contributions, and excessive reliance on any one measure risks creating unintended incentives, narrowing research priorities, and overlooking important dimensions of research excellence.

Accordingly, Washington State University recommends that NIH adopt a comprehensive, multi-indicator framework for assessing scientific impact that recognizes contributions across multiple dimensions, including rigor and reproducibility, data and software sharing, mentorship and workforce development, collaboration and team science, entrepreneurship and translation, foundational scientific exploration, and broader public impact. Metrics should inform, rather than replace, expert judgment and should recognize the many ways biomedical research advances knowledge and benefits society. Evaluation systems should remain adaptable to differences in discipline, institution type, career stage, and research context, while avoiding overly prescriptive measures that may discourage innovative, high-risk, exploratory, or community-focused research. NIH should also carefully monitor implementation to identify and mitigate unintended consequences, including increased administrative burden, incentives that favor easily measurable activities over meaningful scientific contributions, disadvantages for certain disciplines or institution types, and disincentives for high-risk or long-term research. A balanced framework that integrates multiple forms of evidence, relies on existing data sources whenever possible, and incorporates ongoing evaluation and refinement would provide a more comprehensive and fair assessment of scientific impact, encourage diverse pathways to success, and better align researcher incentives with NIH's mission to advance rigorous, collaborative, transparent, innovative, and publicly beneficial biomedical research.

Thank you for the opportunity to provide these comments. Washington State University appreciates NIH's leadership in developing a more comprehensive approach to measuring and rewarding scientific impact and looks forward to continued engagement on these important issues.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/2026-08-18-WSU-Response-to-NIH-RFP-on-Measuring-Scientific-Impact.pdf

Description: Washington State University Response to NOT-OD-26-087 / Measuring & Rewarding Scientific Impact

WASHINGTON STATE UNIVERSITY

Office of Research

August 18, 2026

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 630
Bethesda, Maryland 20892

Re: Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Submitted online at https://osp.od.nih.gov/comment-form-measuring-and-rewarding-scientific-impact/.

On behalf of Washington State University (WSU), I appreciate the opportunity to provide the following comments to the NIH Request for Information (RFI) on measuring and rewarding scientific impact (NOT-OD-26-087).

WSU acknowledges NIH's efforts to identify and incentivize measurable indicators of scientific impact that more accurately reflect the collaborative, rigorous, innovative, and mission-driven nature of modern biomedical research. NIH should adopt a balanced and flexible framework that recognizes the full range of scientific contributions while remaining feasible across career stages, disciplines, and institution types. Because biomedical research encompasses a broad spectrum of activities, including foundational discovery science, translational research, clinical research, implementation science, and community-engaged scholarship, definitions of scientific impact should remain adaptable.

As a land-grant institution, WSU recognizes that impactful research often extends beyond traditional academic outputs to include meaningful contributions to communities, industries, healthcare systems, and public well-being. Evaluation frameworks should also acknowledge that scientifically rigorous research can be impactful regardless of whether findings are positive, negative, or null, as null results play a critical role in refining hypotheses, reducing duplication of effort, and strengthening the overall evidence base. While quantitative indicators can provide useful and comparable measures of impact, overreliance on metrics may create unintended administrative burdens for researchers and institutions and may incentivize behaviors that prioritize measurable outputs over meaningful scientific contributions. NIH should therefore incorporate qualitative measures alongside quantitative indicators, allowing investigators to demonstrate impact through narratives, case studies, mentorship, team science, community engagement, policy influence, and other contributions that may not be fully captured through numerical metrics alone. A balanced framework that integrates quantitative and qualitative evidence will better reflect the diverse pathways through which biomedical research advances science and benefits society.

NIH has specifically requested input on measurable indicators, implementation approaches, benefits, unintended consequences, and feasibility considerations, and our comments below are offered within that context.

Rigor and Reproducibility

Washington State University believes rigor and reproducibility should be recognized as fundamental scientific contributions rather than solely as compliance requirements. For biomedical research assessments, NIH could consider adapting industry principles such as the methodologies of Gage Repeatability and Reproducibility, which evaluate measurement consistency and identify sources of variation. Potential biomedical science indicators could include the percentage of publications with publicly available protocols, analysis plans, data, and code; use of preregistration, blinded analyses, validation cohorts, and replication studies; documentation of assay performance and measurement variation; and dissemination of well-designed studies that yield positive, negative, or null findings. Importantly, null findings are a valuable scientific outcome that can refine hypotheses, prevent duplication of effort, and strengthen the overall evidence base and should therefore be recognized as meaningful contributions to scientific progress.

NIH should also recognize contributions to causal research, defined as research that improves understanding of whether and how specific mechanisms, interventions, exposures, programs, practices, or policies influence health-related outcomes. Because causal questions arise across the biomedical research continuum, measures of causal research should be appropriate to disciplinary context and study design. Indicators could include the use of rigorous causal inference approaches, such as randomized controlled trials, natural experiments, quasi-experimental designs, instrumental variable analyses, difference-in-differences methods, regression discontinuity designs, propensity score methods, mediation analyses, longitudinal studies, mechanistic investigations, and the use of causal frameworks. Additional measures could include replication of causal findings across populations or settings and demonstrated influence on clinical practice, public health interventions, or policy decisions.

NIH could support dedicated replication funding opportunities, establish reproducibility networks, and incorporate rigor, transparency, and causal inference considerations into progress reporting and review criteria. While quantitative indicators can provide useful evidence of rigor and impact, excessive reliance on metrics may increase administrative burden for investigators and institutions, incentivize reporting activities over scientific quality, and may overlook meaningful scientific contributions that are not easily quantified. To minimize burden, NIH should prioritize measures that can be derived from existing grant reports, publications, and public data repositories rather than creating new reporting requirements. Quantitative indicators should be complemented by qualitative assessments, including investigator narratives describing methodological innovations, contributions to community standards, advances in causal understanding, efforts to improve research transparency, and the translation of research findings into practice. Such an approach would better capture meaningful scientific advances while preserving flexibility across disciplines. Potential benefits include greater confidence in scientific findings, stronger evidence for decision-making, more efficient use of research resources, and enhanced public trust. However, excessive emphasis on replication, causal attribution, or resource-intensive validation requirements could discourage innovative or exploratory work and disproportionately burden early-career investigators and smaller institutions with limited administrative personnel and support infrastructure. Expectations should therefore be adapted to disciplinary norms and emphasize transparency, methodological rigor, sound causal reasoning, and meaningful scientific contribution rather than uniform replication requirements or metric-driven evaluation.

Data, Software, and Model Sharing

Washington State University supports expanded recognition of data, software, and model sharing as valuable scientific outputs that accelerate discovery, enhance transparency, improve reproducibility, and increase the return on public research investments. Broad access to high-quality datasets, software tools, and computational models enables researchers to validate findings, conduct secondary analyses, develop new hypotheses, leverage existing data for novel applications, and avoid unnecessary duplication of effort. Data sharing can also facilitate interdisciplinary collaboration, support workforce development, improve research efficiency, and accelerate the translation of scientific discoveries into clinical, public health, and community benefits. To realize these outcomes, NIH should continue investing in infrastructure and tools that make data collection, curation, storage, preservation, and sharing easier for researchers. Effective data-sharing policies depend on systems that reduce administrative burden while ensuring research outputs remain accessible, reusable, interoperable, and discoverable by the broader scientific community. Researchers are more likely to participate in data-sharing efforts when repositories, standards, and submission processes are user-friendly, sustainable, and supported through appropriate training, technical assistance, and dedicated funding for data stewardship activities.

The value of data sharing, however, depends not simply on whether data are deposited in a repository, but on the quality, usability, and reusability of the shared resources. Ensuring data quality is therefore essential to realizing the full benefits of data sharing. NIH should encourage the adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, community-developed data and metadata standards, comprehensive data dictionaries, and clear documentation that enable other researchers to understand and reuse shared datasets. Quality assurance could be further supported through repository-based validation checks, quality control procedures, version control systems, provenance tracking, and the use of certified repositories that provide quality review and long-term preservation. NIH could also encourage independent review of datasets during peer review or repository submission processes to ensure completeness, accuracy, and usability. Because high-quality data sharing requires specialized expertise, continued investment in professional data managers, curators, informatics support, and researcher training in data management and quality assurance practices is essential.

Measurable indicators could include deposition of FAIR-compliant datasets, compliance with NIH Data Management and Sharing requirements, use of persistent identifiers and interoperable metadata, dataset citations and downloads, software adoption metrics, evidence of successful dataset reuse, and demonstrations that shared resources have enabled new collaborations, publications, discoveries, or applications. Assessments should also consider indicators of data completeness, documentation quality, adherence to community standards, reproducibility, and successful reuse by other researchers, rather than relying solely on measures of dataset volume or repository deposits. NIH could further encourage these activities by recognizing datasets, software tools, workflows, computational models, and curated data resources as distinct research products in grant applications, progress reports, and promotion processes; providing supplemental funding for data curation, maintenance, and stewardship; and recognizing investigators who develop widely used community resources. Community-governed data-sharing trust models may provide additional mechanisms for balancing collaboration, stewardship, accountability, and appropriate protections for sensitive data.

The benefits of these approaches include increased data reuse, accelerated validation of findings, more efficient use of research resources, stronger collaboration across institutions, improved reproducibility, and broader dissemination of scientific knowledge. Potential unintended

consequences include incentives for quantity over quality, increased burdens associated with maintaining shared resources, and challenges related to privacy, consent, and the sharing of sensitive human participant data. Flexibility will be required to accommodate disciplinary differences and ensure that early-career researchers receive meaningful credit for creating and maintaining high-quality shared resources regardless of traditional publication volume. By emphasizing data quality, stewardship, and meaningful reuse alongside accessibility, NIH can encourage the sharing of well-curated, reusable datasets while minimizing incentives that prioritize quantity over scientific value.

Training and Mentorship

Washington State University recommends that training and mentorship be recognized as critical scientific contributions that support the long-term strength, sustainability, and diversity of the biomedical research workforce. Effective evaluation should include both leading indicators, such as mentoring engagement, completion of individual development plans, participation in professional development activities, and mentor training, and lagging indicators, including trainee retention, graduation rates, career advancement, workforce placement, research productivity, success in obtaining independent funding, contributions to publications and shared research resources, development of interdisciplinary collaborations, participation in team science, and translation of research into clinical, public health, community, or policy outcomes. Because many important outcomes of training and mentorship are not readily captured through quantitative metrics alone, NIH should also incorporate qualitative assessments. Particularly valuable measures could include confidential trainee and alumni feedback, mentoring effectiveness assessments, evidence of leadership development, case studies demonstrating career progression, examples of successful sponsorship and networking, contributions to creating inclusive and supportive research environments, and narratives describing how mentoring has influenced scientific development, career decisions, collaboration, and professional growth. Standardized and confidential mentoring assessments, including exit surveys, could provide additional evidence of mentoring quality while minimizing reporting burden. Benefits of recognizing training and mentorship contributions include stronger workforce development, improved trainee retention, enhanced research culture, and more sustainable scientific careers. However, overreliance on quantitative outcomes may disadvantage mentors working with populations facing structural barriers or in disciplines with longer training and career timelines. NIH should therefore account for career stage, institutional context, and disciplinary variation when evaluating mentoring impact and adopt a balanced approach that integrates both quantitative and qualitative evidence.

Collaboration

Washington State University encourages NIH to develop mechanisms that recognize meaningful contributions to collaborative, team-based, interdisciplinary, and community-engaged biomedical research rather than relying primarily on conventional authorship metrics. As many of today's complex biomedical challenges require expertise spanning multiple disciplines, sectors, and stakeholder groups, assessment frameworks should reflect the increasingly collaborative nature of scientific discovery. At the same time, the quality, depth, and integration of collaboration should be more highly valued than the number of collaborators involved, because increased team size does not necessarily equate to increased scientific impact.

Potential quantitative indicators include documented individual contributions through standardized contributor role reporting systems; participation in multi-investigator, multi-institutional, and cross-sector collaborations; the number and diversity of disciplines represented on research teams; cross-disciplinary publications; sustained partnerships; development of shared datasets, tools, methods,

software, and research infrastructure; and evidence that interdisciplinary approaches have generated novel research directions, discoveries, or applications. Additional indicators could include achievement of team-specific project goals, establishment of new collaborations, and community participation in research activities.

Because the value of collaborative and interdisciplinary research is not fully captured by quantitative measures alone, NIH should also incorporate qualitative assessments. Useful indicators could include investigator and partner narratives describing how diverse expertise contributed to scientific advances, examples of successful integration of different methodologies or perspectives, evidence of collaboration across traditionally separate disciplines, stakeholder and community feedback, and case studies demonstrating how interdisciplinary approaches improved research relevance, quality, translation, or societal benefit. NIH should also recognize contributions that may not be adequately reflected through authorship or publication metrics, including the development of community partnerships, stakeholder engagement activities, clinical trial recruitment and retention efforts, shared research infrastructure, and other activities that enable successful team science. NIH could support these efforts by incorporating contributor statements into reporting processes and encouraging the use of collaborative impact portfolios that capture diverse forms of participation and scientific contribution.

Such approaches would better reflect the realities of modern biomedical research, promote interdisciplinary problem solving, strengthen community engagement, and recognize expertise that may otherwise be overlooked in traditional assessment systems. Potential benefits include increased innovation, stronger scientific rigor, accelerated translation of discoveries, and improved capacity to address complex biomedical and public health challenges. To minimize administrative burden, NIH should rely whenever possible on information already available through grant applications, publications, progress reports, and publicly available research outputs rather than creating new reporting requirements. Evaluation frameworks should also provide sufficient flexibility to account for disciplinary and institutional differences while maintaining transparency and accountability in the assessment of individual and team contributions.

Entrepreneurship and Translation

Washington State University encourages NIH to recognize innovation and entrepreneurship as important pathways through which biomedical research generates public benefit and advances the agency's mission of improving health. While publications and citations remain important indicators of scholarly activity, they do not fully capture the translational, clinical, economic, and societal impacts of scientific discovery. NIH should therefore recognize a broad range of innovation-related outputs, including patents, licenses, startup companies, software tools, diagnostics, therapeutics, medical devices, implementation strategies, and other technologies that advance research or improve health outcomes. Potential indicators could include invention disclosures, patent and licensing activity, startup formation, participation in and outcomes from Small Business Innovation Research and Small Business Technology Transfer programs, follow-on investment, public-private partnerships, technology adoption, and demonstrated impacts on healthcare practice, public health, or economic development.

Assessment frameworks should also recognize key translational and regulatory milestones that mark progress along the pathway from discovery to application, including Investigational New Drug applications, Investigational Device Exemption filings, Federal Drug Administration submissions and approvals, advancement through clinical trial phases, development of clinical guidelines, and implementation of evidence-based interventions. Because innovation often follows a lengthy and

non-linear trajectory, NIH should consider measures of long-term impact, such as sustained adoption of technologies or interventions, downstream improvements in health outcomes, continued commercial development, market penetration, follow-on scientific advances enabled by the innovation, and enduring contributions to research infrastructure, public health practice, and economic growth.

Economic indicators can provide valuable evidence of the broader return on public research investments and may include follow-on federal, nonprofit, or private-sector funding; licensing activity and revenue; startup company sustainability; job creation and workforce development; healthcare cost savings or efficiency gains; technology adoption; manufacturing activity; and economic activity generated within local or regional innovation ecosystems. However, economic measures should complement, rather than replace, assessments of scientific quality, innovation, and public health benefits.

To ensure fairness, NIH should evaluate innovation and entrepreneurial impact within the context of institutional resources, disciplinary norms, and career stage rather than applying uniform expectations across the biomedical research enterprise. Smaller and rural institutions may have more limited access to technology transfer infrastructure, commercialization support, industry partnerships, venture capital networks, and regulatory expertise, which can affect their ability to translate discoveries into traditional entrepreneurial outcomes despite producing highly impactful research. Similarly, pathways to innovation vary substantially across disciplines, with some fields generating patents, licenses, and commercial products, while others create impact through advances in scientific knowledge, clinical practice, public health, policy, or community outcomes. Early-career investigators may also have fewer opportunities to demonstrate long-term innovation impacts because translational and commercialization outcomes often require many years to mature. NIH should therefore assess innovation relative to opportunity and stage of development and recognize a broad range of translational, scientific, societal, and economic contributions. Such an approach would preserve incentives for innovation across diverse institutional settings, disciplines, and career stages.

Given the diverse pathways through which innovation creates value, quantitative indicators should be complemented by qualitative assessments, including case studies, investigator narratives, peer review, and evidence of contributions to innovation ecosystems, entrepreneurial mentorship, workforce development, and cross-sector partnerships. A balanced approach would encourage translation and entrepreneurship while recognizing the many ways federally funded research produces lasting scientific, health, societal, and economic benefits.

Foundational Scientific Exploration

Washington State University strongly supports NIH's continued investment in foundational scientific exploration, including high-risk, high-reward research that may not generate immediate results but has the potential to produce transformative advances. NIH has long demonstrated its commitment to this area through programs such as the High-Risk, High-Reward Research Program, Pioneer Awards, New Innovator Awards, Transformative Research Awards, and other funding mechanisms designed to support bold, creative, and unconventional ideas. Because paradigm-shifting discoveries often emerge from ambitious scientific questions, novel approaches, and investigations whose outcomes are inherently uncertain, NIH should ensure that foundational and exploratory research remains a visible and valued component of the broader biomedical research enterprise.

Indicators of impact should extend beyond traditional measures of productivity and continue to emphasize novelty, originality, and scientific creativity. Additional indicators could include the development of new theories, methodologies, or enabling technologies; the establishment of entirely new areas of inquiry; follow-on research stimulated by exploratory projects; and expert assessments of scientific risk, innovation potential, and intellectual contribution. NIH should continue to employ review criteria specifically designed for exploratory research and maintain funding mechanisms that provide investigators with the flexibility to pursue ambitious and unconventional ideas, including milestone-based awards, pilot and proof-of-concept programs, and longer-term funding opportunities that recognize the extended timelines often required for transformative discoveries.

Because the impacts of foundational research frequently emerge over years or decades, evaluation frameworks should emphasize long-term influence rather than short-term productivity alone. Quantitative and qualitative measures could include the development of new scientific fields, theories, methods, datasets, software tools, or research platforms; follow-on funding and research activity enabled by an initial discovery; downstream clinical, public health, or technological applications; and expert assessments or case studies documenting how exploratory research influenced subsequent advances. NIH could further embed and sustain high-risk, high-reward research within the academic enterprise by encouraging institutional recognition of such contributions in promotion, tenure, and annual evaluation processes; supporting interdisciplinary collaborations that foster innovative thinking; and providing career-stage protections that allow investigators, particularly early-career researchers, to pursue ambitious research directions without undue concern over short-term publication or funding outcomes. NIH should also promote a research culture that recognizes well-executed studies with negative or inconclusive results as valuable scientific contributions when they advance understanding, refine hypotheses, or open new avenues for investigation.

The principal benefit of this approach is the promotion of creativity, innovation, and breakthrough discoveries that expand the boundaries of knowledge and create opportunities for future scientific, clinical, and societal advances. Potential challenges include the difficulty of measuring impact over shorter timescales and distinguishing thoughtful scientific risk-taking from poorly conceived projects. Evaluation frameworks should therefore account for the extended timeline often associated with foundational discovery and ensure that investigators and institutions are appropriately recognized and rewarded for pursuing innovative research with the potential to generate substantial long-term scientific and public benefit.

Public Impact

Washington State University believes assessments of the public impact of biomedical research should focus on demonstrable contributions to health, society, and public well-being. While publications and citations provide evidence of scholarly productivity, they do not fully capture the many ways research benefits society. Biomedical research creates public value through improvements in health outcomes, clinical practice, healthcare delivery, economic growth, workforce development, public trust in science, accessibility of healthcare and scientific knowledge, and other societal benefits. Researchers also make important contributions through community-engaged research, clinical trial recruitment and retention, stakeholder engagement, mentorship, development of shared research resources, and dissemination of evidence to practitioners, policymakers, patients, and the public.

Evaluation frameworks should therefore incorporate a broad portfolio of indicators that reflect these diverse pathways to impact. Potential measures include contributions to clinical practice guidelines, public health recommendations, healthcare quality improvements, disease prevention efforts, policy development, patient and community engagement, adoption of research-informed practices, workforce development, economic impacts, and evidence of increased public trust and confidence in science. Assessments may also consider longer-term outcomes such as improvements in population health, reductions in disease burden, healthcare cost savings, job creation, technology adoption, and economic activity resulting from research translation and innovation.

Because many public benefits emerge over extended timeframes and cannot easily be attributed to a single project, NIH should employ mixed-method assessment approaches that combine quantitative indicators with qualitative evidence. Implementation could include standardized impact narratives within existing reporting processes, periodic case studies documenting pathways from discovery to application, stakeholder and community feedback, expert review, and the use of evidence already captured through publications, clinical guidelines, policy documents, and public data sources. NIH may also consider periodic retrospective assessments to evaluate the long-term influence of research investments and document how discoveries contribute to health, policy, economic, and societal outcomes over time.

Because no single metric can adequately capture the full public impact of biomedical research, NIH should adopt assessment approaches that integrate multiple forms of evidence. A mixed-method approach offers several benefits, including a more comprehensive assessment of research value, recognition of multiple pathways to impact, reduced reliance on any single metric, and stronger alignment with NIH's mission to improve health and reduce disease burden. However, NIH should also consider potential unintended consequences. Excessive reporting requirements could increase administrative burden, while overly prescriptive impact frameworks could disadvantage foundational research, early-career investigators, and projects whose benefits emerge only over longer time horizons. Attribution may also be challenging when societal outcomes result from the cumulative contributions of multiple investigators, institutions, and funding sources. Flexible impact assessment frameworks that rely primarily on existing data sources can help balance accountability, feasibility, and fairness across disciplines, institutions, and career stages.

Other Areas

Finally, as NIH considers new approaches for assessing and incentivizing scientific impact, it is essential that evaluation frameworks reflect the full breadth of contributions that advance biomedical research and improve human health. The considerations outlined above underscore the need for a balanced, flexible, and context-sensitive approach to impact assessment that recognizes the diverse pathways through which research generates scientific, clinical, economic, and societal value. Washington State University (WSU) further emphasizes that expert peer review should remain a central component of any assessment framework and should be integrated with both quantitative and qualitative indicators. Reviewers should possess the scientific expertise and contextual understanding necessary to evaluate the significance of research contributions, including their methodologies, hypotheses, and intended outcomes. This is particularly important for interdisciplinary, foundational, community-engaged, and emerging areas of research, where conventional metrics may not adequately capture scientific significance, long-term influence, or future potential. No single metric can fully reflect the value of scientific contributions, and excessive reliance on any one measure risks creating unintended incentives, narrowing research priorities, and overlooking important dimensions of research excellence.

Accordingly, WSU recommends that NIH adopt a comprehensive, multi-indicator framework for assessing scientific impact that recognizes contributions across multiple dimensions, including rigor and reproducibility, data and software sharing, mentorship and workforce development, collaboration and team science, entrepreneurship and translation, foundational scientific exploration, and broader public impact. Metrics should inform, rather than replace, expert judgment and should recognize the many ways biomedical research advances knowledge and benefits society. Evaluation systems should remain adaptable to differences in discipline, institution type, career stage, and research context, while avoiding overly prescriptive measures that may discourage innovative, high-risk, exploratory, or community-focused research. NIH should also carefully monitor implementation to identify and mitigate unintended consequences, including increased administrative burden, incentives that favor easily measurable activities over meaningful scientific contributions, disadvantages for certain disciplines or institution types, and disincentives for high-risk or long-term research. A balanced framework that integrates multiple forms of evidence, relies on existing data sources whenever possible, and incorporates ongoing evaluation and refinement would provide a more comprehensive and fair assessment of scientific impact, encourage diverse pathways to success, and better align researcher incentives with NIH's mission to advance rigorous, collaborative, transparent, innovative, and publicly beneficial biomedical research.

Thank you for the opportunity to provide these comments. Washington State University appreciates NIH's leadership in developing a more comprehensive approach to measuring and rewarding scientific impact and looks forward to continued engagement on these important issues.

Sincerely,



Kim Christen, PhD
Vice President for Research
Washington State University

196. Broad Institute

I am responding: On behalf of an organization

Name:

Name of Organization: Broad Institute

Type of Organization: Non-Profit Research Organization

Role: Other

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Please see the attached PDF for Broad Institute's response to the NIH Request for Information on Measuring and Rewarding Scientific Impact.

2. Data, Software, and Model Sharing: Please see the attached PDF for Broad Institute's response to the NIH Request for Information on Measuring and Rewarding Scientific Impact.

3. Training and Mentorship: Please see the attached PDF for Broad Institute's response to the NIH Request for Information on Measuring and Rewarding Scientific Impact.

4. Collaboration: Please see the attached PDF for Broad Institute's response to the NIH Request for Information on Measuring and Rewarding Scientific Impact.

5. Entrepreneurship and Translation: Please see the attached PDF for Broad Institute's response to the NIH Request for Information on Measuring and Rewarding Scientific Impact.

6. Foundational Scientific Exploration: Please see the attached PDF for Broad Institute's response to the NIH Request for Information on Measuring and Rewarding Scientific Impact.

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Broad-Institute-NIH-RFI-Scientific-Impact-Response-1.pdf

Response to NIH Request for Information on Measuring and Rewarding Scientific Impact
NIH Notice Number: NOT-OD-26-087
Date: August 18, 2026
Email: govrelations@broadinstitute.org

I. Rigor and Reproducibility

Rigorous methodologies and open data sharing are foundational parts of the Broad Institute's mission. We welcome the NIH Director's focus on prioritizing agency-wide reproducibility and elevating scientific standards across federally funded research. The recommendations below outline how NIH can build on these established practices by leveraging shared computational infrastructure, encouraging secondary data analysis, and establishing clear frameworks for causal inference.

a. **Embed research integrity into lab culture**: Encourage standardized onboarding materials and clearly documented expectations for data management and intra-laboratory reproducibility.
    i. Maintain standardized onboarding documentation that outlines laboratory norms, expectations, and procedures for intra-laboratory reproducibility.
    ii. Implement standard operating procedures (SOPs) for routine lab-level data checks, error logging, and protocol verification before publishing.
b. **Promote digital data tools and internal quality checks:** Support the adoption of digital tools, standardized review practices, and internal quality checks to strengthen scientific rigor:
    i. **ELN adoption & review SOPs**: Encourage the use of Electronic Lab Notebooks (ELNs) that incorporate NIH Data Management and Sharing (DMS) guidelines and support documented data management and intra-lab reproducibility practices. Ask institutions to report on the use of ELNs and the maintenance of standard operating procedures (SOPs) for reviewing data and handling identified issues.
    ii. **Data sharing in shared computational infrastructure**: Encourage researchers to share detailed methods, protocols, and supporting data to maximize the reach and impact of research findings. Robust data sharing facilitates the reuse of research outputs, improves reproducibility, accelerates new insights, and allows other investigators to build seamlessly on prior discoveries. To fully support this, investing in shared computational infrastructure for storing and sharing/analyzing data can ease technical barriers and ensure these research assets can be fully leveraged across the scientific community.
c. **Establish multi-track mechanisms to validate high-Impact research:** Without systematic validation, researchers risk investing time and federal resources into building on results that are incomplete or unrepeatable. To ensure foundational science is robust

before the field expands on it, the NIH should support validation through two complementary tracks:

i. **Computational validation (secondary analysis)**: Provide dedicated funding for computational scientists to re-evaluate high-impact, publicly deposited datasets using alternative statistical pipelines, novel algorithms, or cross-study cohorts to verify analytical robustness.

ii. **Laboratory validation (experimental replication)**

1. Direct replication: Establish dedicated funding mechanisms for independent labs to repeat and report on high-impact, foundational experiments conducted elsewhere.
2. Foster author collaboration and SOPs: Pair replicating labs with original study authors to troubleshoot methodological nuances, cross-train, and encourage the publication of open-access, standardized operating procedures (SOPs) for the broader scientific community.

II. Data, Software, and Model Sharing

At the Broad Institute, the open sharing of data, software, and predictive models is fundamental to accelerating discovery and enabling peer validation. One of the most efficient and scalable paths to reproducibility is ensuring primary research outputs are AI-ready, standardized, and sustainably funded. Recognizing these assets as core scientific outputs across NIH funding and institutional evaluation will ensure the broader community can seamlessly access, evaluate, and build upon prior discoveries. To build a sustainable ecosystem for data reuse, NIH should focus on the following strategies:

a. **Grant review recognition**: Encourage grant applications to include a summary of the investigator's history of sharing data, software, and resources, if available, and consider this information as part of grant review. This commitment should be evaluated as a positive indicator of scientific impact during peer review.

b. **Sustainable funding models**: Establish direct-cost mechanisms and allowable budget lines for long-term data curation, software maintenance, and storage. Supporting these needs through shared computational infrastructure prevents "data graveyards," relieves individual labs of long-term hosting burdens, and helps ensure data generated with NIH support remains readily available for research.

c. **Academic promotion criteria**: Encourage research institutions to formally recognize the development, maintenance, and sharing of open-source datasets, models, and tools as primary criteria for tenure, promotion, and career advancement.

d. **Evaluate impact of shared resources:** Evaluate the reach and value of shared resources using field-appropriate indicators (such as downstream preprints, clinical trial applications, code forks, and access requests) rather than raw download counts. A specialized tool or dataset in a niche field may have transformative scientific value that raw download numbers may fail to reflect.

III. Training and Mentorship

Mentorship is an important component of scientific impact, so building high-performing training environments are practices embedded in Broad's operational model. We encourage NIH to evaluate mentorship through a comprehensive set of indicators - measuring success by trainee growth, long-term career advancement, transition to scientific independence, and real-world workforce retention rather than grant volume alone. Potential indicators to assess and quantify the impact of mentorship include:

a. **Mentorship training**: Encourage the completion of formal mentorship training programs, such as those offered through the Center for the Improvement of Mentored Experiences in Research (CIMER), as an indicator of effective mentorship.
b. **Trainee-to-PI transition**: Evaluate how effectively training environments support early-career researchers as they navigate the transition into independent research roles and scientific leadership.
c. **Trainee recruitment**: Track whether mentorship environments attract new trainees and support them in successfully completing their training with the PI.
d. **Trainee feedback**: Incorporate anonymous or confidential feedback and exit surveys to assess the quality of mentorship and to truly capture the trainee's experience.

IV. Collaboration

At the Broad Institute, complex biomedical breakthroughs are driven by team-based science, shared platform infrastructure, and interdisciplinary hubs designed to accelerate scientific impact everywhere, not just at Broad. Evaluation frameworks should look beyond raw publication or collaboration counts to capture real functional impact, middle-author leadership, platform engineering, and cross-disciplinary contributions. The approaches below recognize the diverse expertise, leadership, and contributions that enable successful team-based science.

a. **Clear classification of collaboration**: Establish standardized federal definitions for "collaborator" and "active collaborator" to better assess contributions.
b. **Collaborative research engagement:** Evaluate the depth, functional role, and strategic impact of a scientist's contributions within a team, rather than the raw number of collaborative projects.
c. **Funding contributions**: Recognize individuals who are named collaborators on successful funding applications, including sub-awards and collaborative grants.
d. **Collaborator feedback mechanisms**: Incorporate qualitative and quantitative feedback from collaborators and co-investigators regarding an individual's contributions. This feedback could be included in grant progress reports, publication processes, or promotion reviews.

V. Entrepreneurship and Translation

Moving biomedical discoveries from "bench to bedside" requires coordination across research, biomanufacturing, industry partnerships, and clinical organizations. To capture the true breadth of these efforts, evaluation frameworks must recognize high-risk, high-reward research,

especially within regional innovation hubs where proximity to expertise, advanced infrastructure, and commercial partners accelerates biomedical progress. We encourage NIH to consider the following indicators when assessing translational impact across funded institutions:

a. **Startup creation:** Track the formation, growth, and trajectory of spin-off companies created to advance discoveries, evaluated by their technology pipelines and commercial viability.
b. **Industry R&D partnerships**: Acknowledge robust collaborations with pharmaceutical and biotech partners that are focused on developing therapies, diagnostic tools, and biological insights.
c. **External & private sector investment:** Evaluate evidence of meaningful sponsored research funding, co-development capital, and private investments secured from industry partners to advance bench-to-bedside projects.
d. **IP deployment & licensing:** Consider active technology licenses issued to external/commercial parties to transition lab discoveries and inventions into market pipelines and clinics.
e. **Clinical pipeline progression**: Monitor the number of patents filed that successfully lead to licensing agreements, Investigational New Drug (IND) applications, or clinical product launches.

VI. Foundational Scientific Exploration

At the Broad Institute, we believe paradigm-shifting breakthroughs require bold exploration and an inherent tolerance for risk. Transformative science is rarely linear; it takes time, involves unexpected turns, and carries a real possibility of failure. To sustain true foundational discovery, evaluation frameworks must accommodate non-linear progress, account for delayed impact, and treat informative failures as valuable contributions to scientific knowledge.
We encourage NIH to support and incentivize high-risk, high-reward research through the following approaches:

a. **Value learning from failure:** Encourage researchers to share well-documented negative results and preprints from high-risk projects. Publishing these insights saves time and resources across the scientific community, preventing duplicate work and advancing progress overall. Program officers and reviewers should view informative failures as valuable scientific outputs rather than signs of an unproductive lab.
b. **Create flexible funding mechanisms**: Develop short-term seed grants (e.g. 2-3 years) to test high-risk research questions, paired with a streamlined peer-review process to convert promising early results into multi-year funding.
c. **Recognize long-term, delayed impact**: Adopt metrics to recognize publications and discoveries that take years to demonstrate their full impact, reinforcing the value of supporting paradigm-shifting research that may not show immediate impact.
d. **Empower scientific program leadership:** Expand the role of scientifically knowledgeable NIH program officers to identify and champion high-risk opportunities. Program officers and managers with deep expertise in their fields can help guide strategic investments and recognize promising research directions.

197. Emmanuel Okoro

I am responding: On behalf of myself

Name: Emmanuel Okoro

Name of Organization: Meharry Medical College

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: The pure mission of "rigor and reproducibility" leads to designing machines or researching on drugs that target "receptors." In practice, rigor and reproducibility represent purifying or synthesizing chemicals. This approach is disregarding the rising problem: Accumulating nomenclature/catalog based on search-subjective or profitable origin to finish end-related (offer), and alternatively speaking, "force" acronym data. These are evident in multi-conditionality jumbled "pathway and signaling" maps.

2. Data, Software, and Model Sharing: It's not enough to send finished products and call them "sharing." Understanding and communicating natural basics in comprehending these "tools" should be a priority.

3. Training and Mentorship: Training and mentorship need to be diversified and improved by addressing real world problems. Concentrating under old and identical mentality of thought does not nurture growth and independence. It creates stagnation. To advance training, even the mentors (directly or indirectly experienced adults) need to recognize their limitations and seek ways to improve beyond technical and/or work outputs.

4. Collaboration: More people working together does not necessarily represent collaboration. It may simply equal working under the same roof. If collaboration merely increases the same output, perhaps it should be called manufacturing. In order words, is collaboration leading to independently noticeable improvements or merely increasing socializing networks?

5. Entrepreneurship and Translation: Before rushing to entrepreneurship and translation, holistic and basics problems need to carefully addressed.

6. Foundational Scientific Exploration: Addressing real world research challenges is not "high-risk," but a necessity for improvement. High-risk implies skydiving or reckless betting for profits. The outcomes are predictable.

7. Public Impact: It's not enough to present publications results. These are endless and subjective. Leadership and research efforts need to focus on endeavors that improve natural and relational comprehension on all deficiencies.

8. Other Areas: N/A

198. ASM

I am responding: On behalf of an organization

Name: Nicole Zimmerman

Name of Organization: ASM

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Federal Affairs

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Reproducibility is a cornerstone of the scientific process. However, the scientific community has become increasingly aware that some published scientific findings cannot be reproduced. We agree that steps should be taken to address this issue. One way to measure reproducibility and impact is tracking the frequency with which publications are supported or contradicted by subsequent research, using citation statements classified as supporting, contrasting, or merely mentioning prior work. Unlike traditional citation counts, such measures can provide insight into whether findings are being independently validated by the research community. Journal- or researcher-level indicators based on evidence of replication, support, and reproducibility could complement existing metrics and help incentivize more robust and reliable science. Another potential indicator of scientific quality that is more robust than traditional citation counts is tracking where research is published. Major indexing and archiving services, such as Web of Science, Scopus, the Directory of Open Access Journals and PubMed Central, require journals to meet established standards related to editorial quality, transparency, ethics, technical rigor, and accessibility. Incentives for researchers are needed to address reproducibility concerns. Replication studies are frequently overlooked because researchers rarely receive dedicated financial support to undertake research whose primary purpose is replication. Instead, the scientific community maintains scientific rigor through the peer review process, which ensures the quality and reliability of scholarly publications. However, robust peer review is time consuming work, done in addition to other research and service duties, disincentivizing the effort needed to thoroughly review articles prior to publication. ASM suggests the following indicators of rigor and reproducibility: • Incorporate independent, external measures of research reliability and reproducibility into NIH's measures of scientific impact, such as the Scite Index. • Track how often individual researchers publish in journals indexed by major indexing and archiving services with established standards. ASM suggests the following incentives to promote rigor and reproducibility: • Create funding streams for replication studies that supplement and do not supplant existing funding streams for original research. • Recognize and reward researchers' participation in the peer review process by measuring peer review as part of the researcher's overall scientific impact.

2. Data, Software, and Model Sharing: Data, Software, and Model Sharing Scientific advancement occurs through continual expansion of knowledge, which requires access to accurate and complete data for confirmation of findings and building on previously published knowledge. While sharing data, datasets, and software continues to improve, some data or code lack the necessary information and metadata to

make them readily usable by other researchers when they are shared. Measuring the quality of the shared data or software, in addition to the frequency of use, and rewarding groups that consistently publish usable, high-quality data and software can begin to change the open-access scientific data and software landscape. ASM suggests the following indicator to promote data, software and model sharing: • Implementation of a sharing index that measures how often researchers' data and software is used and cited, when it is shared and the completeness and FAIR (Findable, Accessible, Interoperable, Reusable) compliance of their data. ASM supports that agencies provide funding incentives to promote data, software and model sharing for labs that produce and share high-quality, well-annotated public data and datasets that become community resources.

3. Training and Mentorship: Mentorship of trainees is a cornerstone of the scientific enterprise yet is often not evaluated in grant applications and traditional measures of scientific impact and scientific training. ASM suggests agencies provide funding for training of mentors in leadership, mentoring, and managing people.

4. Collaboration: N/A

5. Entrepreneurship and Translation: Although moving basic science discoveries into clinical and field settings is a key goal of translational research, the translation of science extends beyond the development and commercialization of a finished product. Metrics of translational impact should therefore capture the full range of pathways through which research informs practice, policy, technology, and future innovation. Even if the translation fails, capturing the attempt to translate, provided the science was rigorous, can provide indicators of researchers looking to move their science beyond the lab and into practice. This, paired with ongoing funding for programs like the Small Business Innovation Research and Small Business Technology Transfer (SBIR/STTR) programs can help to encourage researchers to dedicate more of their efforts to exploring translation of their findings for the public good. ASM suggests the following indicator to measure entrepreneurship and translation: • Track activities related to the commercialization of research, such as the issuance of patents, investigational new drug applications and validated biomarkers, and manuscript reads that may be linked to commercial interest but not result in citations linked to research interest. • ASM suggests the following incentive to promote entrepreneurship and translation: • Continue robust funding for SBIR/STTR programs and other programs to promote entrepreneurship among researchers, such as the iCorps program.

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/ASM-NIH-Scientific-Impact-RFI-Response.pdf

ASM
Microbes Make Our World

1752 N St. NW
Washington, DC 20036
asm.org

Dr. Lyric Jorgenson
Director
NIH Ofice of Science Policy
9000 Rockville Pike
Bethesda, MD 20892

August 19, 2026

Re: Request for Information (RFI) on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Dear Dr. Jorgensen,

Thank you for the opportunity to respond to the Request for Information (RFI) on Measuring and Rewarding Scientific Impact. ASM is one of the oldest and largest life science societies with over 38,000 members in the U.S. and around the world who are dedicated to conducting robust, rigorous research.

ASM appreciates NIH's interest in promoting scientific rigor and developing alternative measures for scientific success, beyond traditional measures like publication and citation counts. ASM integrates scientific rigor and reproducibility across our programs and our portfolio of 17 peer reviewed journals. Please see below for ASM's comments on rigor and reproducibility, data, software and model sharing, training and mentorship as well as entrepreneurship and translation categories described in the RFI.

Rigor and Reproducibility

Reproducibility is a cornerstone of the scientific process. However, the scientific community has become increasingly aware that some published scientific findings cannot be reproduced. We agree that steps should be taken to address this issue.

One way to measure reproducibility and impact is tracking the frequency with which publications are supported or contradicted by subsequent research, using citation statements classified as supporting, contrasting, or merely mentioning prior work. Unlike traditional citation counts, such measures can provide insight into whether findings are being independently validated by the research community. Journal- or researcher-level

1752 N St. NW
Washington, DC 20036
asm.org

indicators based on evidence of replication, support, and reproducibility could complement existing metrics and help incentivize more robust and reliable science.

Another potential indicator of scientific quality that is more robust than traditional citation counts is tracking where research is published. Major indexing and archiving services, such as Web of Science, Scopus, the Directory of Open Access Journals and PubMed Central, require journals to meet established standards related to editorial quality, transparency, ethics, technical rigor, and accessibility.

Incentives for researchers are needed to address reproducibility concerns. Replication studies are frequently overlooked because researchers rarely receive dedicated financial support to undertake research whose primary purpose is replication. Instead, the scientific community maintains scientific rigor through the peer review process, which ensures the quality and reliability of scholarly publications. However, robust peer review is time consuming work, done in addition to other research and service duties, disincentivizing the efort needed to thoroughly review articles prior to publication.

ASM suggests the following indicators of rigor and reproducibility:

- Incorporate independent, external measures of research reliability and reproducibility into NIH's measures of scientific impact, such as the Scite Index.
- Track how often individual researchers publish in journals indexed by major indexing and archiving services with established standards.

ASM suggests the following incentives to promote rigor and reproducibility:

- Create funding streams for replication studies that supplement and do not supplant existing funding streams for original research.
- Recognize and reward researchers' participation in the peer review process by measuring peer review as part of the researcher's overall scientific impact.

Data, Software, and Model Sharing

Scientific advancement occurs through continual expansion of knowledge, which requires access to accurate and complete data for confirmation of findings and building on previously published knowledge. While sharing data, datasets, and software continues to improve, some data or code lack the necessary information and metadata to make them

1752 N St. NW
Washington, DC 20036
asm.org

readily usable by other researchers when they are shared. Measuring the quality of the shared data or software, in addition to the frequency of use, and rewarding groups that consistently publish usable, high-quality data and software can begin to change the open-access scientific data and software landscape.

ASM suggests the following indicator to promote data, software and model sharing:

- Implementation of a sharing index that measures how often researchers' data and software is used and cited, when it is shared and the completeness and FAIR (Findable, Accessible, Interoperable, Reusable) compliance of their data.

ASM supports that agencies provide funding incentives to promote data, software and model sharing for labs that produce and share high-quality, well-annotated public data and datasets that become community resources.

Training and Mentorship

Mentorship of trainees is a cornerstone of the scientific enterprise yet is often not evaluated in grant applications and traditional measures of scientific impact and scientific training. ASM suggests agencies provide funding for training of mentors in leadership, mentoring, and managing people.

Entrepreneurship and Translation

Although moving basic science discoveries into clinical and field settings is a key goal of translational research, the translation of science extends beyond the development and commercialization of a finished product. Metrics of translational impact should therefore capture the full range of pathways through which research informs practice, policy, technology, and future innovation. Even if the translation fails, capturing the attempt to translate, provided the science was rigorous, can provide indicators of researchers looking to move their science beyond the lab and into practice. This, paired with ongoing funding for programs like the Small Business Innovation Research and Small Business Technology Transfer (SBIR/STTR) programs can help to encourage researchers to dedicate more of their eforts to exploring translation of their findings for the public good.

ASM
Microbes Make Our World

1752 N St. NW
Washington, DC 20036
asm.org

ASM suggests the following indicator to measure entrepreneurship and translation:

- Track activities related to the commercialization of research, such as the issuance of patents, investigational new drug applications and validated biomarkers, and manuscript reads that may be linked to commercial interest but not result in citations linked to research interest.
- ASM suggests the following incentive to promote entrepreneurship and translation:
- Continue robust funding for SBIR/STTR programs and other programs to promote entrepreneurship among researchers, such as the iCorps program.

Conclusion

Thank you for the opportunity to respond to this Request for Information. ASM works to ensure scientific rigor across our programs and looks forward to working with NIH to incentivize and reward high-quality science. If you have any questions or would like to further discuss these comments, please contact Nicole Zimmerman, Senior Federal Afairs Oficer, at nzimmerman@asmusa.org.

Thank you,



Amalia Corby
Director of Federal Afairs
ASM

199. American Medical Association

I am responding: On behalf of an organization

Name: Neda Ashtari

Name of Organization: American Medical Association

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/2026-8-18-NIH-RFI-on-Measuring-and-Rewarding-Scientific-Impact.pdf

AMA
AMERICAN MEDICAL ASSOCIATION

John Whyte, MD, MPH
CEO, EXECUTIVE VICE PRESIDENT

john.whyte@ama-assn.org

August 18, 2026

Lyric Jorgenson, PhD
Associate Director for Science Policy
Director, Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: NOT-OD-26-087, Request for Information on Measuring and Rewarding Scientific Impact

Dear Director Jorgenson:

On behalf of the physician and medical student members of the American Medical Association (AMA), I appreciate the opportunity to respond to the National Institutes of Health (NIH) Request for Information (RFI) on Measuring and Rewarding Scientific Impact (NOT-OD-26-087). The AMA commends NIH's efforts to strengthen the biomedical research enterprise in ways that promote rigorous and reproducible science, accelerate transformative discovery, and ultimately improve patient care and public health. We offer the following recommendations to help advance those shared goals.

1. Fund the Research Questions Patients and Physicians Most Need Answered

Despite major advances in biomedical science, physicians encounter questions every day for which little evidence exists to guide clinical decision making. These gaps persist in part because of misaligned incentives in the research enterprise. Commercial sponsors tend to invest in research with clear commercial applications, while academic funding and promotion structures may place greater value on grant success and publication than on a study's potential to improve health outcomes in the real world.

NIH should ensure that taxpayer dollars fund the research questions with the greatest potential to improve public health, health outcomes, and the value of health care. Physicians, who routinely confront evidence gaps in clinical practice, are well positioned to help identify what those questions are. NIH should thus establish a recurring process through which physicians help identify high-priority research questions and should fund studies designed to answer them. Other countries have demonstrated that this approach is workable: in the United Kingdom, the James Lind Alliance convenes clinicians and patients to rank the most important unanswered questions in a given clinical area, and research funders use those rankings to guide investment. A research portfolio aligned with the questions physicians most need answered would also give NIH a direct, defensible measure of the real-world impact this RFI seeks—evidence generated in response to demonstrated clinical need.

2. Promote Comparative Effectiveness Research and Cost-Effectiveness Analysis

Clinical research in the United States generates substantial evidence about whether individual interventions are efficacious but far less comparative effectiveness research (CER) evaluating the relative

benefits and harms of available treatments or cost-effectiveness analysis (CEA) relating a treatment's cost to the health benefit it provides. This leaves important gaps in the evidence base physicians use to guide treatment decisions and policymakers use to shape value-based pricing policies.

To address these evidence gaps, NIH should increase funding for CER and CEA and strengthen its standards for comparator selection. When an established standard of care exists, the agency should require investigators to compare new treatments against an appropriate active comparator rather than placebo alone; when an active comparator is not feasible, investigators should justify the alternative as methodologically appropriate for the study objective, population, and treatment effect being estimated. NIH should also prioritize post-approval CER when treatments enter clinical use without adequate evidence comparing their benefits and harms with existing care.

3. Strengthen Methodological Standards for NIH-Supported Research

NIH should direct research dollars toward studies designed to generate reliable, clinically useful evidence. The agency can accomplish this by translating its peer review framework into explicit methodological requirements for research. In the following recommendations, we discuss four areas in which stronger standards could substantially improve the reliability and clinical value of study findings: endpoint selection, duration of follow-up, external validity, and prespecification of design and analysis. To apply these standards effectively, NIH should ensure that peer review includes experts in biostatistics and trial methodology and, when appropriate, physicians with expertise in the clinical area under study.

Require the use of clinically meaningful endpoints. NIH should require studies to use endpoints that are clinically meaningful and appropriate to the question under investigation. Clinical trials frequently rely on surrogate or intermediate endpoints that are weakly connected to patient outcomes and distort estimates of treatment benefit. A 2024 JAMA analysis found that most surrogate markers used to support Food and Drug Administration approvals in non-oncologic chronic diseases lacked high-strength evidence linking them to clinical outcomes, and a separate meta-epidemiological study found that surrogate primary outcomes can overestimate treatment efficacy by more than 40 percent. Although surrogates are necessary in select circumstances, NIH should require investigators to select primary endpoints that directly measure outcomes relevant to the clinical question; justify any use of surrogate, composite, or intermediate endpoints with evidence linking them to the patient-important outcomes they are intended to predict; and ensure that conclusions are commensurate with the outcomes measured.

Require sufficient duration of follow-up. Clinical trials of treatments intended for prolonged use often end before long-term effects are adequately assessed. As a result, they may detect early treatment effects without establishing whether benefits persist, harms emerge with continued use, or long-term clinical outcomes improve. To resolve these uncertainties, NIH should require investigators to demonstrate that planned follow-up is sufficient to answer the study question and capture benefits and harms over the relevant time horizon. Where extended follow-up within the original trial is impractical, NIH should support extension studies or linkage to registry and claims data to capture long-term outcomes.

Strengthen external validity. Research can only improve patient care when its findings are generalizable to the patients and settings in which an intervention will be used. Yet many trials today enroll populations that are systematically healthier and less medically complex than patients encountered in routine practice and are conducted in settings that bear little resemblance to those in which most care is delivered. These choices undermine external validity, contributing to the efficacy-effectiveness gap and expending taxpayer dollars on research with limited capacity to improve patient outcomes. To help address this gap, NIH should require applicants to identify the populations and settings a study is intended to represent and to explain how their proposed eligibility criteria and site selection support that generalizability.

Investigators who restrict enrollment should provide sufficient scientific justification and should confine their conclusions to the populations their evidence supports.

Require prespecification of key design and analytic decisions. Clinical trials are required to adhere to detailed protocols, yet published results frequently diverge from what those protocols specify. Analyses have found that roughly one third of trials change, add, or omit a primary outcome before publication, and that statistically significant outcomes appear in the published literature several times more often than nonsignificant ones. Studies that alter outcomes or analyses after results are known can produce false or exaggerated conclusions—distortions that then carry through to systematic reviews, clinical guidelines, and treatment decisions. To prevent these distortions from entering the literature, NIH should require investigators to prespecify outcomes, statistical methods, and planned subgroup analyses before data analysis begins and to finalize protocols and statistical analysis plans before unblinded outcome data become available.

4. Strengthen and Enforce Results-Reporting and Data-Sharing Requirements

Despite NIH policies requiring clinical trial results reporting and data sharing, substantial compliance gaps remain. A 2022 Department of Health and Human Services Office of Inspector General audit found that 51 percent of the 72 NIH-funded clinical trials reviewed failed to report results on time, including 35 percent that had not submitted results at all, and that NIH continued to fund new research by responsible parties with outstanding reporting obligations. These failures limit researchers' ability to reproduce, reanalyze, and build upon federally funded work, reducing its reach and impact.

NIH should strengthen enforcement by making compliance with applicable reporting and data-sharing requirements a condition of continued and future funding. Before issuing a new or continuation award, the agency should verify that existing obligations have been satisfied and, when material noncompliance persists, should consider withholding funding, imposing award conditions, or suspending an award.

Beyond strengthening compliance with data-sharing requirements, NIH should ensure that shared data are readily accessible and usable. The agency should establish minimum usability requirements specifying that data must include sufficient metadata, documentation, and analytic materials to support interpretation and reproduction; use common data elements and interoperable formats where standards exist; and be deposited in established repositories when available. To further facilitate reuse, NIH should link each award in NIH RePORTER to its resulting publications, datasets, protocols, statistical analysis plans, and ClinicalTrials.gov records, creating a single record of the research products generated by each award.

5. Protect Scientific Independence and Federal Research Investments

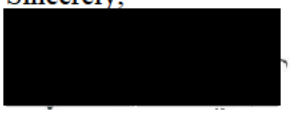
NIH should adopt policies that protect objective, expert-driven decision-making throughout the research award lifecycle, from initial review through continuation, renewal, and termination. These policies should prohibit political officials from selecting reviewers, influencing peer review deliberations, or directing changes to reviewers' scores, and should establish clear standards governing when agency leadership may depart from peer review recommendations. Furthermore, because terminating ongoing research can disrupt scientific progress and waste resources already invested, NIH should provide public notice and an opportunity for comment before terminating a research program for reasons unrelated to noncompliance with award terms or research misconduct and should publish an assessment of the potential consequences for scientific progress, public health, federal investments, and, where applicable, participant safety. These protections should extend to the stability of research funding. NIH should ensure the timely obligation and distribution of available funds, avoid administrative delays or restrictions that impede their use, and preserve annual incremental funding for multiyear grants rather than charging their full value

against a single fiscal year's appropriation. Such an approach would maintain continuity for ongoing studies while leaving resources available for new grants and clinical trials.

Conclusion

The AMA appreciates NIH's commitment to producing rigorous, impactful research that advances biomedical discovery and improves health outcomes. We welcome the opportunity to support the agency in these efforts. Please contact me directly at 312-464-5288 or John.Whyte@ama-assn.org if you have questions or need further information.

Sincerely,



John Whyte, MD, MPH

200. Yale University

I am responding: On behalf of an organization

Name: Sarah Shapiro

Name of Organization: Yale University

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/NIH-RFI-on-Measuring-and-Rewarding-Scientific-Impact-Final.pdf

Yale University
OFFICE OF THE PROVOST

Michael C. Crair, PhD
Vice Provost for Research
William Ziegler III Professor of Neuroscience
Professor of Ophthalmology & Visual Science

P.O. Box 208333
New Haven, CT 06520-8333

2 Whitney Avenue, Suite 400
New Haven, CT 06510

email: michael.crair@yale.edu
Phone: (203) 432-4448

August 18, 2026

Re: Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Dear Director Bhattacharya,

On behalf of Yale University, I appreciate the opportunity to share our feedback for the Request for Information (RFI) on Measuring and Rewarding Scientific Impact (NOT-OD-26-087).

Summary

Yale encourages NIH to modernize its understanding of scientific impact without replacing expert judgment or narrowing the range of science it supports. The appropriate framework is plural, stage-specific, and grounded in evidence. It should recognize rigorous learning, foundational discovery, training (in ethics, rigor, and reproducibility), mentorship, team contributions, responsible sharing, translation, implementation, and public benefit. Across these domains, NIH should identify a limited set of qualitative and quantitative measures; specify whether each applies to projects, investigators, programs, or portfolios; and validate each measure before implementation. Measures should inform expert judgment rather than become formulaic scores and should be revised or discontinued when they add burden, distort incentives, or fail to improve decisions. NIH should be cognizant of the resources of any new systems or processes that it requires during the application, review, funding, and reporting stages to reduce administrative burden.

Yale supports NIH's effort to recognize the defining features of rigorous, collaborative, and mission-driven science. Any new approach should preserve and strengthen the core features that have made NIH successful: sustained bipartisan Congressional investment; a vast extramural network; investigator-initiated science selected through independent expert peer review; expert scientific judgment in the review and conduct of research; high-risk, groundbreaking foundational discovery (such as deciphering how genes encode proteins and developing CRISPR gene editing); and the research workforce and infrastructure that enable innovation and impact.

Scientific impact should be assessed according to the stage of discovery. For each stage, NIH should define relevant indicators prospectively and evaluate results retrospectively over an appropriate time horizon:

- Basic science and methodological research should be assessed by the knowledge, mechanisms, methods, tools, datasets, and capabilities it creates.
- Translational research should be assessed by validation and progression toward practical application.
- Clinical research should be assessed by the quality and relevance of the evidence it produces.
- Implementation research should be assessed by adoption, reach, access, reliability, affordability, sustainability, and population- and individual-level health outcomes.

Retrospective assessments are critical to better understand impacts that may not have been foreseen, which may guide future portfolio assessments. Moreover, projects should not be judged solely by outcomes expected only at a later stage (i.e., development of a new therapeutic intervention). Portfolio priorities may consider disease

burden, unmet need, preventability, potential reach, and scientific opportunity, but these factors should not become formulaic scores for individual proposals. Impact assessments may also consider workforce development, education and training opportunities, and collaborative engagement fostered by projects.

During development, NIH should specify each measure's intended purpose, unit of analysis, and timing of use, distinguishing among projects, investigators, programs, and portfolios. Before implementation, NIH should establish each measure's validity, reliability, feasibility, and incremental value beyond expert review of scientific impact and merit. NIH should also compare new measures with existing quantitative indicators, such as publication count and journal impact factors. Validated measures should first be used retrospectively at the program or portfolio level to evaluate prior investments and inform strategic planning, not to affect individual awards or investigators' future funding. Only after demonstrating value at the relevant level should a measure be considered prospectively in individual award decisions, with advance notice and interpretation through expert review rather than as a formulaic threshold. Validation should include research-intensive universities, smaller institutions, health systems, community organizations, and small businesses, with attention to differences in infrastructure and administrative capacity. NIH should not conflate access to institutional resources with scientific impact and should fund any capacity required to meet new expectations. NIH should discontinue measures that do not improve decisions or that produce excessive burden or unintended incentives.

NIH should match incentives to the barrier being addressed: dedicated awards, supplements, and allowable costs for activities requiring resources; and contribution statements or biosketch recognition for work not captured by conventional outputs. These approaches should precede changes to priority scores or funding criteria, which carry greater risk of distorting behavior and should be considered only after the underlying measures have been validated.

Measures of scientific impact should be distinguished from measures of NIH's operational performance, but both are necessary for efficient, reciprocal accountability. Investigators should report progress they can reasonably influence. Separately, NIH should report review and award timelines, administrative burden, policy stability, and portfolio-level results. New requirements should use existing information whenever possible and should protect research participants, privacy, security, intellectual property, program stability, and scientific independence.

Rigor and Reproducibility

NIH should preserve field-specific expert review and its existing emphasis on sound study design, transparent methods, careful interpretation, and conclusions proportionate to the evidence. Rigor should be defined by whether the design, conduct, analysis, and interpretation are appropriate to the scientific question. This assessment is distinct from determining whether that question is a funding priority. Expected rigor should be assessed prospectively through expert review; reproducibility should be evaluated after completion through transparent reporting, appropriate sharing of data and code, and targeted replication when warranted. Rigor does not require every study to use the same design or every finding to undergo independent replication. The appropriate standard depends on the question, maturity of the evidence, feasibility, and consequences of error.

NIH funding and policies make this possible at scale by supporting not only individual projects but also the shared infrastructure and institutional stewardship on which rigorous research depends. Yale supports NIH's increased attention to replication and reproducibility. Yale's capabilities are targeted rather than centralized. The DISSC (Data-Intensive Social Science Center) Open and Reproducible Research program provides consultation and infrastructure for transparent and reproducible social science; the YODA (Yale University Open Data Access) Project supports responsible clinical-research data sharing; and Yale co-founded medRxiv to enable timely, transparent dissemination of health-science research. Research cores, methodological consultation, and the Yale Center for Clinical Investigation (YCCI) provide broader support for rigorous study design, conduct, analysis, and dissemination. These capabilities serve distinct purposes: DISSC supports computational reproducibility; YODA enables data access and reuse; medRxiv advances transparent dissemination; and research cores and consultation primarily support rigor. None substitutes for dedicated support for independent replication, which requires separate funding and infrastructure.

NIH should assess whether a project's design, conduct, analysis, interpretation, and claims are appropriate to the question and evidence. Relevant indicators may include preregistration and statistical analysis plans when appropriate; transparent reporting of deviations, limitations, negative findings, and credible null results; and adherence to established reporting standards. NIH should not impose a single federal definition of causal research or a universal hierarchy of study designs. Investigators should identify the question being answered, state material assumptions and limitations, and calibrate causal claims to the evidence. Qualitative and exploratory research should be evaluated using field-appropriate standards for methodological fit, transparency, reflexivity, interpretive rigor, and learning value, rather than criteria designed primarily for confirmatory or causal studies.

NIH should integrate and streamline existing reporting requirements to produce a structured, publicly available final scientific report for research awards, proportionate to the award's scale and purpose, while minimizing incremental burden and avoiding a separate reporting process. Reporting fields should reflect the stage of research: knowledge and capabilities for foundational work; validation and progression for translational work; evidence for clinical research; and adoption, reach, access, and outcomes for implementation research. Across stages, reports should link research products and describe the availability of protocols, methods, materials, data, and analytic code; important deviations; null or negative findings; and material limitations. Appropriate exceptions should protect participants, privacy, security, and intellectual property. Machine-readable reports would support reuse and allow the scientific community to develop and validate better measures of impact.

Replication is generally a subsequent scientific activity, often conducted by an independent group, rather than an attribute that every original proposal can demonstrate prospectively. NIH should identify replication priorities at the portfolio level based on scientific importance, uncertainty, downstream reliance, feasibility, and the consequences of an incorrect result. Dedicated awards should support independent replication, while supplements may support validation or methodological work when independence is not essential. These projects should be evaluated for rigor and informativeness rather than conventional novelty.

Replication, null findings, reusable methods, and research-quality infrastructure often receive less recognition than novel positive findings. They also require dedicated time, expertise, and funding. Unfunded documentation or replication mandates could reduce rather than improve scientific productivity. Structured public reporting should consolidate or replace existing requirements, not become an additional reporting layer.

NIH should fund the infrastructure required for rigor, reproducibility, and replication; incorporate relevant activities into allowable budgets; pilot structured reporting before broad implementation; prepopulate it with information already available through existing systems; limit fields to information with demonstrated value; and measure and mitigate investigator and institutional time and cost.

Data, Software, and Model Sharing

Responsible sharing increases the value of federal research by enabling validation, reuse, collaboration, and new discovery. NIH has advanced this objective by making responsible data management and sharing an explicit component of federally supported research, encouraging institutions to build the infrastructure and expertise needed for implementation. Yale supports this work through research-data services, Yale Dataverse, institutional access to Dryad, statistical consultation, secure computing, and clinical-research informatics. These resources help investigators manage, document, preserve, and share research products while meeting ethical and legal obligations.

NIH should recognize data, software, and models as scientific contributions when they are well documented, usable, maintained for an appropriate period, and made accessible under terms consistent with consent, privacy, security, and intellectual-property obligations.

Potential indicators include:

- Use of persistent identifiers, versioning, metadata, data dictionaries, and appropriate documentation.
- Clear access conditions and reasonable time from a valid request to access.
- For software, availability of documentation, testing, dependencies, and reproducible computing environments.
- For scientific models, documentation of provenance, intended use, validation, limitations, version changes, and human oversight.
- Evidence of meaningful reuse or extension, which may need to be interpreted qualitatively rather than through download counts alone.

NIH should fund, either intramurally or with extramural partners, the development of repositories, secure computing, curation, storage, maintenance and personnel required for responsible sharing. Congress should support this infrastructure through dedicated appropriations that supplement, rather than compete with, funding for investigator-initiated research. Controlled-access, federated, or privacy-preserving approaches should receive appropriate recognition when unrestricted sharing would be unethical or unlawful.

NIH should also establish proportionate guidance for generative AI. Disclosure should be required when AI materially influences scientific design, data generation, analysis, results, reproducibility, or interpretation.

Sharing can be limited by participant consent, privacy requirements, cybersecurity risk, licensing terms, intellectual property, inconsistent data standards, and the continuing cost of maintaining research products. Metrics that reward volume or unrestricted openness could encourage poorly documented releases or inappropriate disclosure.

NIH should develop AI and sharing policies with scientific, technical, ethical, security, and patient input. Requirements should be harmonized across institutes and should take effect only when compliant infrastructure and funding are available.

Training and Mentorship

NIH training and career-development programs are essential to the stability and renewal of the biomedical research workforce. Individual fellowships, institutional training awards, career-development awards, protected research time, and sustained mentorship should remain central.

NIH investment in training awards and research infrastructure provides a foundation that institutions can extend through complementary programs tailored to workforce needs across career stages and fields of inquiry. Training in clinical trials, within and outside of the health care system, are important for future scientific rigor and reliability.

For example, Yale's Combined PhD Programs in the Biological and Biomedical Sciences Program coordinates PhD training across nearly all bioscience disciplines and includes numerous specialized NIH-funded training programs, complemented by NIH-supported research in affiliated laboratories. Together with YCCI's training programs, the Office of Physician-Scientist and Scientist Development, and the Yale Physician-Scientist Development Awards, these programs demonstrate a lifecycle approach spanning foundational, clinical, translational, and community-based research. YCCI reports that its programs trained 227 individuals from 2006 to 2025, with a long-standing track record for graduates subsequently receiving substantial independent research support. These outcomes demonstrate the value of long-term infrastructure rather than any single mentorship metric.

Evaluation should distinguish what can be assessed during training from outcomes that emerge years later. Near-term indicators could include:

- Quality and regularity of mentorship.
- Access to protected time, methods training, collaborative networks, and professional development.

- Progress toward individualized scientific and career goals.
- Trainee assessment of the mentoring environment.
- Development of scientific judgment, rigor, independence, and leadership.

Long-term evaluation may consider persistence in research, scientific contributions, mentorship of others, and careers across academia, government, health care, nonprofit organizations, and industry. Subsequent NIH funding should not be treated as the sole definition of success.

NIH should recognize co-mentorship, team mentorship, mentorship by research professionals, and the work required to create productive research environments. Incentives could include funded mentor effort, training supplements, bridge support during career transitions, and qualitative recognition in biosketches and progress reports.

Funding instability, limited protected time, uncertain career pathways, and increasing training costs threaten workforce development. Grant terminations, delayed awards, and the premature ending of mentored career-development support during 2025–26 have disrupted training and transitions to independence, compromising continuity in the biomedical trainee pipeline. Changes to federal graduate and professional education financing may compound these pressures, but their effects should be evaluated using national data on enrollment, completion, debt, and career outcomes. More broadly, NIH should evaluate workforce policy using these structural indicators rather than assuming that mentorship metrics can compensate for structural funding problems.

Collaboration

Modern biomedical research depends on interdisciplinary teams, shared facilities, research professionals, health systems, community organizations, government, nonprofit organizations, and industry. NIH should continue to support multi-investigator awards, research networks, centers, and collaborative infrastructure.

NIH's collaborative awards are most effective when complemented by institutional structures that help teams form, compete for support, and execute complex research programs. Yale's Office of Team Science supports large interdisciplinary proposals, reduces administrative barriers, connects investigators, and provides seed funding and expert review. YCCI similarly connects investigators across Yale's schools, Yale New Haven Health, the VA, community organizations, industry, and other CTSA hubs. These models demonstrate that productive collaboration requires coordination and infrastructure, not simply a larger number of investigators.

NIH should recognize substantive contributions rather than relying primarily on principal-investigator status, authorship position, publication counts, or the number of collaborators. Standardized contribution statements could identify responsibility for:

- Scientific conception and leadership.
- Methods, statistics, data, software, and model development.
- Clinical-trial execution and site leadership.
- Research operations, regulatory work, and participant engagement.
- Core-facility and research-professional expertise.
- Implementation, dissemination, and community partnership.

Contribution information should be entered once, certified by the team, and reused in progress reports and biosketches. It should inform expert review rather than produce another numerical score.

Federal awards should continue to be made through accountable institutions. Institutions provide human-subject protections, financial stewardship, cybersecurity, shared facilities, training, data governance, and continuity that cannot be supplied by an individual investigator alone. Recognition of individual contributions should not come at the expense of the institutional infrastructure that makes team science possible.

Current systems can undervalue site leadership, technical expertise, research operations, and collaborative contributions that do not correspond to conventional authorship. Short funding horizons and fragmented administrative requirements can also make durable partnerships difficult. NIH's proposed cap on simultaneous Research Project Grants should distinguish single-PI leadership from MPI participation and avoid counting collaborative awards in ways that discourage investigators from contributing specialized expertise to large team-science efforts. Policies that destabilize reimbursements for facilities and administrative expenses or substitute political pre-issuance review for expert scientific judgment could weaken the infrastructure and trust on which collaboration and discovery depends.

Entrepreneurship and Translation

NIH should continue supporting a translational ecosystem that includes investigator-initiated discovery, NCATS and other translational programs, SBIR/STTR, academic-industry research, licensing, startup formation, and clinical and implementation partnerships. These mechanisms address different stages and risks. Industry contributes development, regulatory, manufacturing, and commercialization expertise, while federal funding remains necessary for foundational discovery, early validation, and scientific or public needs without an immediate commercial return.

Yale's experience illustrates several complementary pathways to Entrepreneurship and Translation. Craig Crews' academic chemical-biology research, including NIH-supported work advancing PROTAC (proteolysis-targeting chimera) technology, led to university licensing, startup formation, sponsored research, industry partnership, and ultimately FDA approval of *Vepdegestrant,* the first approved PROTAC therapy. Ranjit Bindra's work on DNA-repair vulnerabilities in IDH1/2-mutant tumors provides a complementary example: investigator-led discovery advanced through biomarker-driven early-phase trials and a multisite NCI-sponsored study, illustrating how public support can sustain translation alongside commercial development. In total, Yale research discoveries have resulted in the creation of 186 new startup companies that have collectively created over 11,000 jobs in the United States.

Structured alliances provide another model. The Yale-AstraZeneca Oncology Alliance connects academic cancer biology with industry translational and clinical capabilities. Yale's partnership with Boehringer Ingelheim combines collaborative research with internships, joint seminars, and co-mentored postdoctoral training, including a Biomedical Data Science Fellowship focused on AI-enabled drug discovery, supported by faculty mentors across multiple disciplines. These models allow academic investigators to engage industry knowledge of development needs while preserving scientific judgment. NIH should recognize stage-appropriate evidence across this continuum, including validated targets, sponsored research, shared capabilities, licenses, startups, clinical and regulatory milestones, workforce development, and eventual public benefit, rather than treating patents, revenue, or product approval as the sole measures of translation.

NIH should support the full knowledge lifecycle across its portfolio without requiring every award to span every stage. Measures should be appropriate to maturity:

- Early translation: target, mechanism, assay, or platform validation.
- Product development: readiness, regulatory, manufacturing, licensing, or investment milestones.
- Clinical translation: safety, effectiveness, comparative value, and usability.
- Implementation: acceptability, adoption, reach, affordability, sustainability, and real-world performance.

Patents, licenses, startups, and investment are useful indicators but not ends in themselves. NIH should also recognize clinical guidelines, public-health practices, workflow improvements, open tools, and nonproprietary interventions.

Important work may fall between funding mechanisms or lack adequate commercial incentives, including replication, comparative effectiveness, implementation, open infrastructure, and studies involving small or biomarker-defined populations. NIH should identify gaps through portfolio evidence, improve continuity

between stages, and assess review-to-award time and avoidable administrative delays. Because translational research often combines federal, industry, nonprofit, and international partners, inconsistent and needlessly burdensome disclosure requirements can delay sponsored-research agreements, data and material exchange, licensing, and progression to clinical evaluation. NIH should build on the federal common disclosure forms by harmonizing conflict-of-interest, conflict-of-commitment, and foreign-relationship requirements across agencies; clearly distinguishing gifts, research support, and scientific collaborations; and applying risk-based thresholds tied to overcommitment, duplication, financial conflict, or national-security concerns. International activities presenting no material risk should not trigger duplicative reporting or delay. Greater speed would come from removing administrative processes that do not substantively elevate risk or improve scientific quality, stewardship, or participant protection, not from weakening expert review.

Foundational Scientific Exploration

Foundational science is not an exception to public impact; it is one of its principal sources. NIH's greatest contributions include discoveries whose applications could not have been predicted when the work began. An analysis of 210 drugs approved from 2010 to 2016 found NIH-funded research associated with every drug, with more than 90% of the funding supporting basic research on biological targets rather than the drugs themselves.

Artificial intelligence provides a contemporary illustration. Modern machine learning draws in part on decades of foundational neuroscience, alongside advances in mathematics, statistics, and computer science, much of which preceded any clear biomedical application. These capabilities are now reshaping how scientists generate hypotheses, analyze complex data, design experiments, and pursue new diagnostics and treatments. NIH should sustain foundational research even when its eventual applications cannot yet be predicted.

NIH should also preserve investigator-initiated research, field-specific expert review, stable support, and explicit tolerance for scientific risk. High-risk research should not be confined to special programs. The broader portfolio must retain room for unconventional questions, rare diseases, new investigators, enabling technologies, and work that may challenge prevailing assumptions.

NIH's High-Risk, High-Reward Research programs provide one important model for protecting this space within a broader investigator-initiated portfolio. The 2024 NIH High-Risk, High-Reward portfolio included five awards involving six Yale researchers across the New Innovator, Early Independence, and Transformative Research programs. These mechanisms demonstrate how NIH can create protected pathways for unconventional ideas, cross-disciplinary research, and investigators at different career stages without requiring immediate clinical application.

Foundational proposals should be evaluated for the importance of the question, originality, scientific plausibility, quality of reasoning, learning value, and potential to create new knowledge or capability. Excessive requirements for preliminary data can exclude the very ideas that high-risk programs are intended to support.

Possible long-term portfolio indicators include:

- New mechanisms, hypotheses, or conceptual frameworks.
- Methods, tools, datasets, models, or platforms that enable subsequent work.
- Findings that overturn unsupported assumptions or redirect a field.
- Downstream use across disciplines or applications over an appropriate time horizon.
- Productive knowledge generated from rigorous studies that do not confirm the original hypothesis.
- New constructs, taxonomies, contextual understanding, or hypotheses generated through rigorous qualitative or exploratory research.

These indicators should support retrospective portfolio learning, not near-term scoring of individual investigators. Short-term outcome requirements, unstable funding, and pressure for immediately demonstrable results can narrow the range of questions investigators pursue. Research portfolios require both immediately

demonstrable results and longer-term breakthrough potential. NIH should preserve that balance through expert scientific review, without political or formulaic screens.

Public Impact

NIH's mission encompasses both the creation of knowledge and its eventual contribution to longer, healthier lives. Public impact must therefore include foundational discovery while more deliberately addressing whether mature evidence reaches the American people.

Through NCATS and the Clinical and Translational Science Awards Program, NIH has helped institutions build the durable infrastructure needed to move discoveries across the research lifecycle and into practice. YCCI's clinical and translational infrastructure, dissemination and implementation programs, health-system relationships, and community partnerships demonstrate how academic research can connect discovery with real-world practice. Its community-engagement model includes joint priority setting, research participation, dissemination, and implementation.

NIH should use disease burden, unmet need, preventability, potential reach, and scientific opportunity to inform strategic planning across portfolios, not to impose a uniform public-impact score on individual proposals.

For mature research, relevant indicators may include:

- Incorporation into clinical guidelines, decision support, or public-health practice.
- Adoption and sustained use in academic, community, rural, non-health care, and other real-world settings.
- Reach among the populations expected to benefit.
- Improvements in access, safety, quality, morbidity, mortality, or other patient-centered outcomes.
- Affordability, scalability, sustainability, and time from convincing evidence to routine use.

NIH should make implementation science and community outcomes a major component of its public-impact strategy. A proven intervention has limited public value if it remains unavailable outside specialized research settings. This requires support for pragmatic evaluation, adaptation, de-implementation of ineffective practices, implementation infrastructure, and ongoing assessment after adoption.

NIH should support learning health system partnerships and community partnerships that integrate research with routine care; use clinical, non-clinical, and operational data to generate evidence; and continuously evaluate adoption, outcomes, unintended effects, and opportunities for improvement. These partnerships can shorten the cycle from discovery to implementation, thereby improving the public's return on taxpayer dollars, while rigorously and reliably testing whether interventions work across varied real-world settings.

Public trust should be strengthened through transparent explanation of portfolio priorities, methods, uncertainties, findings, and changes made when evidence evolves, as well as by demonstration of longer-term outcomes of research. NIH should report both successes and credible null results and should involve patients, clinicians, communities, and other partners without converting public engagement into political conformity.

Fragmented funding, slow transitions between research stages, inconsistent real-world infrastructure, and limited incentives for implementation can delay the use of effective discoveries. Conversely, emphasizing immediate health outcomes too early could divert support from the foundational science on which future advances depend.

NIH should therefore judge public impact at the appropriate stage and level. It should assess individual awards on what investigators can influence and assess longer-term health, economic, and societal outcomes across programs and portfolios.

Conclusion

Yale supports a research enterprise that is more rigorous, faster, efficient and accountable while retaining the curiosity, independence, collaboration, and risk tolerance that make transformative discovery possible. Such an approach would advance Yale's commitment to creating and disseminating knowledge and our mission to promote curiosity, advance discovery, nurture scientific leaders, and improve the health of all people. Yale welcomes continued collaboration with the NIH and the American people to advance this mission.

Sincerely,



Michael C. Crair, Ph.D.
William Ziegler III Professor
Vice Provost for Research

201. Research Solutions, Inc.

I am responding: On behalf of an organization

Name: Sean C. Rife

Name of Organization: Research Solutions, Inc.

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: At present, the most commonly-used method of measuring scientific impact is with citation counts. As typically implemented, citation counts treat all citations as equivalent, in that a citation presenting contrary evidence counts the same as one corroborating it. This means existing metrics cannot identify which findings are replicated and which are contested — precisely the information a replication initiative needs.

To facilitate this kind of measurement, citation classification, in which the text surrounding each citation is automatically classified as supporting, contrasting, or mentioning a given claim can be used. This is not a speculative proposal, as it has already been implemented by scite.ai. The methodology is presented in Nicholson et al. (2021; see attached PDF), and over 1.6 billion citation statements from the scholarly literature have already been classified. These data are available through a free API and can also be obtained through a bulk export though arrangements with Research Solutions, Inc. NIH could pilot these indicators on a defined collection of publications (e.g., from a given institution) and compare against expert judgment before broader adoption.
This approach directly addresses misaligned incentives within scientific institutions. Rewarding independently corroborated findings as opposed to raw citation volume does much to align incentives with rigor rather than visibility.

Of course, there are several limitations to this method: First, automated classification is imperfect and subject to error (see Nicholson et al. for F-scores across all citation types). Second, supporting and contrasting citations constitute a minority of all citations, which can limit the ability to derive meaningful signals from papers with a smaller number of citations. Third, any metric is potentially subject to gaming (e.g., authors deliberately using supporting or contrasting language depending on which claims they wish to elevate or demote. Given these limitations, we believe these indicators should inform, rather than replace, expert review.

However, because these indicators operate at the level of individual findings rather than cumulative output, they are less biased against early-career researchers and smaller institutions than other (e.g., h-index-style metrics), provided appropriate time windows and field-normalization techniques are properly applied.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

Description: Paper describing methods for classifying citations

202. Aleksandra Gruslova & Lindsey Macpherson

I am responding: On behalf of myself

Name: Aleksandra Gruslova & Lindsey Macpherson

Name of Organization: The University Of Texas at San Antonio

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Faculty Success Fellows in Response to NIH Request for Information: Measuring and Rewarding Public Impact in Biomedical Research (NOT-OD-26-087)
The comments that follow reflect the perspective of faculty with expertise in STEM fields, including biomedical engineering and neuroscience. They do not represent institutional policy, rather, they are informed by the observations of Faculty Success Fellows at The University of Texas at San Antonio, who participated in an university initiative exploring how faculty define, recognize, and demonstrate Public Impact Research (PIR). Through this collaborative effort, we examined how scientific contributions evolve over the course of faculty careers and how current evaluation systems capture—or fail to capture—the broader value of research.
One theme has emerged from our project: researchers already create public value in many different ways, but these contributions are often not fully recognized by current evaluation systems. Public impact is rarely a single event or an immediate outcome; rather, it often develops gradually over the course of a scientific career as investigators mature, build expertise, establish collaborations, mentor others, and contribute to the broader research ecosystem. Over time, faculty members may generate impact through advances in research methods, development of shared scientific resources, contributions to public policy, education and workforce development, collaborative networks, community engagement, and efforts that strengthen public trust in science.
These following comments below provide our collective observations and offer suggestions in response to NIH's Request for Information on assessing the public impact of biomedical research, with the goal of helping NIH better recognize, encourage and reward the full spectrum of scientific contributions that ultimately benefit society over the long term.
Comment 1: Definition of Public Impact Research, Recognition of Multiple Pathways to Impact, and Long-

Term Scientific Contributions.

A definition of PIR is needed that clearly distinguishes it from translational research while recognizing its important connection to fundamental discovery. Although translation into clinical practice, commercialization, and implementation are important pathways for improving health, they represent only one component of the broader biomedical research ecosystem. Public impact in biomedical research should be defined as research that contributes to improving human health, societal well-being, scientific capability, or evidence-based decision making.

This definition should recognize that public impact can arise from many forms of research contributions, including advancing fundamental knowledge that enables future discoveries, developing new methods, datasets, software, and shared research infrastructure, improving clinical care and public health, informing policy and regulatory decisions, supporting education and workforce development, promoting open science, and strengthening public understanding and trust in science. Research that produces these outcomes may have substantial public value even when the impact is indirect, cumulative, or realized over a longer timeframe.

NIH should also recognize that the public benefit of biomedical research arises through different pathways. For some investigators, impact may be demonstrated through new therapies, diagnostics, or improvements in patient care. For others, it may occur through computational tools, engineering approaches, standardized protocols, research resources, policy guidance, or community-engaged approaches that advance health. These diverse contributions are essential to NIH's mission and should not be evaluated solely through a commercialization or clinical implementation lens.

At the same time, recognizing public impact should not diminish the importance of fundamental, discovery-driven research. Many transformative advances in medicine have originated from basic science discoveries whose eventual applications were not foreseeable at the time of discovery. Fundamental research is often the foundation upon which future clinical and public health advances are built, and the absence of immediate application should not be interpreted as a lack of public value. The definition of PIR should therefore recognize both near-term contributions and long-term scientific investments that enable future improvements in health.

A flexible definition that accounts for differences across biomedical disciplines—including basic science, clinical research, computational biology, biomedical engineering, population health, and community-engaged research—will better capture the full range of ways NIH-supported research advances health and benefits society.

Comment 2: Metrics for Assessing and Recognizing Public Impact in Biomedical Research.

Because public impact is multidimensional, NIH should implement an assessment approach that recognizes multiple forms of evidence rather than relying on a single metric or score. Evaluation of public impact should complement, rather than replace, traditional measures of scientific excellence, such as research quality, rigor, and scholarly contributions. The goal should be to provide a broader view of research value while allowing flexibility across disciplines, research areas, and career stages.

NIH could encourage investigators to document evidence of public impact through an "Impact Portfolio" that captures contributions appropriate to their research program. Such evidence may include, but should not be limited to, open science practices, transparent reporting, sharing of protocols, software, datasets, and other research resources, development of community tools, contributions to standards and guidelines, engagement with patient or community groups, and participation in activities that inform public health policy or government decision making.

Importantly, not every investigator should be expected to demonstrate every type of public impact. The

most meaningful contributions will differ depending on the scientific field, the goals of the research, and the investigator's career stage. Assessment frameworks should therefore allow multiple pathways for demonstrating impact rather than creating incentives that favor only highly visible or immediately measurable outcomes.

NIH could incorporate consideration of public impact across the research lifecycle, including grant applications, peer review, progress reports, competing renewals, investigator recognition programs, and training or career development activities. However, public impact should not be measured primarily through popularity-based indicators such as social media engagement, media attention, or online visibility, as these do not necessarily reflect scientific significance, rigor, or lasting contribution. A balanced approach that considers meaningful evidence of impact will better recognize the diverse ways NIH-supported research contributes to improving health and advancing biomedical science.

Comment 3: Reward the Ecosystem - not Only the Individual Investigator

One important lesson from our Faculty Success Fellow discussions is that public impact rarely results from the efforts of a single investigator or from a single research outcome. Rather, meaningful impact develops gradually over the course of a scientific career as investigators mature, build expertise, and contribute to the broader research ecosystem. While individual researchers make critical contributions to scientific discoveries that improve health, fundamental change often depends on a broader ecosystem that includes regulatory pathways, health system adoption, industry partnerships, public policy, institutional resources, and meaningful engagement with communities. Public impact should therefore be viewed not only as an individual achievement, but also as the outcome of a collaborative research environment.

NIH can promote greater public impact by supporting institutions in building the infrastructure and culture needed for research translation and broader societal benefit. This includes investments in faculty development, interdisciplinary collaboration, mentorship, community partnerships, data-sharing and research infrastructure, and institutional policies that recognize diverse forms of scientific contribution across the research lifecycle. Supporting researchers as they develop expertise, build collaborative networks, and contribute to shared scientific resources may generate greater long-term societal benefit than evaluating investigators solely through individual outcomes. A public impact framework that recognizes both individual contributions and the institutional ecosystems that enable them will better reflect how biomedical advances are developed, sustained, and translated into meaningful benefits for society.

Conclusion: Scientific discoveries improve society through many interconnected pathways. Some lead directly to new therapies, diagnostics, or clinical applications, while others create the foundational knowledge, research tools, infrastructure, policies, and trained workforce that enable future advances. A meaningful framework for Public Impact Research should recognize this full continuum of contributions while maintaining rigorous standards for scientific quality and evidence. NIH has an opportunity to lead the research community by adopting a multidimensional approach that values diverse pathways to impact, strengthens the biomedical research ecosystem, and maximizes the long-term public benefit of federally supported research. Such an approach should remain evidence-based, nuanced, and designed to recognize meaningful contributions without creating incentives that favor visibility or easily measured outcomes over lasting scientific value.

8. Other Areas: N/A

203. Bradley E. Alger, Ph.D.

I am responding: On behalf of myself

Name: Bradley E. Alger, Ph.D.

Name of Organization: University of Maryland School of Medicine, Department of Pharmacology and Physiology

Type of Organization: Healthcare Institution

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Replicability-Comments-Aug-18-2026.docx

Description: Comments in response to request for input from community about rewarding replicability

Comments on “how to measure and reward scientific impact and replicable research practices?”

Bradley E. Alger, Ph.D.

I. Efforts must begin with a few definitions and clarifications.

I. A. What exactly do we mean by “replication” (or “reproducibility, I will use the terms interchangeably)?” Four definitions or types of reproducibility have been identified by the American Society for Cell Biology (ref. 1, for discussion). Briefly, they are:

1. Analytical – reanalyze experimental dataset and attempt to confirm conclusions.
2. Direct – repeat original experiment as exactly as possible and compare results.
3. Systematic – repeat original experiment, but systematically alter one (or more) variable to test generality of conclusions
4. Conceptual – tests previous conclusions using different methods, variables, etc.

Notice that definitions I.A.1 and IA.2 genuinely refer to attempts to confirm or disconfirm original reports, in other words to replicate them.

Definitions I.A.3 and I.A.4 should not be considered “replicability” issues at all because their focus is not on reproducing the original experiments and data, but on the testing the conclusions or inferences that were drawn from the those experiments.

Definitions I.A.3 and I.A.4 are actually about testing hypotheses that arise from, are suggested by, or are consistent with, the original conclusions. If they are not replicated (or confirmed, reproduced, etc.) it may well mean that a hypothesis has been tested and found to be incorrect, “falsified.” Testing and falsifying hypotheses are integral aspects of the scientific method, they are key making scientific progress, not problems to be done away with.

It is important to make such distinctions in order to reduce the confusion often present in discussions of replicability.

I.B. Modern scientific papers typically include many different experiments—ideally they are all related to a central idea. Nevertheless, they are different experiments. If one is not confirmed, does that mean the whole paper is “irreproducible”? Do all of the sub-experiments have to be reproduced in order to conclude that the paper is indeed “reproducible?”

I.C. It is necessary to realize that despite the great importance of reproducibility for science, a single failed attempt at replication of a previous finding is, on its own, generally uninformative:

1. the first study may be wrong and the second right or
2. the second study wrong and the first right or
3. both are wrong or
4. both are right.

The history of science includes numerous examples of I.C.3. or I.C.4. being the case.

I.D. Given the continued rapid growth in the numbers of scientific papers published annually, particularly when most are multi-part reports (see I.B. above), attempts to ensure reproducibility will become increasingly out of reach.

For these reasons and more, a perfect solution to the reproducibility problem is not in the cards.

II. Some suggestions for improving the situation include:

II.A. A definition of "reproducibility" should be developed and provided by a respected scientific committee to address the issues raised under I.A. and I.B. This should help remove some of the confusion present in discussions of reproducibility and focus attempts to improve true reproducibility.

II.B. Given the impossibility of attempting to replicate every experiment in a paper, authors could be required/encouraged to identify explicitly the key finding in their paper. The purpose would be to pinpoint the experiment or experiments that are crucial to its main conclusions. If that experiment could not be replicated, then the authors would acknowledge that either their conclusion(s) were in error, or that some greater problem existed.

II.B.1. Papers that are successfully replicated in follow up publications should be distinguished by a sign or "badge" highlighting that fact.

II.B.2. The widespread practice of announcing in a paper's abstract that such-and-such an issue is "not well understood," as motivation for the paper is generally uninformative—presumably the reader already assumes that. In place of this vague indication of the general topic, authors should be encouraged to be more explicit about the gap(s) in knowledge that they aimed to fill.

II.B.3. Most significantly for improving reproducibility. this practice would provide an explicit focal point for reproducibility attempts. Rather than having to decide how to proceed to check an important result, follow up studies would be directed at the experimental result that the original authors themselves deem to be the main one.

This practice would encourage/require authors to think and communicate more explicitly and clearly about the key significance of their work than they often do.

It would clarify for the community what conclusions could or could not legitimately be drawn from the results obtained.

II.C. The remarks in I.A. imply that the significance of reproducibility or replicability is not the same for every study. For many "kinds" of science— "curiosity-driven, "open-ended questioning," "discovery science," etc., the issue is simply whether or not the original measurements were accurate, thorough, etc. A failure of reproducibility can often be corrected by improved, higher-resolution, or simply greater numbers of measurements. These are critically important matters for scientific validity but normally do not involve major matters of scientific reasoning or thinking.

II.C.1. Definitions I.A.3. and I.A.4. directly involve the broader issues of the generality or interpretations of experimental results. Addressing these matters usually involves strategies that are lumped under the rubrics of "hypothesis-testing, or hypothesis-driven," science. This category is fundamentally different from the others. The purpose of a hypothesis is to *explain* data or other observations. It tries to account for why they are what they are.

Hypothesis-testing experiments are not *opposed* to the other kinds of science. Rather, hypothesis-testing fundamentally depends on the existence or experimental results or observations that need to be explained. Not infrequently the results of an initial "curiosity-driven" effort leads naturally into hypothesis testing.

The hypothesis is an inescapable part of scientific procedure. Efforts to characterize a field give rise to prompt attempts to explain the data collected. This does not mean that every paper or project must be explicitly aimed at testing a hypothesis. Observations, curiosity, discovery science, questioning, etc.—will always play a vital role. These investigations provide the bedrock information necessary for science. They answer the questions: "what?" "where?" "how much?" "when?" etc.

In contrast, the answer to the question, "why?" depends on reasoning and interpretation, not bare observation or measurement, no matter how motivated. Attempts to say "why" begin with one or more, suspected, proposed, or putative, explanations. These cognitive constructs cannot truly be tested "directly" by an experiment or observation of the world. Instead, each proposed explanation, if true, predicts outcomes of experiments or observations that *must* be true if the hypothesis is true. Otherwise, the hypothesis is false.

II.C.2. These basic points are intimately intwined with the replicability problem.

Falsification of hypotheses is often considered a "null result." Null results are often considered "negative data," and not published.

A recent report in *The Journal of Reproducibility in Neuroscience*[2] on the impact of various "questionable research practices," identifies "not publishing null results," as "perhaps the most problematic" contributor to the "false discovery rate," which is a major driver of irreproducibility.

II.C.3. Encouraging and rewarding the publication of “null results” will therefore help improve the reproducibility of published science. Recognizing the often-hidden existence of hypotheses as inherent components of investigations (as suggested by the discussion in I.A.3. and I.A.4.) can foster this practice in several ways:

a. It will decrease the stigma attached to “null results” and “negative data” by calling attention to the value of such results as tests that can falsify a hypothesis. The process of hypothesis proposal and falsification is a major factor in scientific progress.

b. As a falsification of a significant hypothesis, “null” results will find more ready acceptance by journals.

c. This will motivate investigators to undertake reproducibility studies by rewarding their efforts with journal publications and their associated benefits to their CVs, grant and promotion prospects etc.

Conclusions: The problem of scientific reproducibility is vast and will not be solved by any single recommendation or change. More comprehensive solutions will include improved statistical practices, greater funding levels, alterations in the reward systems of science and many more. Nevertheless, clarification of the reproducibility problem and related matters of scientific reasoning and communication can help alleviate its detrimental effects.

References

1. Alger, BE (2019) *Defense of the Scientific Hypothesis: From Reproducibility Crisis to Big Data* (Oxford; New York) Chp. 7, The Reproducibility Crisis.

2. Kravitz DI, Mitroff SR (2023) Quantifying and correcting for, the impact of questionable research practices on false discovery rates in psychological science. Journal for Reproducibility in Neuroscience, 5:e44.

204. AUTM- Association of University Technology Managers, Inc.

I am responding: On behalf of an organization

Name: Stephen J. Susalka

Name of Organization: AUTM - Association of University Technology Managers, Inc.

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings. NIH also seeks comments on rigor, including with regard to defining causal research.

First, validated patient recruitment biomarkers can lead to more focused clinical trials, thereby enabling more reproducible research.

Patient recruitment biomarkers have been shown to increase the probability of success, thus making pre-clinical research and earlier clinical trial results more likely to be reproduced in later stage results. A 2018 paper (Estimation of clinical trial success rates and related parameters) finds: "... trials that use biomarkers in patient-selection have higher overall success probabilities than trials without biomarkers."

A 2021 report ("Clinical Development Success Rates and Contributing Factors 2011–2020") confirms: "Indications are increasingly segmented by biomarkers in order to match patients with the treatments most likely to show the greatest benefit, according to the underlying drug mechanism and disease pathophysiology. Drug development programs with trials employing patient preselection biomarkers have a two-fold higher LOA (likelihood of approval) (15.9%) than those that do not (7.6%)."

Second, AUTM supports incentivizing discovery and publication of biomarkers by i) restoring broad patent eligibility as the Patent Eligibility Restoration Act (PERA) is intended to do, and ii) creating and using regulatory and funding incentives for data publication and sharing.

With respect to restoring broad patent eligibility, AUTM's prior comments summarize our position:
• August 16, 2023 letter to Senators Tillis and Coons: https://autm.net/AUTM/media/About-AUTM/Documents/AUTM_PERA_Support_August_2023.pdf
• September 2024 letter, joint with the Association of American Universities (AAU), The Association of Public Land Grant Universities (APLU) and ACE (American Counsil on Education) to the Senate Judiciary Committee: https://autm.net/AUTM/media/About-Tech-Transfer/Documents/AAU-Patent-Bill-Markup-Letter.pdf
• August 29, 2025 letter to Senators Thom Tillis and Senator Adam Schiff: https://autm.net/AUTM/media/About-Tech-Transfer/Documents/Support-for-PERA-2025-FINAL.pdf

Third, NIH could create incentives, such as funding sources, for confirmatory studies, publication of negative results, and transparent and enabled method sections.

Finally, with respect to research using electronic health records containing protected health information (PHI), AUTM suggests using synthetic datasets to enable replication of results, as discussed in Section II.e. of AUTM's Comments on Intellectual Property Protections for Artificial Intelligence Innovation (Docket No. PTO-C-2019-0038).

Creating regulatory and funding incentives for data publication and sharing are discussed below.

2. Data, Software, and Model Sharing: Sharing scientific data and tools accelerates biomedical research discovery, in part by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies. NIH seeks to incentivize and reward these activities to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.

To accomplish these goals, NIH could:

(A) Incentivize data sharing by creating databases for unique or important datasets and providing in-kind access to researchers. This would promote growth of the databases and encourage data sharing.

Examples of such special data sets include i) normative pediatric databases containing, for example, normal ranges of certain parameters for pediatric populations, ii) validation sets, such as the protein structure database used to help train and validate protein folding predictions (Protein Structure Prediction Center), and iii) authoritatively labeled and annotated databases, such as the OPTIMAM Mammography Image Database.

(B) Incentivize data sharing via regulatory incentives, such as priority review vouchers for sharing de-identified clinical trial data. Disclosure of de-identified secondary outcome data could also be incentivized. Often, secondary outcomes include biomarkers, which may be useful in subsequent clinical trials.

(C) Incentivize data sharing by encouraging publication of data structures or Electronic Relational Diagrams (ERDs) of interest. Known data structures could aid interinstitutional data sharing or multi-team collaborations as the databases would be inherently easier to combine.

(D) Convene stakeholders to foster consensus on best practices for data management. AUTM has previously commented on certain similarities between biological materials and data. (See paragraph III.c. of AUTM's Comments on Intellectual Property Protections for Artificial Intelligence Innovation (Docket No. PTO-C-2019-0038). AUTM noted that each resource requires time and commitment to produce and maintain, and there is a desire to share it with others, typically with conditions such as receiving appropriate credit. As data becomes an increasingly important research resource, it is timely for stakeholders to share their experience managing such resources, both commercially and for research purposes.

(E) Advise researchers on the most up-to-date best practices for i) data management, and ii) de-identification of PHI.

(F) Incentivize/ease data sharing via licensing by allowing managers of curated data to claim a sui generis data right in line with the EU Database Directive, which requires that such databases meet the conditions of i) systematic arrangement, ii) substantial investment, and iii) that the investment went toward obtaining, verifying, or presenting the data (and not the costs to create the data itself, e.g. cost of clinical trials). Such a sui generis right can be time limited, in which case extending the permitted period of the sui generis right can become one regulatory incentive. See in particular the discussion of data management in Section III of AUTM's Comments on Intellectual Property Protections for Artificial Intelligence Innovation (Docket No. PTO-C-2019-0038).

(G) To the extent that datasets provide support for patented inventions, the USPTO could allow the use of synthetic datasets to show possession and demonstrate enablement of such inventions. (See AUTM's Comments in Response to the United States Patent and Trademark Office's Request for Comments Titled "Unlocking the Full Potential of Intellectual Property by Translating More Innovation to the Marketplace" (Docket No. PTO-C-2024-004).

3. Training and Mentorship: NIH recognizes that advancing its mission requires promoting the growth and stability of the biomedical research workforce. While NIH supports training through existing policies and programs, additional indicators are needed to assess and quantify the impacts of these efforts. For example:

(A) The NIH can track and publish statistics on results of researchers utilizing entrepreneurs-in-residence and similar programs. How many lead to that researcher forming a start-up and engaging with industry, and how many of those engagements ultimately resulted in a new product?

(B) The NIH could ask for feedback from participating researchers to learn which aspects they found most helpful.

(C) The NIH could ask researchers to document interactions with industry they attribute to NIH programs, such as the NIH SEED program. These connections can be counted or mapped using Pajek or network diagrams to show the growth of connections across sectors from the NIH to industry.

4. Collaboration: As biomedical research becomes increasingly interdisciplinary and convergent, collaboration is growing in both scale and complexity. NIH seeks new approaches for assessing and recognizing contributions to team-based science. AUTM suggests:

(A) To streamline collaborations, retain the Bayh-Dole framework in all agreements, including Other Transaction Agreements (OTAs).

(B) Make federal funding available for multi-institutional or team projects committed to developing i) a sharable consensus on data structures and ii) data sharing.

(C) Create an ORCID system for multi-researcher and multi-institutional teams. The acronym ORCID stands for “Open Researcher and Contributor ID” and refers to individual people. Researchers use this throughout their career to show the impact of their work as evidenced by citations to it. There is no equivalent for multi-researcher and multi-institutional teams.

Looking for publications or other outputs associated with multi-researcher and multi-institutional team can be challenging because of spelling variations, misspellings, use of different abbreviations, or name changes. Teams could name themselves and register for a team ORCID to enhance their visibility.

However, AUTM cautions against overvaluing near-term metrics, such as publication counts, as these are not always reliable and dispositive indicators of the long-term public benefits we all want.

5. Entrepreneurship and Translation: To advance the translation of discoveries into actual products for people, NIH fosters the maturation of discoveries from bench to bedside and promotes innovation across sectors. More can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission. For example:

(A) Committing stable SBIR/STTR funding is crucial, as these are important vehicles for partnerships between universities and industry.

(B) NIH could encourage researchers to work with Entrepreneurs in Residence (EiRs), other business consultants, or mentors through industry programs or workshops. As part of the program/workshop, industry partners could offer access to proposal review or potential investment through licensing or sponsored research.

(C) NIH could host workshops on AI-assisted invention, customer discovery, and business development.

(D) Using data in iEdison, NIH could track and publish aggregated data on the percentage of grants leading to new products, and time elapsed between invention, SBIR/STTR funding, and product introduction.
For additional recommendations, see AUTM’s December 22, 2025 Response to the Notice of Request for Information; Accelerating the American Scientific Enterprise (No. 2025-21150) (https://autm.net/AUTM/media/About-Tech-Transfer/Documents/OSTP-RFI-Response-by-AUTM-Dec-2025.pdf).

6. Foundational Scientific Exploration: Basic research and high-risk, high-reward initiatives represent foundational scientific exploration with an inherent risk of failure and the potential for paradigm-shifting breakthroughs. Further embedding these approaches within the academic enterprise is essential to ensuring continued support for bold and innovative research directions.

AUTM supports predictable and stable grant funding. AUTM supports and enables public-private partnerships between organizations with different and complementary skill sets. Most basic research in the U.S. is performed by nonprofits, such as the universities, academic medical centers, and research institutes that comprise AUTM’s membership. Most applied research and experimental development is performed by industry. We believe that basic research and high-risk, high-reward initiatives are well

embedded within the academic enterprise and look forward to continuing our role in America's vibrant innovation ecosystem.

7. Public Impact: NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.

It is more straightforward to assess causation in focused, tightly controlled circumstances, such as lab experiments, or well-designed clinical trials, than in highly complex, interconnected systems, such as the public impact of innovation ecosystems, which is the subject of the last question in this RFI.

The AUTM data referenced above clearly show demand for public-private partnerships and provide information on the economic impact of these public-private partnerships (https://autm.net/AUTM/media/SurveyReportsPDF/AUTM-2025-US-Infographic-Fillable.pdf). However, the data focus on license agreements and lack details regarding product development timelines.

iEdison, however, may provide an opportunity to track invention and product development timelines and outcomes by funded project or by funding theme. The iEdison data include information on the entity type taking the first license, the names of the licensed products and the numbers sold, and the number of sublicenses granted. This data could be made available to select researchers for analysis of factors contributing to research programs producing visible outcomes, such as publications and patents, attracting public-private partnerships, and products on the market. Still, with respect to causal inference, such analysis should be interpreted with caution, since there is often more than one funding source per invention, and more than one invention per product. Additionally, the applied research and experimental development conducted by industry partners needed to bring products to market is an essential "cause" of the product being on the market.

Macroeconomic impacts of innovation are well documented, including notably in the 2025 Nobel Prize in Economic Sciences (https://www.nobelprize.org/prizes/economic-sciences/2025/press-release/), though the cited work was not limited to biomedical innovations and encompassed innovation overall. The first paper on "A Model of Growth Through Creative Destruction" was published in 1992.

The Bureau of Economic Analysis (BEA) identifies contributions to GDP by industry (https://www.bea.gov/data/gdp/gdp-industry). It is possible to isolate and compare over time, if desired, contributions to GDP by groups of North American Industry Classification System (NAICS) coded industries, such as research-intensive industries, chemical industries, or pharmaceutical industries. These data identify trends and relationships among industries, and shed light on factors contributing to a healthy ecosystem. The entire system of national accounts was built over decades, starting in 1947. The practical work needed to identify research-intensive industries started in 1992 and was deemed, by international consensus, complete enough to start using in 2008. See "The Evolving Treatment of R&D in the U.S. National Economic Accounts."

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Response-to-RRI-on-Measuring-and-Rewarding-Scientific-Impact-AUTM-Final_081826-1.pdf

Description: For you conviencence, please see AUTM's attached comments in response to the Request for Information on Measuring and Rewarding Scientific Impact, Notice Number: NOT-OD-26-087 Sincerely, Stephen J. Susalka AUTM CEO

August 13, 2026

Response to Request for Information on Measuring and Rewarding Scientific Impact Notice Number: NOT-OD-26-087

Introduction

AUTM is the non-profit leader in technology transfer, representing more than 3,000 technology transfer professionals at universities, research centers, and hospitals. Our members turn basic scientific discoveries into the life-saving medicines, transformative technologies, and advanced computing engines of tomorrow.

AUTM is committed to enabling and implementing well-informed, transparent technology transfer and intellectual property policies and practices for public benefit. AUTM started collecting information about the public benefits of technology transfer in 1991 and continues to do so today. The AUTM database (https://autm.net/surveys-and-tools/databases/statt/new-accessing-statt-database) includes information on research funding, invention disclosures, patents, licenses, start-ups formed, and license revenue received, including royalty income from product sales.

The first AUTM Public Benefits Survey was conducted in 1994. It tabulated 177 products associated with 440 license agreements and grouped them by type of license (exclusive or nonexclusive), by type of business (small business or large business), by state, and by responding institution. AUTM continues to gather data on licensed technologies, which it publishes annually, and also collects university innovation narratives via its Better World Project (https://autm.net/about-tech-transfer/better-world-project).

Because this RFI seeks comments on rigor and causal research, it is helpful to note the widely acknowledged Bradford-Hill criteria for causal inference used in the scientific process, and randomized clinical trials in particular. See "The Environment and Disease: Association or Causation?" (available at https://pmc.ncbi.nlm.nih.gov/articles/PMC1898525/). The identified criteria include temporality, strength, consistency, and specificity. Moreover, there should be a plausible mechanism explaining how the cause leads to the effect, and the mechanism should be consistent with prior knowledge.

This methodical perspective has been used broadly in biology and economics. With the surge in the amount of available data, a "cause" is often a combination of causes. The effects—whether good health, a robust economy, or scientific impact—evolve over years, if not decades. Many causes interact with each other and are interdependent. Therefore, it is important not to isolate and overvalue any one factor and to keep long-term effects in mind.

(1) Rigor and Reproducibility. Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings. NIH also seeks comments on rigor, including with regard to defining causal research.

First, validated patient recruitment biomarkers can lead to more focused clinical trials, thereby enabling more reproducible research.

Patient recruitment biomarkers have been shown to increase the probability of success, thus making pre-clinical research and earlier clinical trial results *more likely to be reproduced in later stage results.* A 2018 paper (Estimation of clinical trial success rates and related parameters) finds: "... trials that use biomarkers in patient-selection have higher overall success probabilities than trials without biomarkers."

A 2021 report ("Clinical Development Success Rates and Contributing Factors 2011–2020") confirms: "Indications are increasingly segmented by biomarkers in order to match patients with the treatments most likely to show the greatest benefit, according to the underlying drug mechanism and disease pathophysiology. Drug development programs with trials employing patient preselection biomarkers have a two-fold higher LOA (likelihood of approval) (15.9%) than those that do not (7.6%)."

Second, AUTM supports incentivizing discovery and publication of biomarkers by i) restoring broad patent eligibility as the Patent Eligibility Restoration Act (PERA) is intended to do, and ii) creating and using regulatory and funding incentives for data publication and sharing.

With respect to restoring broad patent eligibility, AUTM's prior comments summarize our position:

- August 16, 2023 letter to Senators Tillis and Coons: https://autm.net/AUTM/media/About-AUTM/Documents/AUTM_PERA_Support_August_2023.pdf
- September 2024 letter, joint with the Association of American Universities (AAU), The Association of Public Land Grant Universities (APLU) and ACE (American Counsil on Education) to the Senate Judiciary Committee: https://autm.net/AUTM/media/About-Tech-Transfer/Documents/AAU-Patent-Bill-Markup-Letter.pdf
- August 29, 2025 letter to Senators Thom Tillis and Senator Adam Schiff: https://autm.net/AUTM/media/About-Tech-Transfer/Documents/Support-for-PERA-2025-FINAL.pdf

Third, NIH could create incentives, such as funding sources, for confirmatory studies, publication of negative results, and transparent and enabled method sections.

Finally, with respect to research using electronic health records containing protected health information (PHI), AUTM suggests using synthetic datasets to enable replication of results, as discussed in Section II.e. of AUTM's Comments on Intellectual Property Protections for Artificial Intelligence Innovation (Docket No. PTO-C-2019-0038).

Creating regulatory and funding incentives for data publication and sharing are discussed below.

(2) Data, Software, and Model Sharing. Sharing scientific data and tools accelerates biomedical research discovery, in part by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies. NIH seeks to incentivize and reward these activities to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.

To accomplish these goals, NIH could:

(A) Incentivize data sharing by creating databases for unique or important datasets and providing in-kind access to researchers. This would promote growth of the databases and encourage data sharing.

Examples of such special data sets include i) normative pediatric databases containing, for example, normal ranges of certain parameters for pediatric populations, ii) validation sets, such as the protein structure database used to help train and validate protein folding predictions (Protein Structure Prediction Center), and iii) authoritatively labeled and annotated databases, such as the OPTIMAM Mammography Image Database.

(B) Incentivize data sharing via regulatory incentives, such as priority review vouchers for sharing de-identified clinical trial data. Disclosure of de-identified secondary outcome data could also be incentivized. Often, secondary outcomes include biomarkers, which may be useful in subsequent clinical trials.

(C) Incentivize data sharing by encouraging publication of data structures or Electronic Relational Diagrams (ERDs) of interest. Known data structures could aid interinstitutional data sharing or multi-team collaborations as the databases would be inherently easier to combine.

(D) Convene stakeholders to foster consensus on best practices for data management. AUTM has previously commented on certain similarities between biological materials and data. (See paragraph III.c. of AUTM's Comments on Intellectual Property Protections for Artificial Intelligence Innovation (Docket No. PTO-C-2019-0038). AUTM noted that each resource requires time and commitment to produce and maintain, and there is a desire to share it with others, typically with conditions such as receiving appropriate credit. As data becomes an increasingly important research resource, it is timely for stakeholders to share their experience managing such resources, both commercially and for research purposes.

(E) Advise researchers on the most up-to-date best practices for i) data management, and ii) de-identification of PHI.

(F) Incentivize/ease data sharing via licensing by allowing managers of curated data to claim a sui generis data right in line with the EU Database Directive, which requires that such databases meet the conditions of i) systematic arrangement, ii) substantial investment, and iii) that the investment went toward obtaining, verifying, or presenting the data (and *not* the costs to create the data itself, e.g. cost of clinical trials). Such a sui generis right can be time limited, in which case extending the permitted period of the sui generis right can become one regulatory incentive. See in particular the discussion of data management in Section III of AUTM's Comments on Intellectual Property Protections for Artificial Intelligence Innovation (Docket No. PTO-C-2019-0038).

(G) To the extent that datasets provide support for patented inventions, the USPTO could allow the use of synthetic datasets to show possession and demonstrate enablement of such inventions. (See AUTM's Comments in Response to the United States Patent and Trademark Office's Request for Comments Titled "Unlocking the Full Potential of Intellectual Property by Translating More Innovation to the Marketplace" (Docket No. PTO-C-2024-004).

(3) Training and Mentorship. NIH recognizes that advancing its mission requires promoting the growth and stability of the biomedical research workforce. While NIH supports training through existing policies and programs, additional indicators are needed to assess and quantify the impacts of these efforts. For example:

(A) The NIH can track and publish statistics on results of researchers utilizing entrepreneurs-in-residence and similar programs. How many lead to that researcher forming a start-up and engaging with industry, and how many of those engagements ultimately resulted in a new product?

(B) The NIH could ask for feedback from participating researchers to learn which aspects they found most helpful.

(C) The NIH could ask researchers to document interactions with industry they attribute to NIH programs, such as the NIH SEED program. These connections can be counted or mapped using Pajek or network diagrams to show the growth of connections across sectors from the NIH to industry.

(4) Collaboration. As biomedical research becomes increasingly interdisciplinary and convergent, collaboration is growing in both scale and complexity. NIH seeks new approaches for assessing and recognizing contributions to team-based science. AUTM suggests:

(A) To streamline collaborations, retain the Bayh-Dole framework in all agreements, including Other Transaction Agreements (OTAs).

(B) Make federal funding available for multi-institutional or team projects committed to developing i) a sharable consensus on data structures and ii) data sharing.

(C) Create an ORCID system for multi-researcher and multi-institutional teams. The acronym ORCID stands for "Open Researcher and Contributor ID" and refers to individual people. Researchers use this throughout their career to show the impact of their work as evidenced by citations to it. There is no equivalent for multi-researcher and multi-institutional teams.

Looking for publications or other outputs associated with multi-researcher and multi-institutional team can be challenging because of spelling variations, misspellings, use of different abbreviations, or name changes. Teams could name themselves and register for a team ORCID to enhance their visibility.

However, AUTM cautions against overvaluing near-term metrics, such as publication counts, as these are not always reliable and dispositive indicators of the long-term public benefits we all want.

(5) Entrepreneurship and Translation. To advance the translation of discoveries into actual products for people, NIH fosters the maturation of discoveries from bench to bedside and promotes innovation across sectors. More can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission. For example:

(A) Committing stable SBIR/STTR funding is crucial, as these are important vehicles for partnerships between universities and industry.

(B) NIH could encourage researchers to work with Entrepreneurs in Residence (EiRs), other business consultants, or mentors through industry programs or workshops. As part of the program/workshop, industry partners could offer access to proposal review or potential investment through licensing or sponsored research.

(C) NIH could host workshops on AI-assisted invention, customer discovery, and business development.

(D) Using data in iEdison, NIH could track and publish aggregated data on the percentage of grants leading to new products, and time elapsed between invention, SBIR/STTR funding, and product introduction.

For additional recommendations, see AUTM's December 22, 2025 Response to the Notice of Request for Information; Accelerating the American Scientific Enterprise (No. 2025-21150) (https://autm.net/AUTM/media/About-Tech-Transfer/Documents/OSTP-RFI-Response-by-AUTM-Dec-2025.pdf).

(6) Foundational Scientific Exploration. Basic research and high-risk, high-reward initiatives represent foundational scientific exploration with an inherent risk of failure and the potential for paradigm-shifting breakthroughs. Further embedding these approaches within the academic enterprise is essential to ensuring continued support for bold and innovative research directions.

AUTM supports predictable and stable grant funding. AUTM supports and enables public-private partnerships between organizations with different and complementary skill sets. Most basic research in the U.S. is performed by nonprofits, such as the universities, academic medical centers, and research institutes that comprise AUTM's membership. Most applied research and experimental development is performed by industry. We believe that basic research and high-risk, high-reward initiatives are well embedded within the academic enterprise and look forward to continuing our role in America's vibrant innovation ecosystem.

(7) Public Impact. NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.

It is more straightforward to assess causation in focused, tightly controlled circumstances, such as lab experiments, or well-designed clinical trials, than in highly complex, interconnected systems, such as the public impact of innovation ecosystems, which is the subject of the last question in this RFI.

The AUTM data referenced above clearly show demand for public-private partnerships and provide information on the economic impact of these public-private partnerships (https://autm.net/AUTM/media/SurveyReportsPDF/AUTM-2025-US-Infographic-Fillable.pdf). However, the data focus on license agreements and lack details regarding product development timelines.

iEdison, however, may provide an opportunity to track invention and product development timelines and outcomes by funded project or by funding theme. The iEdison data include information on the entity type taking the first license, the names of the licensed products and the numbers sold, and the number of sublicenses granted. This data could be made available to select researchers for analysis of factors contributing to research programs producing visible outcomes, such as publications and patents, attracting public-private partnerships, and products on the market. Still, with respect to causal inference, such analysis should be interpreted with caution, since there is often more than one funding source per invention, and more than one invention per product. Additionally, the applied research and experimental development conducted by industry partners needed to bring products to market is an essential "cause" of the product being on the market.

Macroeconomic impacts of innovation are well documented, including notably in the 2025 Nobel Prize in Economic Sciences (https://www.nobelprize.org/prizes/economic-sciences/2025/press-release/), though the cited work was not limited to biomedical innovations and encompassed innovation overall. The first paper on "A Model of Growth Through Creative Destruction" was published in 1992.

The Bureau of Economic Analysis (BEA) identifies contributions to GDP by industry (https://www.bea.gov/data/gdp/gdp-industry). It is possible to isolate and compare over time, if desired, contributions to GDP by groups of North American Industry Classification System (NAICS) coded industries, such as research-intensive industries, chemical industries, or pharmaceutical industries. These data identify trends and relationships among industries, and shed light on factors contributing to a healthy ecosystem. The entire system of national accounts was built over decades, starting in 1947. The practical work needed to identify research-intensive industries started in 1992 and was deemed, by international consensus, complete enough to start using in 2008. See "The Evolving Treatment of R&D in the U.S. National Economic Accounts."

Conclusion

Some research questions can be focused, such that it is possible to observe causes and effects. Systems, such as the economic innovation systems noted in this response evolve over time as a result of many interactions. In due course, they reveal trends and groups of factors contributing to a particular outcome.

Some of these suggestions, such as more and better managed data sharing, are likely to produce near-term improvements. Others, such as a continuing commitment to fund basic research, take longer to produce public impact.

Thank you for the opportunity to respond to this RFI, and we welcome the opportunity to work with you on ways to measure and grow the impact of federally funded scientific research.

Sincerely,



Stephen J. Susalka
AUTM CEO

205. Kevin Jon Williams, M.D.

I am responding: On behalf of myself

Name: Kevin Jon Williams, M.D.

Name of Organization: Lewis Katz School of Medicine at Temple University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: The NIH should focus on the creation, preservation, and re-establishment of viable career paths for physician-scientists, often described as an endangered species.

NIH support for physician-scientist training should equally emphasize both training paths – namely, the MD-PhD route (which remains disproportionately over-emphasized in my opinion), as well as the path by which physicians complete their clinical training, without a PhD, and then undertake a research fellowship. This latter path is not as well supported financially as the MD-PhD route. Advantages of this latter path include choosing a research area when the physician is a fully trained clinician (instead of a second-year medical student) and being able to stick with the research without interruption (instead of years away from a focus on research to complete clinical training, which starts at the end of the PhD period of an MD-PhD program and lasts through the third and fourth years of medical school, then internship and residency, and then a clinical subspeciality fellowship).

Moreover, the NIH's current emphasis on basic science disfavors physician-scientists. Full-time PhD scientists have inherent advantages under the current system, such as no clinical duties to compete for their time and fewer clinically related interests to distract them from what is most likely to be funded, e.g., minutiae in animal and cell-culture models.

4. Collaboration: Big Science is not always the answer. The NIH should not neglect the individual investigator with a small laboratory.

One approach would be a formal split in funding amongst large, medium, and small laboratories or groups, so that allocations to big vs small science become a matter of policy rather than mere preferences of study section members.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: For foundational scientific exploration, I would again emphasize that the NIH should adopt policies to explicitly support and protect individual investigators with small laboratories.

7. Public Impact: Regarding your request for "specific examples of measurable indicators", NIH-supported biomedical research has a terrible problem with entrenched Big Ideas that underperform, or do not perform, for human health. The measurable indicators would be preventive strategies, medicines, and/or diagnostic tests that arise from these ideas.
Often, there are none, even after decades of support. Yet study sections still lavishly support these ideas. It's a self-reinforcing loop that relies on "friend groups" whose unelected members use their power to direct taxpayer dollars to fund each other. Based on this extensive record, I no longer believe that peer review can solve this problem.

Here is a possible solution. Grant proposals, existing extramural grants, and intramural projects studying ideas that have received US taxpayer support for ≥20 years, with little or no clinical benefit, should be subjected to special scrutiny – but outside of peer review. Ideas that are 10-19 years old, with little or no clinical benefit, should also be re-evaluated, but less stringently. Existing literature supports these thresholds, e.g., “Promising technologies that did not lead to a published human study within 10 to 12 years were unlikely to be tested in humans subsequently.” (doi: 10.1016/s0002-9343(03)00013-5)

Of course, persistent lavish funding for under- or non-performing Big Ideas means less support, or no support, for approaches that might actually improve human health.

Here are relevant citations that contain specific examples of the problem:

1. Charlton BG. Zombie science: a sinister consequence of evaluating scientific theories purely on the basis of enlightened self-interest. Med Hypotheses. 2008;71:327–329. doi: 10.1016/j.mehy.2008.05.018.
2. Nicholson JM, Ioannidis JPA. Research grants: conform and be funded. Too many US authors of the most innovative and influential papers in the life sciences do not receive NIH funding. Nature. 2012;492:34–36. doi: 10.1038/492034a.
3. Joyner MJ, Paneth N, Ioannidis JPA. What happens when underperforming big ideas in research become entrenched? JAMA. 2016;316:1355–1356. doi: 10.1001/jama.2016.11076. Open-access at https://www.nature.com/articles/492034a .
4. Williams KJ. Inflammation in atherosclerosis: a Big Idea that has underperformed so far. Curr Opin Lipidol. 2025;36:78-87. doi: 10.1097/MOL.0000000000000973. Open access hypertext at https://journals.lww.com/co-lipidology/fulltext/9900/inflammation_in_atherosclerosis__a_big_idea_that.106.aspx ; open-access PDF at https://pmc.ncbi.nlm.nih.gov/articles/PMC11888836/pdf/colip-36-78.pdf (which has remarks on the citations that were highlighted with bullets).
5. Williams KJ. Anti-inflammatory agents in atherosclerosis—and a need for reform: Extraordinary claims require extraordinary evidence. J Intern Med. 2026;300:258-283 (ePub 2026 July 12). doi: 10.1111/joim.70128. Open-access at https://onlinelibrary.wiley.com/doi/10.1111/joim.70128 .

References #4 and #5 deal with the Big Idea that inflammation must be a definitive therapeutic target in atherosclerotic arterial disease, the biggest killer in the USA and worldwide. My search today of the NIH RePORTER database for "inflammation" and "atherosclerosis" identified over 16,000 grants with those keywords, spanning decades and totaling well over US$6,500,000,000. PubMed lists over 34,000 articles with those keywords. Reference #5 listed at least 57 (fifty-seven) published human clinical trials of anti-

inflammatory agents in this disease, with no signs of slowing.

And yet, there is nearly nothing to show for human health. There is just one anti-inflammatory medication – colchicine – with FDA approval for patients with atherosclerotic arterial disease. But cardiologists are not using it because the most recent clinical trials of this agent showed no benefit or even harm (see ref. #5). Reference #5 suggests that these new trial data should prompt the FDA to re-examine its cardiovascular indications for colchicine.

The field of inflammation in atherosclerosis has produced just one approved diagnostic marker – serum levels of CRP – but CRP adds only slight predictive power beyond conventional risk factors for ASCVD events. Far more useful are recent advances in technologies to image the arterial wall, which outperform every additional marker in plasma or serum in the clinical assessment of ASCVD event risk (see ref. #4).

Many of these points are summarized in the Graphical Abstract of reference #5.

Shortly after reference #5 was published online, Novo Nordisk announced that its much-anticipated ZEUS trial of IL6 inhibition (an anti-inflammatory strategy) in patients with ASCVD, CKD, & elevated CRP levels showed no benefit. see
6. https://finance.yahoo.com/healthcare/articles/novo-nordisk-stock-tumbles-10-120558178.html and
7. https://www.nytimes.com/2026/08/06/health/heart-disease-drug-fails-novo-nordisk.html , among others.

These sentences from reference #5 seem relevant.
"How many dozen clinical trials that failed to bring a new therapy to patients in need, over how many decades, and at what personal and financial cost will it take for us to move past a concept that has produced so little for patient care? Table 1, Table S1, and this text list a total of at least 57 human clinical trials of anti-inflammatory agents in patients with atherosclerotic arterial disease, spanning more than the past third of a century, without clear benefit and with no signs of stopping. Do we need a 58th trial? A 98th?"

Well, now we have our 58th.

Moreover, 2026 is on track to being another banner year of NIH support for "inflammation" and "atherosclerosis": 400 projects so far. Every year since 2007, two decades ago, shows at least 400 NIH-supported grants with those keywords.

Reference #5 also briefly reviewed the closely related theory of LDL oxidation in atherosclerosis. That idea has also been lavishly supported by the NIH. Yet "Decades of highly funded work on that hypothesis ended in ashes: no approved antioxidant medications in atherosclerosis and no approved diagnostic tests" (ref. #5).

"Sometimes peer review fails in this way." Why are theories that come to nothing still supported?

Because no outside authority that answers directly to the voting, taxpaying public ever steps in to stop this self-serving waste.

8. Other Areas: N/A

206. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: Improving scientific rigor and reproducibility is urgent, and it requires strengthening the institutional review process. The program officer is the most important hub for tracking trends within a scientific field and for overseeing both the scientists in a portfolio and new applicants. Experienced program officers have a better understanding of the field than out-of-field reviewers on study sections, and are therefore better placed to judge fairly.

Science today cannot be assumed to run on goodwill alone. NIH needs the resolve to confront the minority of scientists who damage public trust in science and to impose real consequences. Once a scientist is flagged for failures of rigor or reproducibility, program officers can weigh that record in the next grant review or terminate an existing award. To equip program officers to judge fairly, NIH should build a system that collects and analyzes the information they need.

1. Poor data quality in publications:
Establish an office that, at a program officer's request, analyzes data quality in the publications of portfolio scientists and applicants. The problems extend well beyond duplicated images. For example, animal behavior experiments in high-profile journals frequently rely on very small sample sizes (n = 7–8). In my experience, animal behavior is highly variable, so samples of that size cannot support firm conclusions. Statistical anomalies are a further warning sign of fabrication.

2. Misconduct by PIs within their laboratories:
Establish an anonymous reporting system for research misconduct.
a) Data fabrication in published papers. PubPeer records should be incorporated as additional input.
b) PIs pressuring subordinates to produce favorable data. Complaints of this kind are heard regularly from graduate students and postdocs, including in laboratories that publish in high-profile journals. In most cases, the complainants leave the laboratory quietly.

3. False collaborations:
Inflating the number of authors in publications dilutes achievements and obscures responsibility when fraud is identified. NIH should require a fractional measure of each author's contribution, with the contributions for any publication summing to 1. The distribution should be agreed upon among the authors at the time of publication. Until such a system exists, credit should be divided equally, at 1/n per author. When a paper is retracted, all authors should be flagged equally. Critics may object that this

would discourage collaboration; in my view, genuine collaboration would be unaffected.

4. Institutional incentives favor the business of science over its quality:
University administrators pressure PIs to have graduate students complete PhD theses within three years. By the time a thesis reaches the defense committee, it is difficult to intervene even when the work is immature. The academic research enterprise is oriented toward expansion: enrolling more graduate students and securing more federal funding. Indirect cost recovery, meanwhile, appears to be reinvested in recruiting additional faculty to sustain that expansion. The anonymous reporting system should also capture institutional behavior of this kind to inform HHS re-negotiations of indirect cost rates.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

207. Griffin M Weber

I am responding: On behalf of myself

Name: Griffin M Weber

Name of Organization:

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: The NIH recently ran a Data Sharing Index (S-Index) Challenge to advance data sharing across biomedical research by developing a metric to measure, incentivize and reward sharing of high-quality data. Our team won third place in the Challenge with a project focused on engaging key stakeholder communities to promote data sharing, and to understand impact through a composite S-Index. We have attached our winning proposal. Our team members include Griffin M Weber, M.D., Ph.D. (Beth Israel Deaconess Medical Center); Katy Börner, Ph.D. (Indiana University); Amy Brand, Ph.D. (MIT); Kristi Holmes, Ph.D. (Northwestern); Daniel Hook, Ph.D. (Digital Science, Holtzbrinck); John Inglis, Ph.D. (bioRxiv, medRxiv); Chonnettia Jones, Ph.D. (Addgene); Arjun Krishnan, Ph.D. (University of Colorado Anschutz); and Steinn Sigurdsson, Ph.D. (Penn State, arXiv). Please contact Griffin Weber (weber@hms.harvard.edu) with questions. An NIH announcement about the Challenge can be found at https://www.nei.nih.gov/research-and-training/research-news/nih-announces-winners-s-index-challenge-advance-data-sharing-across-biomedical-research.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/SIndex_Challenge_Submission_GriffinWeber.pdf

Description: We have attached our winning proposal for the NIH Data Sharing Index (S-Index) Challenge.

Introduction of Team

Thank you for the opportunity to participate in the NIH Data Sharing Index Challenge. We have devoted our careers to promoting open science and facilitating the sharing of data and scientific knowledge. Together, we represent *and have the ability to convene* **key stakeholder communities** required to advance NIH's goals of incentivizing and rewarding data sharing and data reuse and driving a culture of openness, transparency, and explainability in science. We are a multidisciplinary team of experts from academia (MIT, Northwestern U, PSU, U Colorado, Indiana U); publishing (arXiv, bioRxiv, medRxiv, MIT Press, CSHL Press); research materials and data sharing (Addgene); research information management systems (Digital Science, VIVO, Profiles); identity management (Crossref, ORCID, InCommon); and scientific funding (former associations with HHMI, Wellcome Trust, Health Research BC) and policy making. Our team includes inventors of widely used quantitative metrics (Dimensions, Altmetric) and data visualizations (Atlas of Science) related to scientific productivity and impact. We have attached **Letters of Support** from additional organizations that will join us in this effort, including the Gates Foundation, Novo Nordisk Foundation, Sloan Foundation, HHMI Janelia, Navigation Fund, DataCite, Data.org, Make Data Count, CRediT, Invest in Open Infrastructure, Zenodo, GREI, FASEB, Accountability in Research, CTSA Eval., NNLM-NEC, and university librarians. (A list of acronyms used in this document can be found in **Table S1**.)



Katy Börner, PhD
Professor of Engineering & Information Science, Indiana University; Atlas of Science; Human Reference Atlas
Data analysis and visualization

Amy Brand, PhD
Director and Publisher of the MIT Press; previously on Boards of Crossref, ORCID; Faculty Promotion and Tenure
Linguistics, cognitive science

Kristi Holmes, PhD
Assoc. Dean, Northwestern; Advisor: HELIOS Open, Make Data Count; InCommon, US Repository Network; Zenodo GREI lead
Evaluation, data repositories, libraries

Daniel Hook, MBA, PhD
CEO, Digital Science (Figshare, Altmetric, Symplectic, Dimensions, GRID); Research on Research Institute (RoRI)
Quantum statistical physics

John Inglis, PhD
Co-founder, bioRxiv, medRxiv; Founding Exec Dir., CSHL Press; previously *The Lancet Immunology*

Chonnettia Jones, PhD
Executive Director, Addgene; formerly HHMI, Wellcome Trust, Health Research BC and Chan Zuckerberg Initiative Open Science Advisory Board member
Research funding, policy & assessment

Arjun Krishnan, PhD
Assoc. Professor of Biomedical Informatics, Center for Health Artificial Intelligence, University of Colorado Anschutz
Biomedical informatics, machine learning, AI

Steinn Sigurdsson, PhD
Scientific Director, arXiv; Professor of Astrophysics, Pennsylvania State University
Theoretical physics

Griffin Weber, MD, PhD *(Team Captain)*
Assoc. Professor of Biomedical Informatics, Beth Israel Deaconess Medical Center; Profiles; Electronic Health Records (i2b2)
Medicine, clinical data, computer science

Engaging key stakeholder communities to promote data sharing and reuse and to understand impact through a composite S-Index

Introduction

Each member of our team has made significant contributions to advancing openness, transparency, and explainability in science. We understand the importance of this to performing high quality reproducible research, accelerating new discoveries, and understanding the meaningful impact of research. Even so, we also recognize that there is a need for creative new ways of incentivizing researchers to not only share data, but also to increase the "value" of their data (e.g., ensuring high-quality data, detailed metadata, and adherence to FAIR principles). The specific task of this NIH Challenge is to propose a person-level metric that rewards sharing high-value datasets. However, as the Landscape Report leading to this Challenge showed, defining a data-sharing index, or "S-Index", is not enough. We believe the larger impact of this Challenge will be bringing key stakeholder communities in the NIH Data Sharing and Scholarly Reward Ecosystem ("Data Ecosystem") together to catalyze a real culture change in which researchers, their research organizations, funders, editors, and the general public develop a greater appreciation for the scholarly contributions of sharing and reusing high value datasets, on par with publishing in high-impact journals. We start with a conceptual framework representing different components of a shared dataset's value and an initial composite S-Index prototype. Through a series of stakeholder workshops, we will refine the framework and its individual quantitative and qualitative measurements, which may be weighted differently based on context, and are linked to clear data sources. We will work with stakeholders on a phased implementation to integrate the S-Index into key stages in the data lifecycle. Maintaining trust with stakeholders throughout this process will be critical for success.

Conceptual Innovation

What makes our model unique? Our model uses a **participatory approach** that engages stakeholders in the co-development and implementation of the S-Index to ensure that it is understandable, usable, and meaningful (**Table 1**). Close collaboration with stakeholders has been essential to the widespread adoption of open science innovations and the S-Index will be no different. An important part of this is **trust**, because as with other metrics, there is a risk that an S-Index will have unintended negative consequences. For example, researchers must be assured that the S-Index won't favor certain disciplinary perspectives, negatively impact their promotion and tenure, or encourage competing labs to use their data without proper credit. The S-Index cannot achieve its intended utility, impact, and cultural change without this trust.

Stakeholder Group	Understandable	Usable	Meaningful
Researchers and data producers	Links data sharing to impact	Integrated into familiar tools; automatically generated	Shows academic value (for P&T); leads to collaboration
Funders and policy makers	Summarizes impact in policy terms	Pulls from existing reporting & monitoring systems	Demonstrates innovation & health impact; informs strategy
Data repositories and infrastructure providers	Measures usage/reuse clearly	Embedded in analytics/metrics; uses existing data processes & PIDs	Informs service; enhances metadata
Healthcare providers	Indicates real-world clinical relevance	Highlights trusted datasets; requires no additional input	Supports evidence-based practice
Patients and the public	Shows data; demonstrates clear health outcomes	No action required; openly available; presented in accessible formats	Builds trust & consent; accountability to taxpayers
Publishers, journals, aggregators	Connects datasets to publications, research objects	Integrated into editorial & peer review workflows	Enhances transparency & integrity
Universities and research organizations	Shows cross-institutional collaboration; links to patents	Can be aggregated at the person, department, or organization level	Enables benchmarking; helps promotion & tenure decisions

Table 1. Factors that will drive adoption of the S-Index while maintaining stakeholder trust.

How does it differ from existing methods? Our S-Index is a composite metric based on a **conceptual framework** (**Table 2**) that considers four distinct facets of the data: (1) *datasets as objects* that can be viewed, downloaded, or cited; (2) the *metadata* describing the quantity of

data and characteristics that indicate its "FAIRness"; (3) the *contents* of the data in terms of quality and informativeness; and (4) the broader *context* that shapes the impact and reputation of the data. Each facet has components with quantitative measures. These are complemented with additional qualitative assessments of the *context*.

Facet	Component	Measurements	Sources	Examples
1: Datasets as individual objects	Usage	Number of datasets, views, downloads	GREI Repositories	Figshare; Zenodo
	Reuse	Number of literature citations	Publishers and aggregators	Dimensions, Other aggregators
		Number of other citations	Internet, workspaces	Media mentions (Altmetric), Code (GitHub), Patents (IFI)
2: Metadata	Quantity	Number of records, features, bytes of data	Analysis of the metadata	A.I. & Machine Learning, NLP entity recognition, Knowledge graphs, Network analysis, Large language models
	FAIRness	Findable, Accessible, Interoperable, Reusable		
3: Contents	Quality	Machine readable, completeness, plausibility	Analysis of the dataset	
	Informativeness	Information content, entropy		
4: Context	Impact & Reputation	Supporting and contrasting citations	Analysis of other datasets	
		Derived datasets and other works		
		Qualitative assessment	Health, social, economic context	

Table 2. Conceptual framework of a composite S-Index.

How do these methods enhance the evaluation of data sharing? Although the composite S-Index is a single person-level metric, our framework and prototype emphasize transparency, trust, and explainability by providing easy access to the individual components, their measurements, and their underlying data sources. It makes the factors that contribute to the value of a dataset easier to understand; it allows the relative weights of these components to be adjusted by discipline, data type, use cases, or objectives; and it lowers the chance of misuse.

What creative methods are involved? We will host a series of **Stakeholders' Workshops** in Washington, DC, with attendees drawn from the wide range of stakeholder communities that our team represents (see our **Letters of Support** for examples), and other submitting teams to the S-Index Challenge. This will be critical to gaining acceptance of the S-Index, refining its framework, promoting its use, encouraging the sharing of high value data, and communicating meaningful impact. The NIH is considering an initial public event after the completion of the Challenge. In that case, our workshops will build on that momentum by working with the NIH to reconvene stakeholders regularly throughout the S-Index implementation process.

Validation and Testing

How will you validate and test your S-Index? We implemented as our main interpretable data sharing metric a **"GREI-Dimensions" Prototype**, available at (https://DataSharingIndex.org, password "sindex2024"). It aggregates datasets across more than 1,620 repositories, leveraging existing infrastructure and data made available by Daniel Hook, the CEO of Digital Science, whose portfolio of companies include **Figshare**, **Dimensions**, and **Altmetric**; and Kristi Holmes, **Zenodo** lead on the **NIH Generalist Repository Ecosystem Initiative** project (**GREI, Figure S1**) and Make Data Count advisor (**Figure S2**). This prototype demonstrates two facets of our composite S-Index framework, measuring ***reuse*** of datasets by adding both direct citations (to the datasets) and indirect citations (to the datasets' publications) and controlling for ***context*** by normalizing these citation counts by field and year.

However, stakeholder groups have different perspectives on what high value data sharing means and it will be essential to work with them in refining the S-Index. In line with this participatory approach, we have also implemented an **"Alternative PubMed" Prototype** (https://DataSharingIndex.org/Alt) to demonstrate to stakeholders the flexibility of our framework and to provide concrete examples of additional measurements that can be added to the final **consensus composite S-Index**. The prototype calculates a weighted sum of nine measurements representing each of the facets and components of our framework. To make these easy to explore, replicate, and validate, the "datasets" in this prototype are journal articles

in PubMed with associated Data Bank records in their metadata. The top Data Banks in PubMed are ClinicalTrials.gov, figshare, Dryad, and GenBank. The individual metrics include an h-index-like score (number of articles with data and at least that many citations); whether the articles are in PubMed Central (FAIRness); the number of other datasets the article cites (provenance, integration of existing data); relevance to human health; and others. All the data for these alternative measures are available at https://DataSharingIndex.org/Data.

Have you included preliminary results? Our GREI-Dimensions Prototype calculates an S-Index for 349,446 authors with 34,780,590 datasets; 1,690,784 direct citations; 132,517,596 indirect citations. The top fields of the datasets include physical sciences (61%), biological sciences (20%), engineering (16%), computer science (8%), and chemical sciences (8%). The mean S-Index is 0.677 (StDev 4.66, max 350.81). Our Alternative PubMed Prototype calculates an S-Index for 232,968 authors (with ORCID iDs) in PubMed who published 110,655 articles associated with 193,182 Data Bank records and cited 1,367,392 times. Most authors have very few data articles (mean 1.53) and 74.6% have just one (**Figure S3**). However, a composite S-Index with multiple measurements creates a clear spread (mean 18.8; StDev 6.84), which is important for comparing individuals and measuring change as they share more or higher value datasets (**Figure S4**). **Figure S5** shows the relationship between authors' h-index and S-Index.

Implementation considerations to be discussed at the Stakeholder Workshops: Publication metrics like h-index benefit from their context: there is general agreement on what a publication is; journal articles are relatively similar in length and number of references; and they are already aligned with criteria used for promotion, tenure, and professional awards and honors (providing a "gold standard" for metric validation). In contrast, there are many open questions about defining a "dataset" and how to handle certain types of data and ways of sharing data (**Table 3**).

Dataset Definitions	Assigning Credit	Policies
• Do we measure at the level of entire datasets and data collections, or at individual data records or files? • Are regular updates to an existing dataset as new data objects? • Do we count the digital objects required to use the data like software code, protocols, and codebooks?	• How do we distribute credit for massive datasets with thousands of contributors? • How do we handle the vast differences in data sizes, from KBs to PBs? • Should data with restricted access, such as hospital electronic health records, or data stored in secure enclaves be penalized?	• Should we require that data be indexed in a NIH or GREI repository? • Should specific types of data (molecular, clinical/health, behavioral, imaging, population, wearables) or data characteristics (simulated/synthetic, human/non-human, raw/processed, primary/secondary) be measured or weighted differently?

Table 3. Implementation considerations.

Feasibility

Can your idea be realistically implemented? What technologies and resources are needed? Are these attainable? As illustrated by our prototypes, our framework and composite S-Index can adapt to different data sources and measurements that can be expanded and refined over time. GREI research repositories, including Figshare and Zenodo (built on InvenioRDM), already compute basic metrics (e.g., number of shared datasets, usage statistics), and have adopted the GREI recommendations to support consistent practices to collect, expose and aggregate citations to open data (**Figure S6**). This is extended with citation data from Dimensions (**Figure S7**); open resources such as DataCite and CrossRef Event Data and EuropePMC, already integrated into Zenodo. The NLM Dataset Catalog has ingested GREI repositories.

What is the likelihood of acceptance by key stakeholders? How does the model align with current needs and contexts? Stakeholder buy-in and usability are central to the design and implementation of the composite S-Index, laying the groundwork for widespread acceptance, adoption, and impact. Our Stakeholder Workshops offer an opportunity to support active, meaningful, responsive, and sustainable collaboration with stakeholders from across the Data Ecosystem. Katy Börner, who has organized more than 40 conferences and workshops over the past decade, will lead the planning of these events.

User-Friendliness

How will you ensure the S-Index is easily accessible to researchers across disciplines and requires minimal training for effective use? Our prototypes have a simple intuitive interface that emphasizes transparency and enables exploration of the individual measurements that combine into the S-Index. We will conduct usability testing and obtain feedback on the prototypes at our Stakeholder Workshops. Our GREI-Dimensions Prototype covers all disciplines and normalizes by field. Our Alternative PubMed Prototype takes inspiration from Altmetric's use of colorful "badges" (**Figure S8**) to create one that concisely summarizes the components of the S-Index and provides context by listing the primary discipline (**Figure S9**). Help text defines each metric and lists the behaviors they are expected to encourage and their potential risks (**Figure S10**).

Scalability

Can your metric scale to handle large datasets, different types of data, formats and disciplines? Our GREI-Dimensions Prototype scans real-time internet sources using scalable infrastructure and computational methods developed by Altmetric. It normalizes by field, and our alternative prototype shows normalizing by data type. As we move beyond the prototypes we will use AI to assist in parsing metadata and examining the contents of datasets to measure their quality and informativeness. For example, Arjun Krishnan developed software that uses natural language processing and embeddings to map unstructured text descriptions of genomic data to controlled vocabularies (PMID:36347858); Griffin Weber implemented a machine learning model to characterize data completeness in electronic health records (PMID:37632234); and Kristi Holmes looked at the impact of generative AI in scholarly evaluation (in press).

Interoperability

How will your S-Index promote data standards and interoperability and integrate with existing data sharing tools, platforms and tools? The 7 GREI repositories represent some of the largest research repositories across commercial and non-profit organizations. Many universities and domain repositories use these platforms, and there will be future adoption through USRN. GREI repositories have implemented the COUNTER-based 'Make Data Count' standards that enable collection of standardized data usage (views and downloads) and data citation metrics. Key to interoperability will be machine readability; adherence to FAIR Principles; persistent identifiers; widely used metadata schemas and tools to support metadata mapping; standard formats (e.g., JSON-LD); APIs; and case study examples of successful implementations and use cases.

Adoption Strategy

What is your plan to promote the S-Index, involve the research community in adoption and use, and provide outreach and education to encourage widespread use? We have a five-year plan with a high-level goal to incentivize and reward high-value data sharing while producing a lasting cultural shift in how research outputs are assessed and valued. The phased plan will achieve (1) governance and pilots; (2) broad adoption and integration; and (3) global, trusted, FAIR-aligned, integrated open infrastructure (**Table 4**). Our team has collectively been successful in doing this across many initiatives. For example, Amy Brand helped develop governance frameworks for ORCID and CRediT. As Director of MIT Press she will coordinate outreach to journals and Stein Sigurdsson from arXiv and John Inglis from bioRxiv/medRxiv will engage with preprint servers.

Sustainability

What is your plan for long-term sustainability and potential funding partnerships? How will you ensure that the S-Index remains relevant and effective over time? The sustainability plan is aligned with the adoption strategy to ensure that the S-index remains open, trusted and adaptable to evolving research practices and community needs (**Table 4**). Chonnettia Jones is leading efforts to achieve sustainable financing through diversified funding sources. Several funders and partners have already indicated willingness to offer support (**Letters of Support**).

Phase	Adoption	Sustainability
Year 1 Governance and pilots	• Establish a governance ecosystem that builds trust, early institutional adoption, and coordination with the NIH • Present prototypes at stakeholder workshops • Reach out to journals to pilot extended metadata to better capture authors' contributions to data • Publish an article about the S-Index in a prominent journal • Identify influential academics to help us spread the word	• Achieve financing through diversified funding sources, including pooled funding from public and private funders and in-kind resources from partners • Commit a proportion of the Challenge Prize to fund knowledgeable staff with experience of managing stakeholder relationships
Years 2-3 Broad adoption and integration	• Expand to 30-50 institutions through partnerships with HELIOS Open, CoARA, others • Add S-Index to researcher profiling systems (e.g., ORCID, Profiles, Symplectic Elements) • Implement APIs for institutions, funders, and publishers • Use dashboards and benchmarking to highlight reproducibility and institutional cost-savings via data reuse	• Implement fee-for-service models, in collaboration with research institutions, libraries, infrastructure providers, and publishers, such as premium analytics and benchmarking, while upholding our collective commitment to openness and transparency
Years 4-5+ Global, trusted, FAIR-aligned, integrated open infrastructure	• Incorporate into P&T, funder reporting, and the global benchmarking ecosystem • Be on a standards pathway (similar to CRediT's NISO) • Implement real-time updating and predictive analytics. • Enable major longitudinal reports on research reproducibility and cost savings	• Establish an independent entity (inspired by the nonprofit 501(c)3 openRxiv) to manage pooled funds and structured to ensure broad stakeholder representation and to coordinate agile adaptation to changes in technology, research practices, standards

Table 4. Phased approach to adoption and sustainability.

Metrics for Success

How will you assess success and impact; track and report the performance of the S-Index and its effectiveness in promoting data sharing? Led by Kristi Holmes, we will have a clear and comprehensive approach to monitoring, reporting and evaluating the S-Index. Proposed key indicators and metrics (**Table 5**) will need input from the parties impacted by the S-Index (e.g., researchers, institutions, publishers, repositories, funders, etc.).

Success and Impact	Performance	Promoting Data Sharing
• Implementation (discoverable repositories/datasets, frictionless connection and onboarding) • Behavior (shared datasets, improved metadata quality and completeness) • Reuse outcomes (dataset citations, downloads, cross-disciplinary reuse) • Adoption (incorporation in P&T workflows) • Downstream impact (citations in policy documents, guidelines, patents)	• Monitoring the prototype interface usage and feedback • Dashboards for longitudinal context on dataset production and reuse • Quality control to ensure efficient data streams • Bias audits to monitor career stage, discipline, funders	• Increases in publications with linked datasets and higher rates of dataset citations • Improved linking of cited datasets in the PID graph to the works and people responsible • Institutional adoption of composite S-Index data in evaluation and P&T processes • Adoption demonstrated in NIH Biosketch, RPPR workflows, SciENcv

Table 5. Indicators for evaluation of the S-Index.

Impact

How will your model incentivize data sharing and impact researchers' behavior? Our framework offers stakeholders a nuanced tool to drill into **key facets of interest**, depending on need. Stakeholders' choice of weights in the composite S-Index will incentivize different behaviors. The Stakeholder Workshops will consider, for example, which components, if emphasized, would be valued in academic promotion or would encourage researchers to generate better metadata. How will your model identify, and reward widely used and impactful data resources? Our framework considers a range of different metrics to quantify use and impact. We have found that creative ways of using and visualizing metrics greatly aids in their adoption. These include adding eye-catching badges to expert finder systems (**Figure S11**), rendering dataset citation network graphs (**Figure S12**), and illustrating institution-level data sharing trends (**Figure S13**).

Conclusion

We present a conceptual framework and composite S-Index prototype that captures different facets of data sharing. However, key stakeholders must be involved in refining it and in its future implementation. They must trust both the final metric and the process that will get them there to (1) ensure its long-term impact, scalability, and sustainability and (2) lead to cultural change.

References

Open Science

1. Aspesi C, **Brand A**. In pursuit of open science, open access is not enough. Science. 2020;368(6491):574-7.
2. Borgman CL, **Brand A**. Data blind: Universities lag in capturing and exploiting data. Science. 2022;378(6626):1278-81.
3. Borgman CL, **Brand A**. The Future of Data in Research Publishing: From Nice to Have to Need to Have? Harvard Data Science Review. 2024.
4. **Brand A**, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publishing. 2015;28(2):151-5.
5. Champieux R, Solomonides A, Conte M, Rojevsky S, Phuong J, Dorr DA, Zampino E, Wilcox A, Carson MB, **Holmes K**. Ten simple rules for organizations to support research data sharing. PLoS Comput Biol. 2023 Jun 15;19(6):e1011136. doi: 10.1371/journal.pcbi.1011136.
6. Contaxis N, Clark J, Dellureficio A, Gonzales S, Mannheimer S, Oxley PR, Ratajeski MA, Surkis A, Yarnell AM, Yee M, **Holmes K**. Ten simple rules for improving research data discovery. PLoS Comput Biol. 2022 Feb 10;18(2):e1009768. doi: 10.1371/journal.pcbi.1009768.
7. Draux H, Fane B, **Hook DW**, Wastl J, Lewis P, Jones MM, Roblero P, Wilsdon J, Understanding the importance of SHAPE to the UK research ecosystem (2025) https://doi.org/10.48550/arXiv.2501.16701
8. Herzog C, **Hook DW**, Hahnel M, Konkiel S and Williams DE, Digital Science Use Cases: Enriching context and enhancing engagement around datasets in Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure (2020) Eds. Julia Lane, Ian Mulvany, Paco Nathan. ISBN: 978-1-5297-0586-7.
9. **Hook DW** & Porter SJ, AI versus IA in Artificial Intelligence in Libraries and Publishing, Eds. Ruth Pickering and Matthew Ismail. Charleston Briefings (2022) https://doi.org/10.3998/mpub.12669942
10. **Hook DW**, Hahnel M, Ian Calvert, The Ascent of Open Access, Digital Science Report (2019) https://doi.org/10.6084/m9.figshare.7618751.v2
11. **Hook DW**, Hahnel M, Open by Default in the State of Open Data, Digital Science (2016) https://doi.org/10.6084/m9.figshare.4036398.v1
12. Hosseini M, Horbach SPJM, **Holmes K**, Ross-Hellauer T. Open Science at the generative AI turn: An exploratory analysis of challenges and opportunities. Quant Sci Stud. 2025;6:22-45. doi: 10.1162/qss_a_00337. Epub 2025 Jan 27. PMID: 40124128; PMCID: PMC11928019.
13. Make Data Count. GREI recommendations to support consistent practices to collect, expose and aggregate citations to open data. Feb 6, 2024. Blog post. https://makedatacount.org/read-our-blog/grei-data-citation-recommendations/

14. Porter SJ and **Hook DW**, Recategorising research: Mapping from FoR 2008 to FoR 2020 in Dimensions, Quantitative Science Studies (2023) 4 (1): 127–143, https://doi.org/10.1162/qss_a_00244
15. Puebla, I., Neumann, S., Stathis, K., Gauthier, J., Hamelers, A., Nielsen, L. H., Belletti, L., Elliott, F., & Hahnel, M. (2024). GREI Data citation best practices for repositories. Zenodo. https://doi.org/10.5281/zenodo.10562429
16. Sharp PA, Bonvillian WB, **Brand A**, Stebbins M. The future of open research policy should be evidence based. Proc Natl Acad Sci U S A. 2024;121(32):e241268.
17. Vasilevsky NA, Hosseini M, Teplitzky S, Ilik V, Mohammadi E, Schneider J, Kern B, Colomb J, Edmunds SC, Gutzman K, Himmelstein DS, White M, Smith B, O'Keefe L, Haendel M, **Holmes KL**. Is authorship sufficient for today's collaborative research? A call for contributor roles. Account Res. 2021 Jan;28(1):23-43. doi: 10.1080/08989621.2020.1779591. Epub 2020 Jun 30. PMID: 32602379; PMCID: PMC7736357.
18. Hosseini M, Kerridge S, Allen L, Kiermer V, **Holmes K**. Enhancing Understanding and Adoption of the Contributor Roles Taxonomy (CRediT). *Learned Publishing*. (in press)

Metrics, Methods, and Evaluation

1. Herzog C, **Hook D**, Adie E, Reproducibility or producibility? Metrics and their Masters. STI Conference Proceedings (2018) https://scholarlypublications.universiteitleiden.nl/handle/1887/65257
2. **Holmes K**. (2025, April 14). Translational Science Impact. Translational Science 2025, Washington, DC. Zenodo. https://doi.org/10.5281/zenodo.15351011
3. **Holmes K**. (2022, July 12). Translating Computable Knowledge to Meaningful Action. Zenodo. https://doi.org/10.5281/zenodo.6990294
4. **Hook DW**, Porter SJ, Herzog C, Dimensions: Building Context for Search and Evaluation, Frontiers in Research Metrics and Analytics (2018) https://doi.org/10.3389/frma.2018.00023
5. **Hook DW** and Porter SJ, Scaling Scientometrics: Dimensions on Google BigQuery as an infrastructure for Large-scale analysis, Frontiers in Research Metrics and Analytics (2021) https://doi.org/10.3389/frma.2021.656233
6. Zeitlyn D, **Hook DW**. Perception, prestige and PageRank. PLoS One. 2019 May 28;14(5):e0216783. doi: 10.1371/journal.pone.0216783. PMID: 31136588; PMCID: PMC6538150.
7. Kane C, Trochim W, Bar H, Vaught A, Baker H, Khan M, Wagner R, **Holmes K**, Herzog K, Doyle JM. Navigating the road ahead: using concept mapping to assess Clinical and Translational Science Award (CTSA) program goals. Front Public Health. 2025 Mar 31;13:1562191. doi: 10.3389/fpubh.2025.1562191. PMID: 40231183; PMCID: PMC11994586.
8. Liu R, **Krishnan A**. Open Biomedical Network Benchmark, a Python toolkit for benchmarking datasets with biomedical networks. Proceedings of Machine Learning Research (2023) 240:23.
9. Make Data Count. Opening Up Translational Science Data Impact Through the Data Citation Corpus. May 13, 2025. Blog post.

https://makedatacount.org/read-our-blog/opening-up-translational-science-data-impact-through-the-data-citation-corpus/

10. Porter SJ and **Hook DW**, Connecting Scientometrics: Dimensions as a Route to Broadening Context for Analyses, Frontiers in Research Metrics and Analytics (2022) https://doi.org/10.3389/frma.2022.835139
11. Sarli CC, Dubinsky EK, **Holmes KL**. Beyond citation analysis: a model for assessment of research impact. J Med Libr Assoc. 2010 Jan;98(1):17-23. doi: 10.3163/1536-5050.98.1.008. PMID: 20098647; PMCID: PMC2801963.
12. Sarli CC, **Holmes KL**. Update to "Beyond Citation Analysis: A Model for Assessment of Research Impact". J Med Libr Assoc. 2012 Apr;100(2):82. doi: 10.3163/1536-5050.100.2.002. PMID: 22514502; PMCID: PMC3324805.
13. Yuan H, Hicks P, Ahmadian M, Johnson KA, Valtadoros L, **Krishnan A**. Annotating publicly-available samples and studies using interpretable modeling of unstructured metadata. Brief Bioinform. 2024 Nov 22;26(1). PubMed PMID: 39710433.

AI Tools for Scalability

1. Hawkins NT, Maldaver M, Yannakopoulos A, Guare LA, **Krishnan A**. Systematic tissue annotations of genomics samples by modeling unstructured metadata. Nat Commun. 2022 Nov 8;13(1):6736. doi: 10.1038/s41467-022-34435-x. PMID: 36347858; PMCID: PMC9643451.
2. Klann JG, Henderson DW, Morris M, Estiri H, **Weber GM**, Visweswaran S, Murphy SN. A broadly applicable approach to enrich electronic-health-record cohorts by identifying patients with complete data: a multisite evaluation. J Am Med Inform Assoc. 2023 Nov 17;30(12):1985-1994. doi: 10.1093/jamia/ocad166. PMID: 37632234; PMCID: PMC10654861.
3. Hosseini M, Earp B, Porsdam S, **Holmes K**. Generative Artificial Intelligence Could Create Momentum for Research Evaluation Reform. Research Evaluation. (accepted with minor revisions)

Visualizing Scientific Impact

1. **Börner K**. 2010. Atlas of Science: Visualizing What We Know. Cambridge, MA: The MIT Press.
2. **Börner K**. 2015. Atlas of Knowledge: Anyone Can Map. Cambridge, MA: The MIT Press.
3. **Börner K**. 2021. Atlas of Forecasts: Modeling and Mapping Desirable Futures. Cambridge, MA: The MIT Press.
4. **Börner K,** Bueckle A, Ginda M. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. Proc Natl Acad Sci U S A. 2019 Feb 5;116(6):1857-1864. doi: 10.1073/pnas.1807180116. PMID: 30718386; PMCID: PMC6369751.
5. **Börner K**, Elizabeth G. Record, Todd Theriault. 2025. Atlas of Macroscopes: Interactive Data Visualizations. Cambridge, MA: The MIT Press.

6. **Weber GM**. Identifying translational science within the triangle of biomedicine. J Transl Med. 2013 May 24;11:126. doi: 10.1186/1479-5876-11-126. PMID: 23705970; PMCID: PMC3666890.
7. **Weber GM**, Yuan LA. “The Power of Research Networking Systems to Find Experts and Facilitate Collaboration.” Book chapter in Strategies for Team Science Success: Handbook of Evidence-Based Principles for Cross-Disciplinary Science and Practical Lessons Learned from Health Researchers, edited by Hall KL, Vogel A, Croyle R, Springer International Publishing, 2019, pp. 541-562.

Trust, Privacy and Ethical Considerations

1. **Hook DW**, Linda S O’Brien and Stephen Pinfield, Editorial: Trust and Infrastructure in Scholarly Communications, Frontiers in Research Metrics and Analytics (2022) https://doi.org/10.3389/frma.2022.925500
2. Hosseini M, Hong S, **Holmes K**, Wetterstrand K, Donohue C, Amaral LAN, Stoeger T. Ethical considerations in utilizing artificial intelligence for analyzing the NHGRI's History of Genomics and Human Genome Project archives. J Escience Librariansh. 2024 Mar;13(1):e811. doi: 10.7191/jeslib.811. Epub 2024 Mar 5. PMID: 39552744; PMCID: PMC11566842.
3. Hosseini M, Gordijn B, Wafford QE, **Holmes KL**. A systematic scoping review of the ethics of Contributor Role Ontologies and Taxonomies. Account Res. 2024 Aug;31(6):678-705. doi: 10.1080/08989621.2022.2161049. Epub 2023 Jan 14. PMID: 36641627; PMCID: PMC11795455.
4. Murphy SN, **Weber G**, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30. doi: 10.1136/jamia.2009.000893. PMID: 20190053; PMCID: PMC3000779.
5. Yu YW, **Weber GM**. Balancing Accuracy and Privacy in Federated Queries of Clinical Data Repositories: Algorithm Development and Validation. J Med Internet Res. 2020 Nov 3;22(11):e18735. doi: 10.2196/18735. PMID: 33141090; PMCID: PMC7671849.
6. Hosseini M, Eisenman D, Riddle J, Pyle S, Peters A, Cobb N, & **Holmes K**. (2026). Guidelines needed for the use of AI in the preparation or review of IRB, IBC, and IACUC applications. *Accountability in Research*. https://doi.org/10.1080/08989621.2025.2612564

Research Organizations and Support for Adoption

1. Champieux R, Soper D, Roesch E, Hatch A, Tao D, Blake M, Puebla I, **Holmes K**, & Implementing data evaluation in academia Working Group (HELIOS Open & Make Data Count). (2025). Institutional Data Evaluation and Assessment (IDEA) Maturity Model. Zenodo. https://doi.org/10.5281/zenodo.17432837
2. Implementing data evaluation: A resource toolkit for institutions. Implementing data evaluation in academia Working Group (HELIOS Open & Make Data Count). https://makedatacount.org/read-our-blog/implementing-data-evaluation-a-resource-toolkit-for-institutions/

3. Gonzales S, Champieux R, Contaxis N, Denton AH, Hosseini M, Surkis A, Wescott A, **Holmes K**. Clinical and Translational Science Personas: Expansion and use cases. J Clin Transl Sci. 2023 Jun 8;7(1):e147. doi: 10.1017/cts.2023.572. PMID: 37456266; PMCID: PMC10346007.
4. Gutzman KE, Bales ME, Belter CW, Chambers T, Chan L, **Holmes KL**, Lu YL, Palmer LA, Reznik-Zellen RC, Sarli CC, Suiter AM, Wheeler TR. Research evaluation support services in biomedical libraries. J Med Libr Assoc. 2018 Jan;106(1):1-14. doi: 10.5195/jmla.2018.205. Epub 2018 Jan 2. PMID: 29339930; PMCID: PMC5764574.
5. **Holmes KL**, Lyon JA, Johnson LM, Sarli CC, Tennant MR. Library-based clinical and translational research support. J Med Libr Assoc. 2013 Oct;101(4):326-35. doi: 10.3163/1536-5050.101.4.017. PMID: 24163607; PMCID: PMC3794691.
6. Luo Y, Mao C, Sanchez-Pinto LN, Ahmad FS, Naidech A, Rasmussen L, Pacheco JA, Schneider D, Mithal LB, Dresden S, **Holmes K**, Carson M, Shah SJ, Khan S, Clare S, Wunderink RG, Liu H, Walunas T, Cooper L, Yue F, Wehbe F, Fang D, Liebovitz DM, Markl M, Michelson KN, McColley SA, Green M, Starren J, Ackermann RT, D'Aquila RT, Adams J, Lloyd-Jones D, Chisholm RL, Kho A. Northwestern University resource and education development initiatives to advance collaborative artificial intelligence across the learning health system. Learn Health Syst. 2024 Apr 15;8(3):e10417. doi: 10.1002/lrh2.10417. PMID: 39036530; PMCID: PMC11257059.
7. Make Data Count. Announcing Make Data Count and HELIOS Open Collaboration: 'Implementing data evaluation at institutions' Working Group. Jan 6, 2025. Blog post. https://makedatacount.org/read-our-blog/implementing-data-evaluation-in-academia-working-group/
8. Make Data Count. Leveraging data citations to respond to libraries' data evaluation needs. Jan 6, 2025. Blog post. https://makedatacount.org/read-our-blog/leveraging-data-citations-to-respond-to-libraries-data-evaluation-needs/
9. Ragon B, Seymour A, Whipple EC, Surkis A, Haberstroh A, Muilenburg J, Rethlefsen ML, Aspinall EE, Deaver J, Dexter N, Barger R, Contaxis N, Glenn EJ, Hinton E, Kern B, Little M, Pickett K, Sevetson E, Tao D, von Isenburg M, Werner DA, Wheeler TR, **Holmes K**. The role of information science within the clinical translational science ecosystem. J Clin Transl Sci. 2024 Nov 27;8(1):e224. doi: 10.1017/cts.2024.664. PMID: 39801680; PMCID: PMC11713432.

Supplementary Information and Figures

API: Application Programming Interface

arXiv: a free distribution service and an open-access archive

bioRxiv: an open-access preprint repository for the biological sciences

CoARA: Coalition for Advancing Research Assessment

COUNTER: Counting Online Usage of Networked Electronic Resources

CRediT: Contributor Role Taxonomy

CSHL: Cold Spring Harbor Laboratory Press

CTSA: Clinical and Translational Science Awards Program

EuropePMC: Europe PubMed Central

FAIR: Findable, Accessible, Interoperable, Reusable

FASEB: Federation of American Societies of Experimental Biology

GREI: Generalist Repository Ecosystem Initiative

HELIOS Open: Higher Education Leadership Initiative for Open Scholarship

HHMI: Howard Hughes Medical Institute

Indiana U: Indiana University

InvenioRDM: Invenio Research Data Management software

IOI: Invest in Open Infrastructure

JSON-LD: JavaScript Object Notation for Linked Data

MDC: Make Data Count

medRxiv: preprint server for health sciences

MIT: Massachusetts Institute of Technology

NIH: National Institutes of Health

NISO: National Information Standards Organization

NLM: National Library of Medicine

NNLM-NEC: Network of the National Library of Medicine National Evaluation Center

Northwestern U: Northwestern University

ORCID: Open Researcher & Contributor ID

PSU: Pennsylvania State University

P&T: Promotion and Tenure

RPPR: Research Performance Progress Report

SciENcv: Science Experts Network Curriculum Vitae.

U Colorado: University of Colorado Anschutz

USRN: United States Repository Network

Table S1. Acronyms used in this proposal

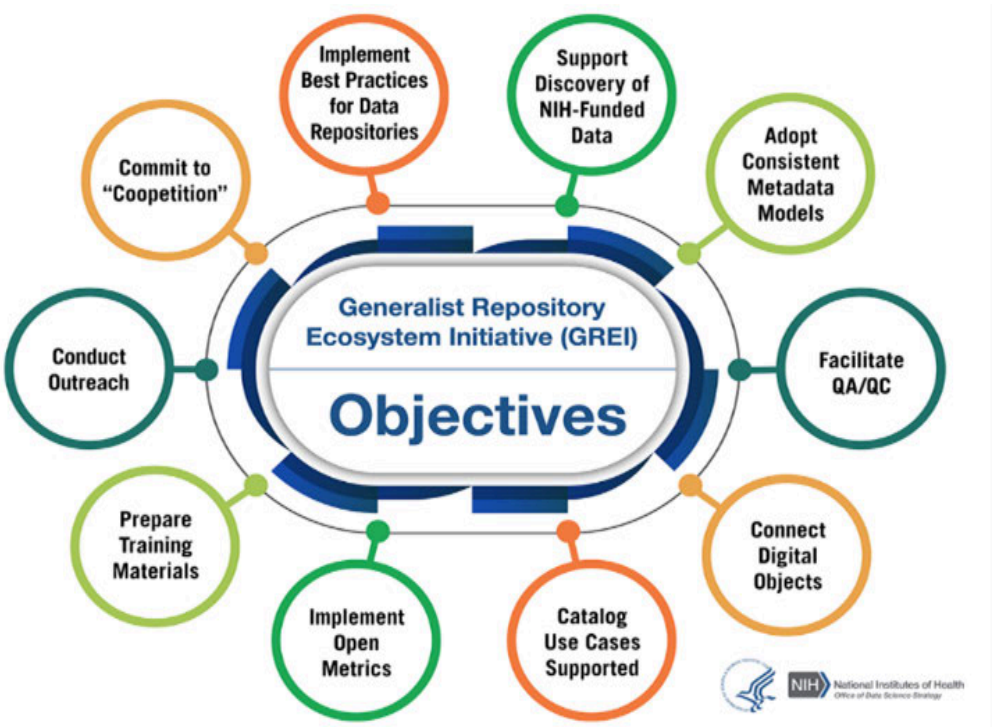


Figure S1. Vision of the NIH Generalist Repository Ecosystem Initiative (GREI) and established generalist repositories that work together to establish consistent metadata, develop use cases for data sharing, and educate researchers on FAIR data and the importance of sharing.



Figure S2. Make Data Count, a community-led initiative focused on driving transparent and meaningful open data metrics so that research data are valued research outputs across scholarly activities, evaluation, and communications.

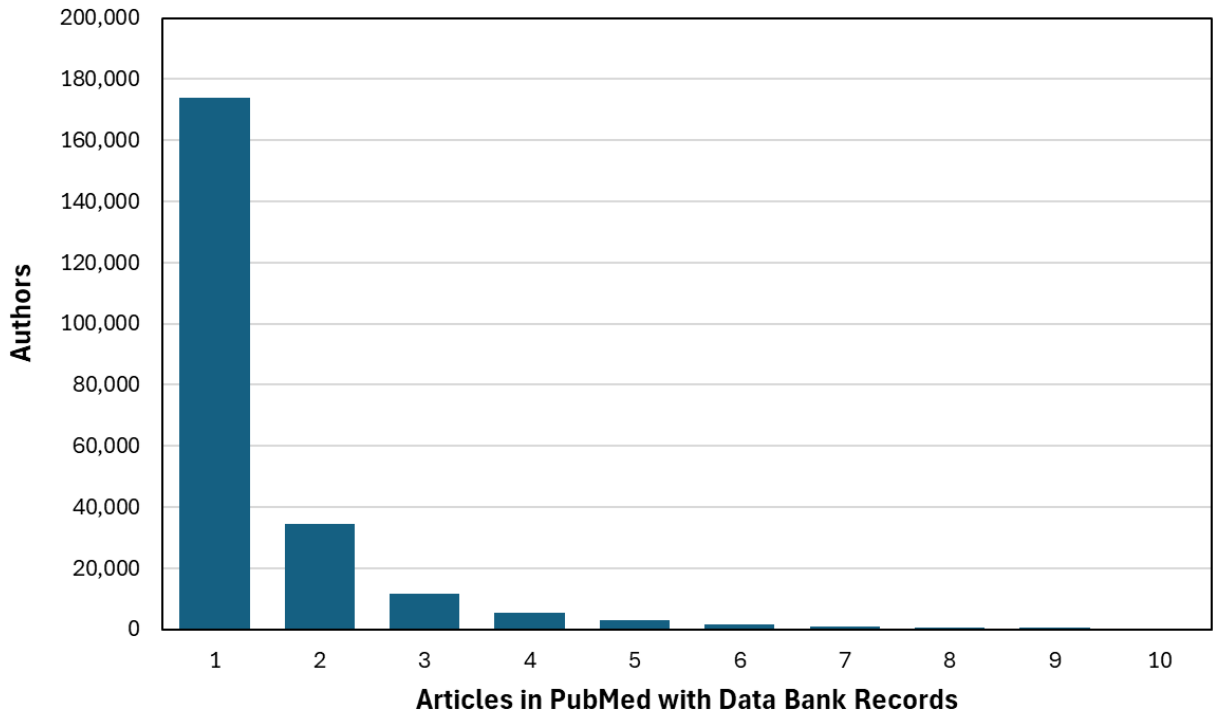


Figure S3. Distribution of the number of articles authors have published in PubMed with associated data bank records (from the Alternative PubMed Prototype). Note that 74.6% of authors have a single article, with an overall average of 1.53. As a result, an S-Index based only on “counting” metrics might show too little difference between people to be useful. A more nuanced composite metric is needed to compare the relative value of each article.

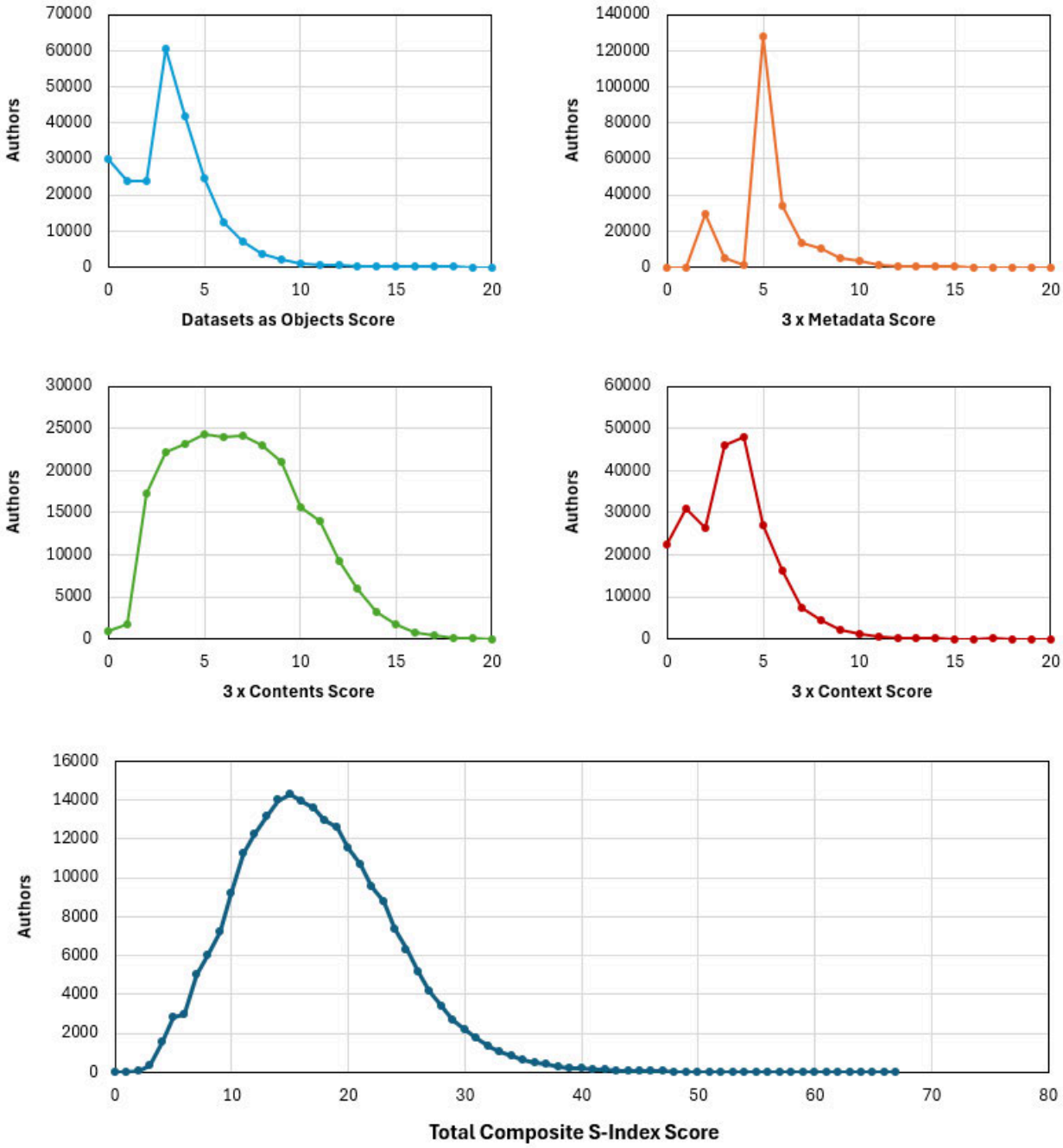


Figure S4. Distribution of S-Index values from the Alternative PubMed-based prototype. The top four graphs show the score distributions for each facet of data sharing. The bottom graph is the sum of individual scores, or the **Total Composite S-Index**. Note that a composite index creates a distribution that is smoother and with a greater spread than any facet alone. Its range is also similar to the familiar h-index (roughly 1 to 100), making it easy to interpret and remember.

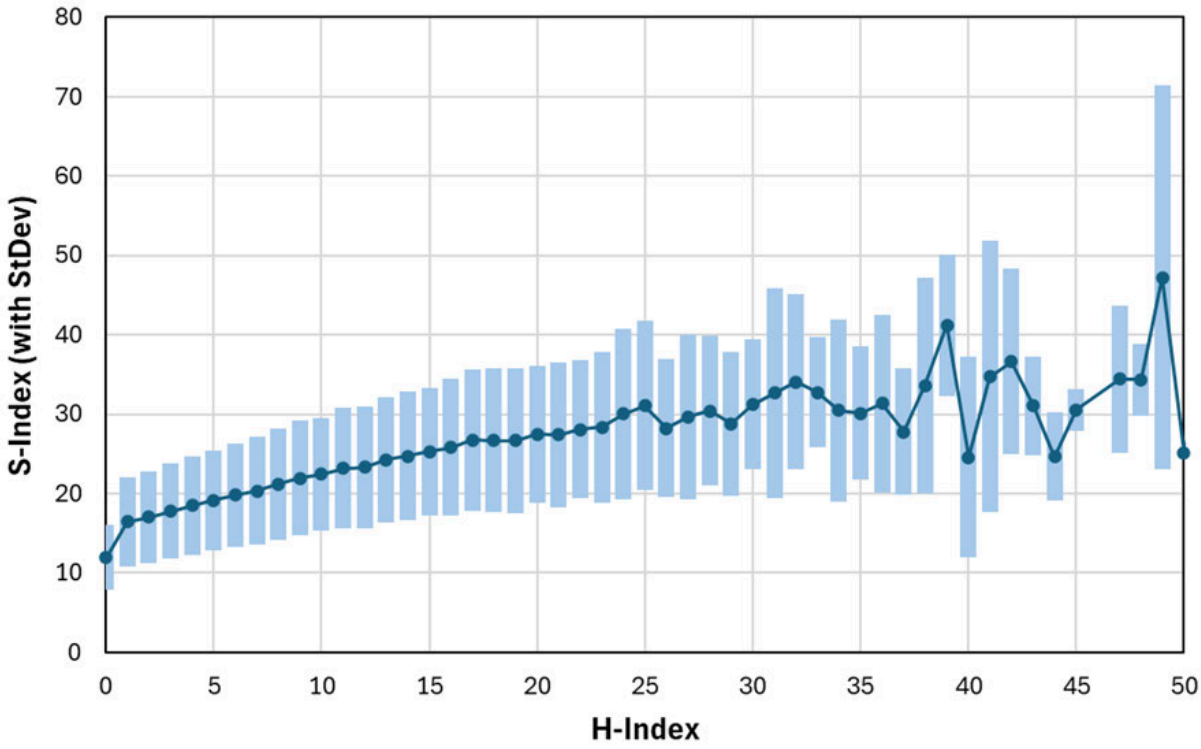


Figure S5. Relationship between authors' h-index and S-Index in the Alternative PubMed-based prototype. Bars show standard deviation. Although there is some correlation between authors' h-index and S-Index, authors with a small h-index can have an above average S-Index and vice versa. In other words, it emphasizes the need for an S-Index by demonstrating that sharing high value data is a unique characteristic of scholarly impact that isn't fully captured by the h-index.

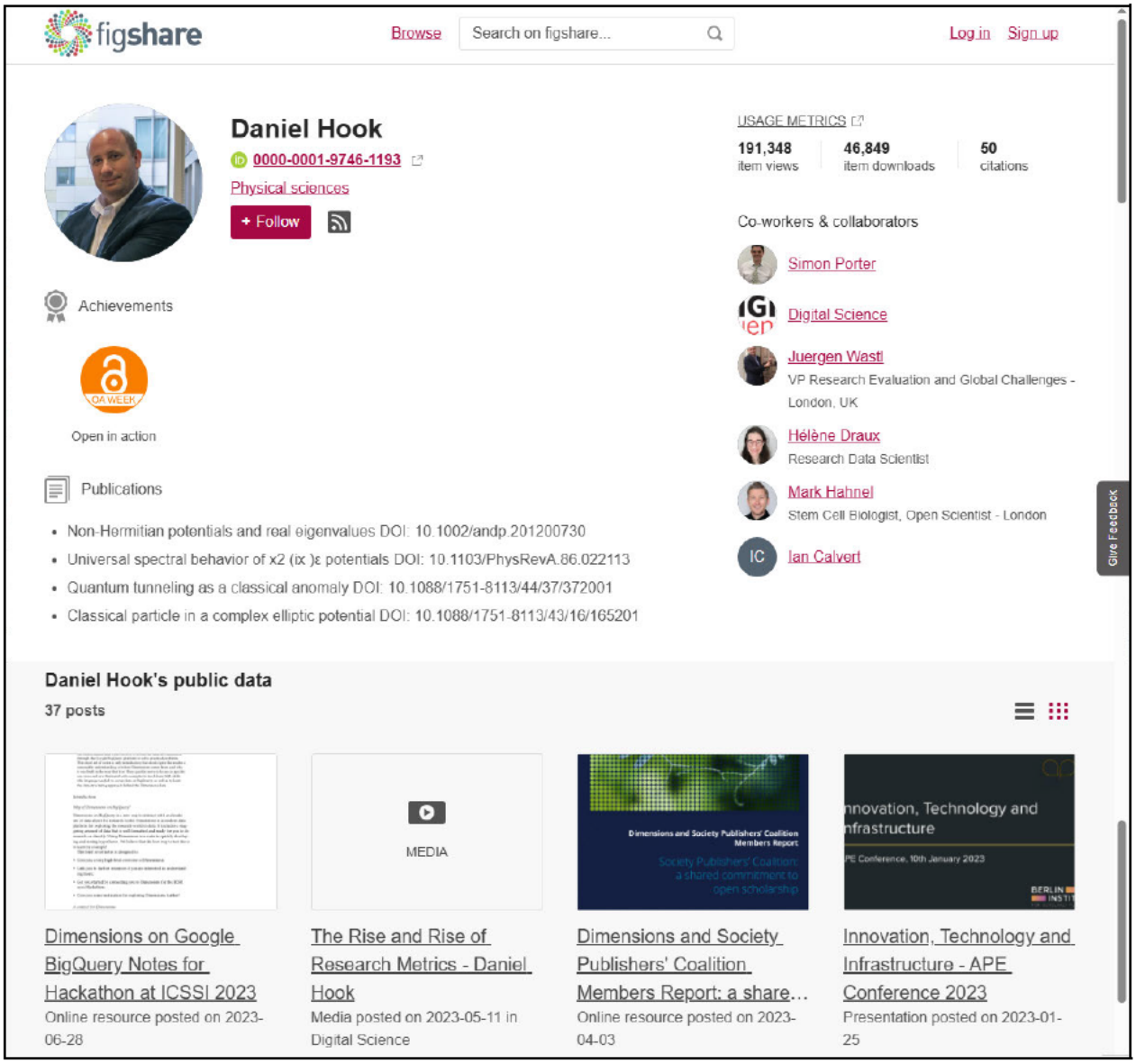


Figure S6. Author page on the Figshare repository website, an example GREI repository, showing posted data files and person-level usage and citation metrics.

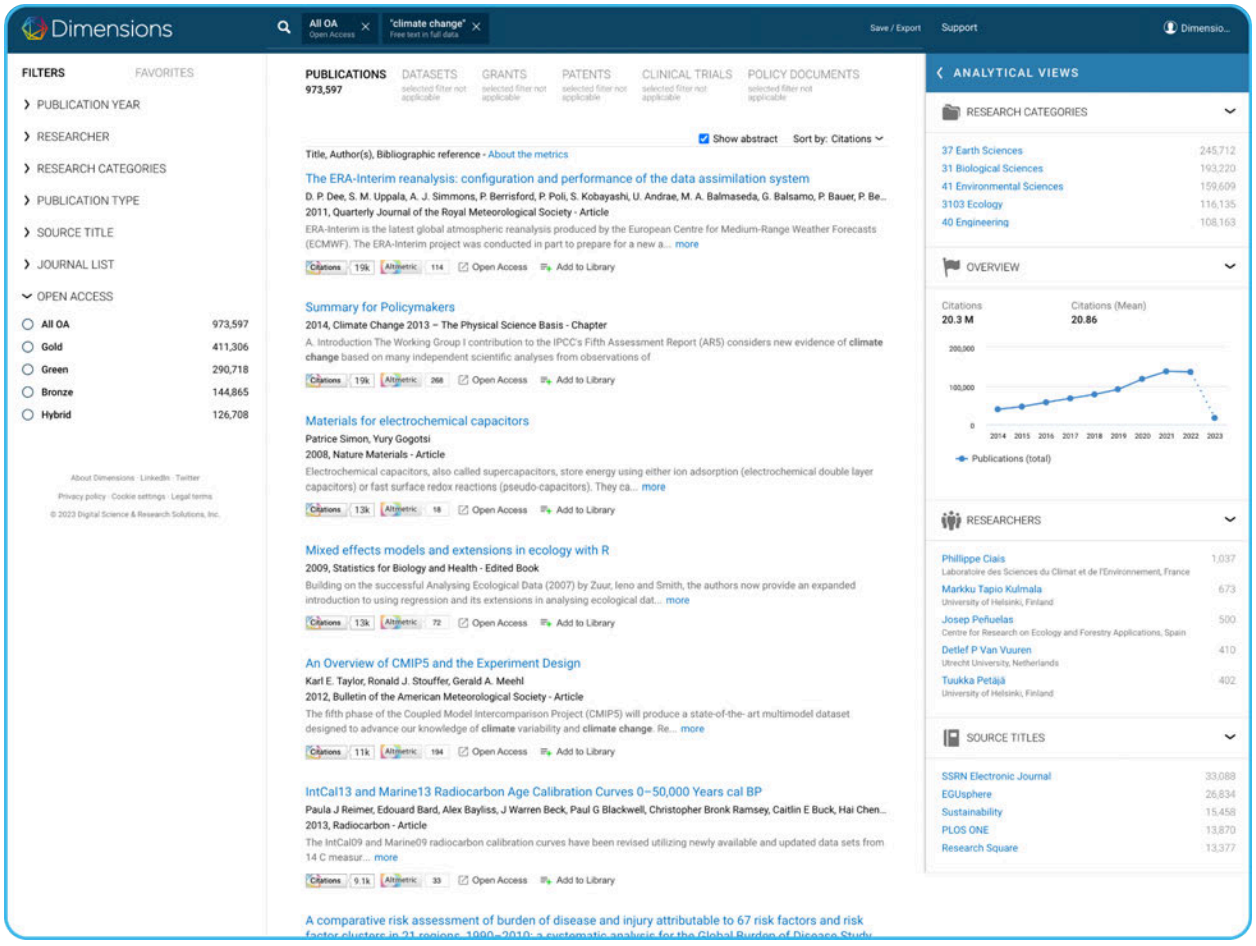


Figure S7. Search results in Dimensions, an example citation database with 29 million datasets and 140 million publications with linked funding, patents, clinical trials, and policy documents.

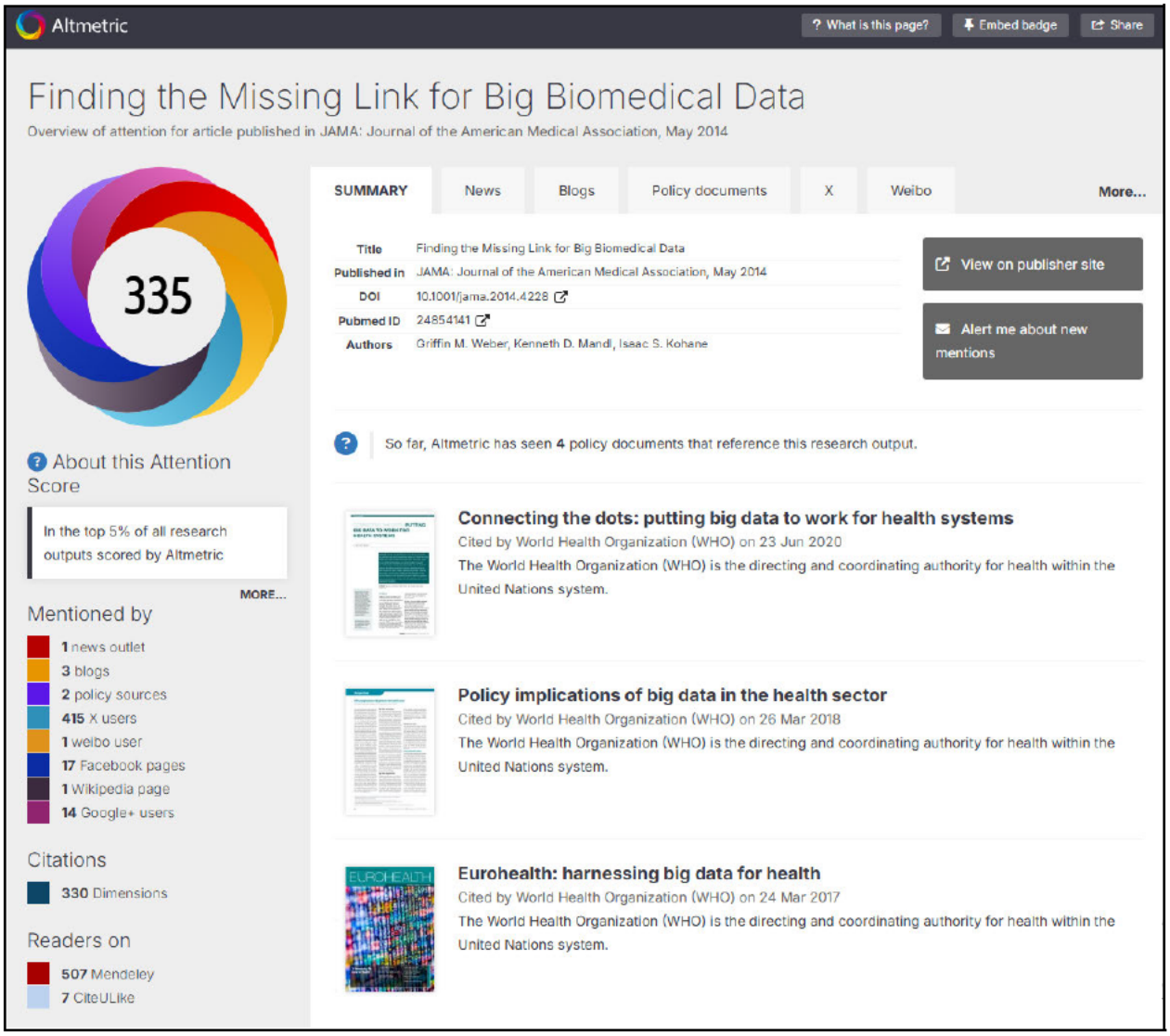


Figure S8. Journal article summary page on Altmetric, which uses eye-catching graphics to highlight metrics that track mentions in social media, news outlets, policy documents, and more.

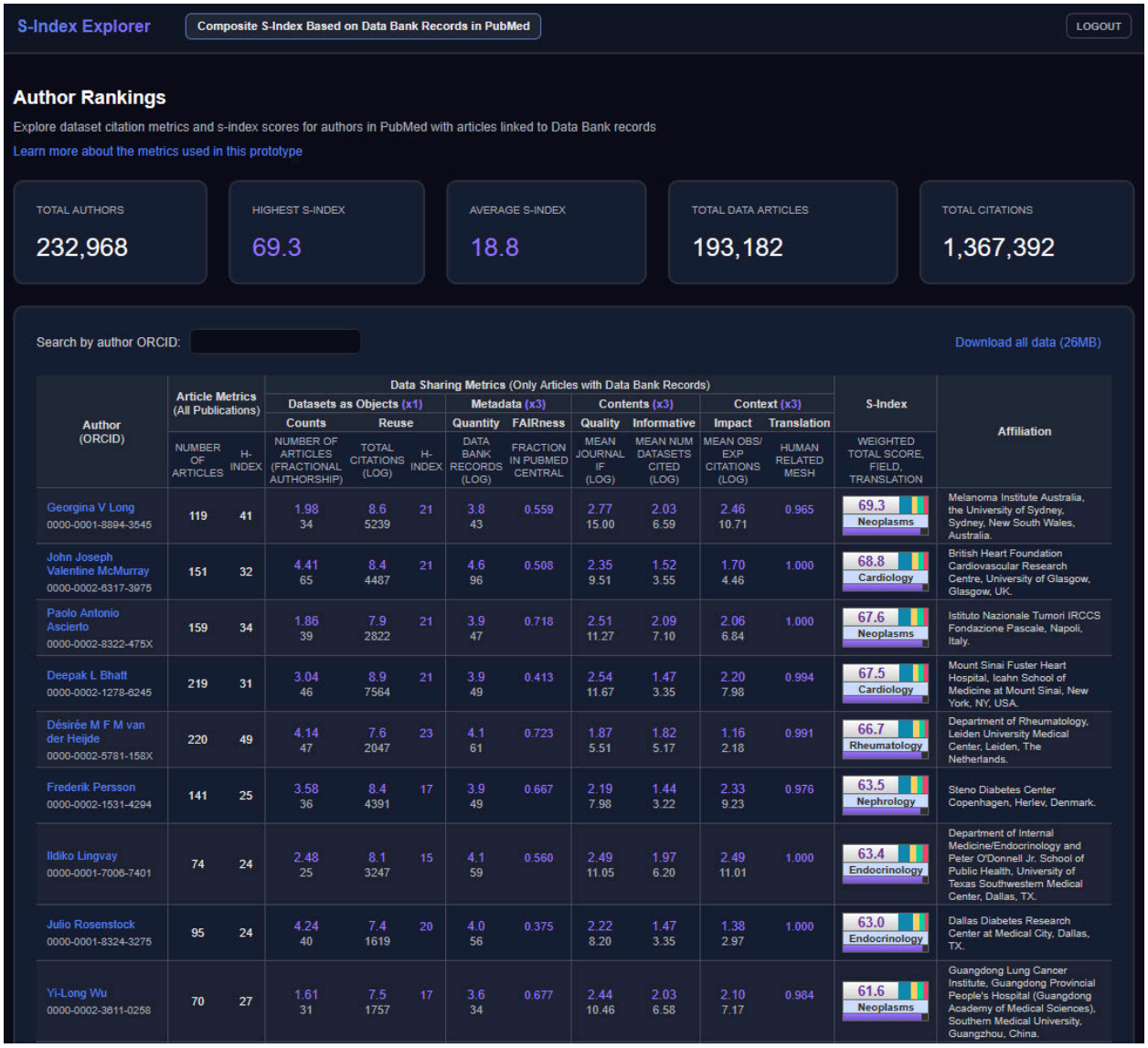


Figure S9. User interface of the Alternative PubMed S-Index prototype, based on data from PubMed, showing a list of top authors ranked by composite S-Index total score. (https://DataSharingIndex.org/Alt)



METRICS

Note that for several metrics, the natural log of the value plus one is used to reduce the effect of extreme values on the overall S-Index score.

COMPONENT	METRIC	DEFINITION	ENCOURAGES	POTENTIAL RISKS
Datasets as Objects (weight = 1)				
Counts	Number of Articles Fractional Authorship	The number of articles with at least one Data Bank record ("data articles"). For this metric, fractional authorship is used, where each article is weighted by the inverse of the number of authors. For example, an article with two authors is counted 0.5 and an article with ten authors is counted 0.1. This assumes authors contributed equally to creating and sharing the dataset. Alternative approaches might use CRediT (Contributor Roles Taxonomy) information in the article.	Sharing more datasets	Intentionally splitting data across multiple articles; sharing smaller low quality datasets; collaborating with fewer people
Reuse	Total Citations (Log)	The total number of citations of all the author's data articles.	Sharing more usable datasets	Self citations; measuring citations of the publication rather than the dataset
	H-Index	The h-index of the author's data articles. The h-index is calculated in the usual way, indicating the number of data articles that have at least that many citations. The metric combines both number of datasets and impact.	Adoption of the S-Index due to similarities to the well-known h-index	Similar limitations as the h-index, such as not controlling for field differences or career stage
Metadata (weight = 3)				
Quantity	Data Bank Records (Log)	The total number of Data Bank records for all the author's data articles.	Sharing larger amounts of data	Penalizes data formats that cannot be divided into smaller individual records
FAIRness	Fraction In PubMed Central	The fraction of the author's data articles that are in PubMed Central.	Open access to the articles describing the data	Does not measure the accessibility of the actual data
Contents (weight = 3)				
Quality	Mean Journal IF (Log)	The average impact factor (IF), calculated from PubMed citation data, of the journals that published the author's data articles. Note that this metric is a proxy for data quality, assuming that a high IF journal has screened out datasets with data quality problems during the peer review process.	Sharing higher quality datasets	Penalizes speciality journals with a lower IF that might have presented the data to a more targeted audience
Informative	Mean Num Datasets Cited (Log)	The average number of references to other data articles in the author's data articles. For example, an author's data article might cite three other data articles. Note that this is a proxy for novelty or integration of different data sources, without direct knowledge of how those data were used.	Reuse of existing data; documenting provenance; acknowledging others' contributions to data sharing	Authors are incentivized to cite many datasets even if they did not use them in a meaningful way
Context (weight = 3)				
Impact	Mean Obs/Exp Citations (Log)	The average number of citations of the author's data articles, normalized by Data Bank type and article year. For example, if a data article from 2021 contains GenBank records, its citation count (observed citations) is divided by the average citation count of all data articles with GenBank records in 2021 (expected citations). If the data article contains records from multiple Data Banks, then a weighted average of the expected citations is used.	Sharing data that is useful to many scientists	Discourages sharing niche data that might be essential to certain scientists, but is unlikely to have many citations; sensitive to the normalization method
Translation	Human Related MeSH	The average fraction of MeSH headings on the author's data articles that are related to humans, compared to MeSH headings on the article related to molecules, cells, or animals. Specifically, these are MeSH headings "Humans" and "Persons" and all their descendants in the MeSH hierarchy (e.g., "Adult", "Child", etc.). This metric indicates where the data articles fit on the translational spectrum from basic science to clinical medicine. The method is described in Weber 2013.	Generating data with direct applications to human health	Penalizes basic science research

Figure S10. Screenshot of help text within the Alternative PubMed S-Index prototype, with definitions of each metric, along with the behaviors they are expected to encourage and potential risks. (https://DataSharingIndex.org/Alt)

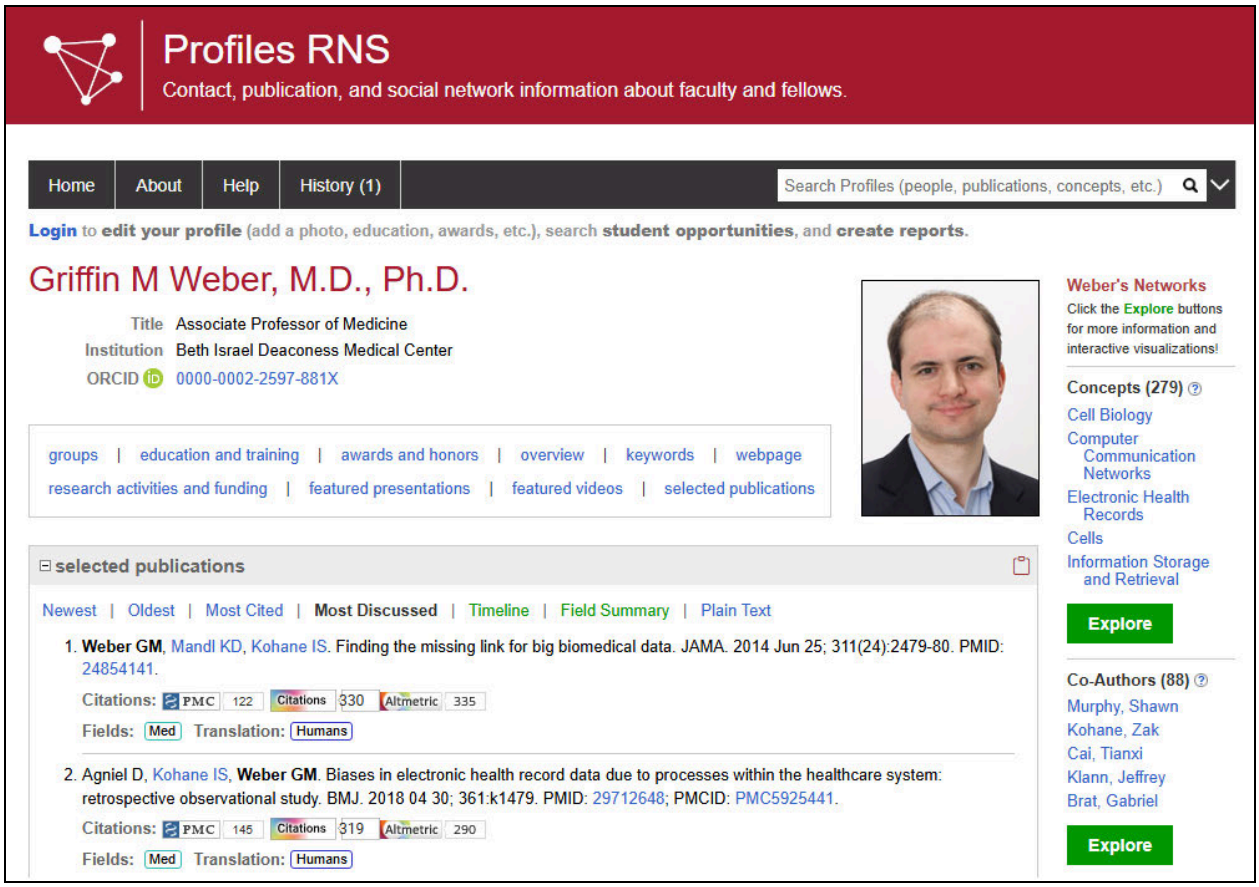


Figure S11. An investigator page on Profiles RNS, a widely-used open-source expert finder system, highlighting Altmetric and Dimensions badges and other publication metrics.

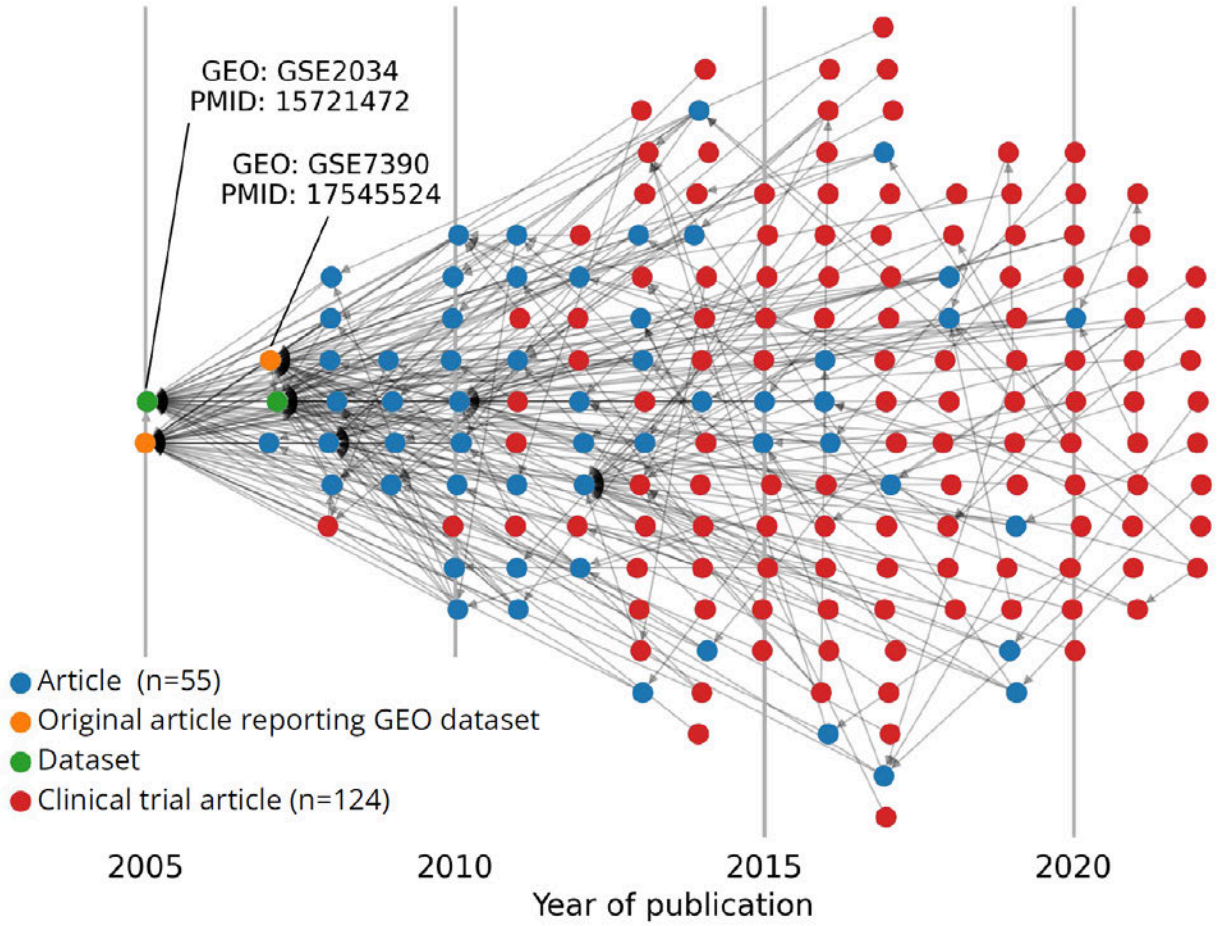


Figure S12. Tracing impact of data sharing using a network diagram of data citations for two Gene Expression Omnibus (GEO) datasets (green circles), one a derivative of the other, in the Make Data Count Data Citation Corpus.

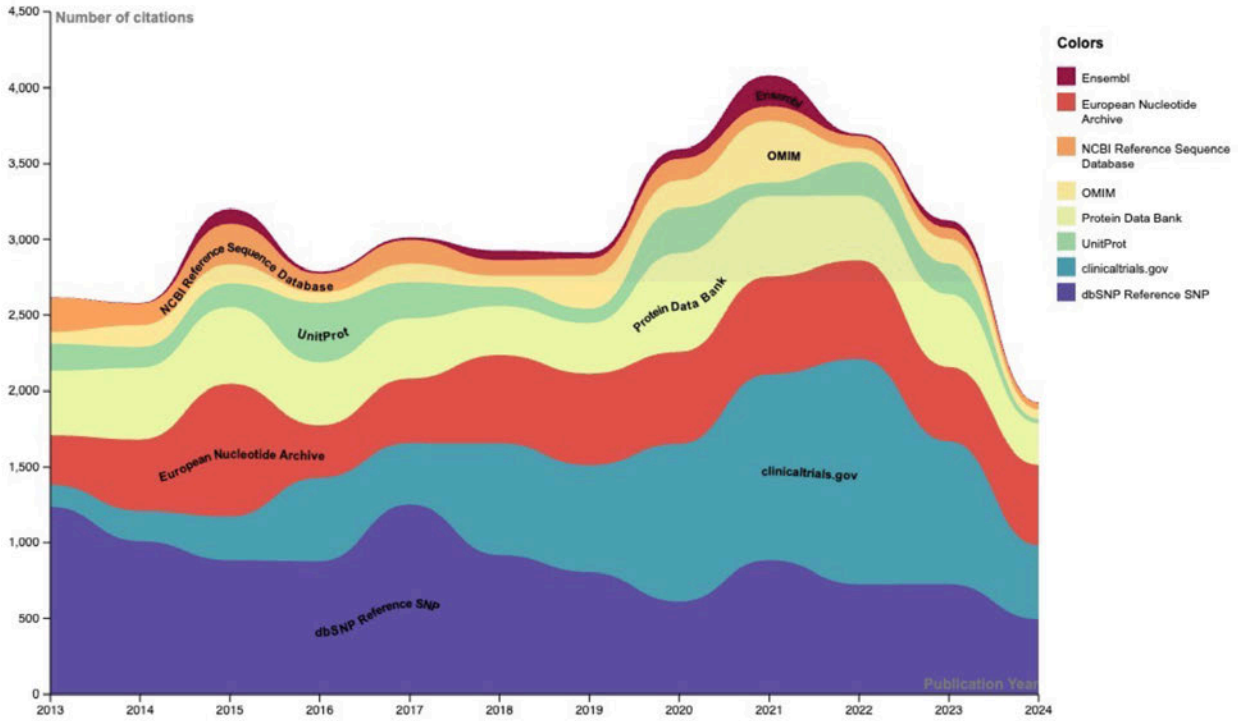


Figure S13. Breakdown of dataset citations by data type between 2013-2024. Example from Northwestern and University of Colorado Boulder analysis of impact of their life science data repositories.

Letters of Support

We have included the following letters of support as Supplementary Information:

Community	Organization	Signatories	Title
Evaluation	National Evaluation Center (NEC) for the Network of the National Library of Medicine (NNLM)	Verónica Hoyo	Executive Director
Evaluation	NIH CTSA Program Evaluators	Joe Hunt	Co-Chair
Funder	Alfred P. Sloan Foundation	Daniel Goroff	VP and Program Director
Funder	Gates Foundation	Ashley Farley	Senior Officer
Funder	Novo Nordisk Foundation	Thomas Chirstensen	Senior VP
Funder	The Navigation Fund	Karthik Ram	Open Science Program Officer
Infrastructure	Contributor Role Taxonomy (CRediT) Project	Liz Allen; Veronique Kiermer	Co-Chairs
Infrastructure	DataCite	Matt Buys	Executive Director
Infrastructure	NIH Generalist Repository Ecosystem Initiative (GREI)	Julie Wood; Ana Van Gulick	Co-Chairs
Infrastructure	Invest in Open Infrastructure	Kaitlin Thaney	Executive Director
Infrastructure	Make Data Count	Iratxe Puebla	Director
Infrastructure	Zenodo (operated by CERN)	Jose Benito Gonzalez Lopez	Senior Staff
Infrastructure	Data.org	Danil Mikhailov	Executive Director
Research Institute	Howard Hughes Medical Institute, Janelia Research Campus	Wyatt Korff; Mike Perham	Directors
Research Integrity, Ethics	Accountability in Research; Northwestern University	Mohammad Hosseini	Associate Editor; Assistant Professor
Researchers	Federation of American Societies for Experimental Biology (FASEB) [22 scientific societies representing over 110,000 Researchers]	Frank Krause; Darla Henderson	Executive Director and CEO; Chief Publishing, Open Science, and Research Integrity Officer
University Health Sciences Libraries	U Colorado Anschutz Med Campus; Harvard Medical School; Northwestern University; University of Washington; University of Virginia; University of New Mexico; Indiana U School of Medicine; Johns Hopkins University; University of Illinois at Chicago; University of Chicago; Weill Cornell Medicine	Melissa De Santis; Len Levin; Kristi Holmes; Jenny Muilenburg; Bart Ragon; Melissa L. Rethlefsen; Gabe Rios; Anne Seymour; Donghua Tao; Debra A. Werner; Terrie Wheeler	Library Directors

Katy Börner, Indiana University
Amy Brand, MIT Press
Kristi Holmes, Northwestern University
Daniel Hook, Digital Science
John Inglis, bioRxiv/medRxiv
Chonnettia Jones, Addgene
Arjun Krishnan, University of Colorado
Steinn Sigurdsson, arXiv
Griffin Weber, Harvard

RE: NIH S-Index Challenge Proposal Letter of Support

Dear Katy, Amy, Kristi, Daniel, Chonnettia, Arjun, Steinn and Griffin

It is my pleasure to write on behalf of the National Evaluation Center (NEC) for the Network of the National Library of Medicine (NNLM) to offer strong support and partnership for your proposal submitted to the NIH Data Sharing Challenge "S-Index". We bring many years of multidisciplinary practice and extensive experience in evaluating a wide range of biomedical programs and assessing their impact. The NEC is committed to advancing responsible and transparent evaluation practices, and we are enthusiastic about the S-Index's potential to encourage meaningful, measurable, and socially valuable data sharing.

As Executive Director of the NEC, I affirm our commitment to collaborating with your S-Index development team in ways that emphasize the core principles of responsible metrics, including transparency, contextual relevance, fairness, and interestholder involvement. We value the project's commitment to making data impact more visible and understandable not only to researchers and institutions but also to the public, helping to build trust in scientific research and in the open sharing of data as a public good.

We envision several specific opportunities for potential collaboration and look forward to considering other ways we might work together:

- NEC staff can contribute to design and feedback for the collaborative workshops, offering expertise in evaluation methodology, responsible metrics, and the practical application of indicators in diverse research and community settings.
- Drawing on our deep understanding of NNLM member organizations, including academic centers, libraries, community health partners, and other community organizations, we can help ensure the S-Index reflects the range of values and priorities of information providers and the different communities they serve.
- We can help inform outreach efforts to engage public interestholders, emphasizing the importance of clear, accessible communication about what data sharing impact means, and how it supports community health, equity, and scientific progress.
- We can also contribute our evaluation experience to this effort. For example, we can consider prototypes of the S-Index, providing structured input on usability, interpretability, and alignment with responsible, public-facing metrics. We can also co-develop evaluation tools to help assess uptake, relevance, and public understanding, and relatedly, offerour expertise in tailored messaging of these outputs.

In summary, the NEC is proud to collaborate with you on this important initiative. The S-Index stands to advance a more transparent, inclusive, and effective data sharing culture. We are excited about the opportunity to grow a data sharing and reuse culture that recognizes meaningful contributions and serves both scientific and larger public interests. We look forward to supporting the project through the challenge and into broader implementation. Please don't hesitate to reach out if we can be of assistance.

Best regards,

Verónica Hoyo, PhD
Executive Director, NNLM National Evaluation Center

INDIANA CTSI
Clinical and Translational Sciences Institute

Improving Health through Research

May 27, 2025

Katy Börner, Indiana University
Amy Brand, MIT Press
Kristi Holmes, Northwestern University
Daniel Hook, Digital Science
John Inglis, bioRxiv/medRxiv
Chonnettia Jones, Addgene
Arjun Krishnan, University of Colorado Anschutz Medical Campus
Steinn Sigurdsson, arXiv
Griffin Weber, Harvard

To Whom It May Concern:

I am writing to express my support and collaboration with your team for your proposal to the NIH Data Sharing Challenge (S-Index). I serve as Co-Chair of the Clinical and Translational Award Program Evaluators group, Co-Chair the Association for Clinical and Translational Science Evaluation Special Interest Group, and I also serve as the Director of the Tracking and Evaluation Program at the Indiana Clinical and Translational Sciences Institute, Indiana University School of Medicine.

In addition to extensive experience with formative, summative, and process evaluation methodologies, I have completed extensive work to understand clinical and translational research impact. Datasets play an important role in realizing this impact. I am excited about your team's efforts to advance innovative solutions that assess and incentivize responsible data sharing in biomedical research. The S-Index presents a compelling approach to quantifying the impact and reuse of shared datasets. I am certain this effort will be of interest to my CTSA evaluation colleagues across the country.

The NIH Clinical and Translational Science Awards (CTSA) Program, led by the National Center for Advancing Translational Sciences (NCATS), supports a national network of medical research institutions working to improve the efficiency, quality, and impact of translational research. By fostering innovation, collaboration, and workforce development, the CTSA Program accelerates the translation of scientific discoveries into clinical applications to improve human health.

I am eager to participate in your proposed collaborative workshops where I and other CTSA evaluators can provide input on the design and implementation of the proposed S-Index, and offer insights on metric usability for CTSA hubs, and alignment with translational research goals.

I can help engage the broader CTSA evaluation community, including the CTSA Translational Impact Working Group, to explore use cases, address implementation challenges, and identify opportunities for integration. Additionally, I can assist in disseminating findings through CTSA-affiliated channels to help broaden awareness and uptake of the S-Index across the translational science landscape.

The S-Index project aligns strongly with the CTSA mission to accelerate translation through data-driven decision-making and a culture of transparency and accountability.

Please do not hesitate to contact me should additional information be required. I look forward to the opportunity to support this project.

Sincerely,

[redacted]

Joe Hunt
Director, Tracking and Evaluation
Indiana Clinical and Translational Sciences Institute
Co-Chair, CTSA Program Evaluators
Co-Chair, Association for Clinical and Translational Science Evaluation Special Interest Group

Daniel L. Goroff
Vice President and Program Director

ALFRED P. SLOAN
FOUNDATION

Driven by the promise
of great ideas.

May 26, 2025

Dear Chonnettia,

It is my pleasure to support your team's proposal to the NIH Data Sharing Index Challenge entitled "Engaging key stakeholder communities to promote data sharing with an S-Index."

The Alfred P. Sloan Foundation, where I serve as Vice President and Program Director, is a dedicated funder of Open Science. One recent and successful example, led by my colleague Josh Greenberg, has been our support for the establishment of Open Source Program Offices (OSPOs) on many university campuses. Another recent investment of this kind by Sloan is our involvement with Open Athena, a high-powered initiative to develop transparent foundational models for use by academic scientists.

Personally, I am especially excited about how the approach this team is taking emphasizes stakeholder engagement. It is a truism that you cannot manage what you cannot measure. But an S-Index will only "work" if it takes into account factors that may not be apparent to everyone. The eclectic and very distinguished team working on this Challenge represents many points of view—from academics to advocates and from commercial data providers to nonprofit data stewards.

With this remarkable array of experience and expertise, the team is exceptionally ready to take on all the challenges entailed by the Data Sharing Index Challenge—including thorny issues that I had to confront when promoting Open Science during my service as a government official. Examples of these are privacy and proprietary considerations, intellectual property rights, sustainable timing and costs, societal benefits measurement, as well as international research security protections.

For all these reasons, I would be delighted to see this project make progress on encouraging and evaluating better data practices among scientific researchers. I have every confidence this team is uniquely poised to make significant, welcome, and enduring progress. Above all, I am eager to participate in these activities and do what I can to make this work successful.

Sincerely yours,

Danny
Dr. Daniel L. Goroff

Gates Foundation

PO Box 23350 Seattle, WA 98102, USA
T: +1.206.709.3100
F: +1.206.494.7100
www.gatesfoundation.org

Dear Katy Börner (Indiana U), Amy Brand (MIT Press), Kristi Holmes (Northwestern), Daniel Hook (Digital Science), John Inglis (bioRxiv/medRxiv), Chonnettia Jones (Addgene), Arjun Krishnan (U Colorado), Steinn Sigurdsson (arXiv), Griffin Weber (Harvard),

I am writing in support of your proposal to the NIH Data Sharing Index Challenge titled "Engaging key stakeholder communities to promote data sharing with an S-Index".

The Gates Foundation is committed to advancing global health and development through the responsible sharing and reuse of data. We support policies and efforts that promote open access to high-quality data, foster data interoperability, and encourage collaboration across sectors. By enabling researchers, policymakers, and practitioners to build on existing evidence, we aim to accelerate innovation, increase transparency, and drive more equitable outcomes for communities around the world.

For these reasons, we understand both the benefits and challenges in incentivizing open science and sharing high value datasets. We are pleased to support your team's exciting plans to promote data sharing and data reuse to further drive a culture of openness, transparency, and explainability in science.

I am looking forward to participating in your stakeholder workshops and providing my guidance and expertise.

Best wishes,

Ashley Farley
Senior Officer, Knowledge & Research Services, Gates Foundation



foundation



navigation.org
info@navigation.org

Katy Börner, Indiana University
Amy Brand, MIT Press
Kristi Holmes, Northwestern University
Daniel Hook, Digital Science
John Inglis, bioRxiv/medRxiv
Chonnettia Jones, Addgene
Arjun Krishnan, University of Colorado
Steinn Sigurdsson, arXiv
Griffin Weber, Harvard University

RE: NIH S-Index Challenge Proposal Letter of Support

Dear Katy Börner, Amy Brand, Kristi Holmes, Daniel Hook, John Inglis, Chonnettia Jones, Arjun Krishnan, Steinn Sigurdsson, and Griffin Weber

On behalf of the Navigation Fund, I would like to express our support for your team's pitch to the NIH S-Index Data Sharing Challenge. As a philanthropy focused on open science, we support initiatives that explore ways to encourage broader data sharing in research.

Should you advance in this challenge, we would be pleased to participate in collaborative workshops to help ensure the S-Index reflects the needs and values of the research communities it serves. We can also help connect the project to our broader networks across open science and academic reform communities.

We look forward to potential collaboration with your team and the NIH in support of more open and sustainable science.

Best regards,

[redacted]

Karthik Ram,

Open Science Program Office

Katy Börner, Indiana University
Amy Brand, MIT Press
Kristi Holmes, Northwestern University
Daniel Hook, Digital Science
John Inglis, bioRxiv/medRxiv
Chonnettia Jones, Addgene
Arjun Krishnan, University of Colorado Anschutz Medical Campus
Steinn Sigurdsson, arXiv
Griffin Weber, Harvard

Dear colleagues,

On behalf of the Contributor Role Taxonomy (CRediT) project, we are pleased to offer our support for your submission to the NIH Data Sharing “S-Index” Challenge.

CRediT is a high-level taxonomy of contributor roles that enables transparent and standardized attribution for individuals involved in scholarly work. Originally developed to bring clarity to author contributions in journal articles, CRediT includes 14 distinct roles (Data Curation, Software, Methodology, and Supervision) designed to capture the full spectrum of activities essential to research. Notably, Amy Brand and I led the original development of CRediT, which is now a formal NISO standard (ANSI/NISO Z39.104-2022), and used by hundreds of journals and systems worldwide to improve the transparency and discoverability of research contributions.

Your proposal closely aligns with the values that underpin CRediT - namely, the importance of finding practical ways to recognize the diversity of research contributions. The collaborative approach to developing the S-Index that you outline, presents an exciting opportunity to develop quantitative and qualitative ways to better measure and incentivize data stewardship and reuse, and has the potential to ensure that data sharing achieves the status of a first-class scholarly activity.

The CRediT team looks forward to the opportunity to engage with you and the larger scholarly community in your workshops. We also welcome future opportunities to collaborate – for example to explore how CRediT can be optimized for research data activities. We commend your leadership in pushing this work forward and look forward to the impact your submission will have on the policy and practice of data sharing.

With appreciation and full support

[redacted]

Liz Allen, PhD
CRediT project Co-Chair

Veronique Kiermer, PhD
CRediT project Co-Chair

May 27th, 2025

DataCite

CONNECTING RESEARCH, ADVANCING KNOWLEDGE

Matt Buys
Phone +49 762-17642
matt.buys@datacite.org
datacite.org

May 27, 2025

RE: Letter of Support | NIH Data Sharing Challenge

Dear Katy Börner (Indiana U), Amy Brand (MIT Press), Kristi Holmes (Northwestern), Daniel Hook (Digital Science), John Inglis (bioRxiv/medRxiv), Chonnettia Jones (Addgene), Arjun Krishnan (U Colorado), Steinn Sigurdsson (arXiv), and Griffin Weber (Harvard)

I am writing in support of your proposal to the NIH Data Sharing Index Challenge titled "Engaging key stakeholder communities to promote data sharing with an S-Index".

As you know, DataCite is dedicated to making research more effective by connecting research outputs and resources—from data and preprints to images and samples. We support the creation and management of DOIs and metadata records, enhance research workflows with service integration, and enable the discovery and reuse of research outputs and resources. Our advanced tools leverage the metadata provided to enable discovery, analytics, and reporting capabilities, ensuring that research impact is visible and measurable through our DataCite Commons platform.

For these reasons, we understand both the benefits and challenges in incentivizing open science and sharing high value datasets. We are pleased to support your team's exciting plans to promote data sharing and data reuse to further drive a culture of openness, transparency, and explainability in science.

DataCite brings deep expertise and a strong track record in supporting global research infrastructure, and may be well positioned to contribute across several shared areas of interest—offering opportunities to strengthen and align with your initiatives.

- **DOI and metadata management:** With years of experience as a trusted provider of DOI registration services, DataCite ensures that datasets and other digital objects are persistently identified, citable, and discoverable in accordance with FAIR principles. Our robust APIs—including the DataCite REST API, GraphQL API, and associated services—are widely adopted and can be integrated into your systems to support automated DOI creation, enrich metadata quality, and improve interoperability with existing research

repositories.

- **Analytics and reporting:** DataCite Commons and our advanced analytics tools provide insights into the reach and use of research outputs. These tools can support your efforts to track link-outs and citations, assess usage patterns, and demonstrate the scientific impact of your resources.
- **DataCite Metadata Schema:** As a leader in metadata standards development, DataCite has shaped community-wide practices around metadata quality—spanning accuracy, completeness, consistency, and context. We welcome the opportunity to share this experience and explore how our approaches can inform your efforts to assess and improve metadata quality, including the impact of AI models that account for connectivity, context, and broader impact across disciplines.

DataCite is committed to supporting our shared mission to democratize data discovery and advance biomedical research through innovative metadata collection and management. Our team at DataCite looks forward to contributing to your stakeholder workshops where possible and providing our guidance and expertise in this regard. We are confident that our collaboration on this effort will significantly enhance the functionality and reach of this effort for biomedical researchers worldwide.

We are excited about the opportunity to collaborate on this transformative initiative. We look forward to working closely with you and the stakeholder community to realize the full potential of this opportunity and to contribute to the advancement of human health through enhanced data discoverability and reuse.

Yours sincerely,

Matt Buys
Executive Director



May 27, 2025

Katy Börner, Indiana University
Amy Brand, MIT Press
Kristi Holmes, Northwestern University
Daniel Hook, Digital Science
John Inglis, bioRxiv/medRxiv
Chonnettia Jones, Addgene
Arjun Krishnan, University of Colorado Anschutz Medical Campus
Steinn Sigurdsson, arXiv
Griffin Weber, Harvard

Dear Colleagues,

On behalf of the NIH-sponsored Generalist Repository Ecosystem Initiative (GREI), we are pleased to offer our support for your submission to the NIH S-Index Challenge, and express our interest in collaborating with you on this important effort to improve the recognition and reward of data sharing in biomedical research.

GREI is a collaborative effort among seven NIH-recognized generalist repositories (Dataverse, Dryad, Figshare, OSF, Mendeley Data, Vivli, and Zenodo) working to enhance the discoverability, interoperability, and reusability of research data. We share a common commitment to building interoperable data infrastructure, promoting FAIR data principles, and advancing a research culture that values data as a first-class scholarly output. The goals of the S-Index Challenge align closely with GREI's mission to support researchers in sharing high-quality, well-documented data across disciplines and platforms.

GREI is well-positioned to contribute to this effort by participating in your proposed stakeholder workshops where we can contribute our gained perspectives from working

GREI is made possible by support from the NIH Office of Data Science Strategy/Office of the NIH Director pursuant to OTA-21-009, "Generalist Repository Ecosystem Initiative (GREI)" through Other Transactions Agreement (OTA) Numbers OT2DB000001, OT2DB000002, OT2DB000003, OT2DB000004, OT2DB000005, OT2DB000006, and OT2DB000013.

together as a “coopetition” to support the advancement of the NIH data sharing ecosystem. We can also provide technical guidance on data deposit practices, metadata standards, and persistent identifiers across generalist repositories. We can further contribute insights into cross-repository usage metrics, discoverability, and data citation practices, helping to ensure the proposed metric is grounded in real-world data sharing behaviors.

We believe the NIH Data Sharing S-Index Challenge will contribute meaningfully to advancing a more FAIR and sustainable data-sharing ecosystem and we look forward to supporting your team in these efforts.

Best wishes,



Julie Wood
Co-chair GREI Collaboration, 2025 and 2022



Ana Van Gulick, PhD
Co-Chair GREI Collaboration, 2025 and 2023
Digital Science Head of Customer Engagement for Government, Funders, and Nonprofits

GREI is made possible by support from the NIH Office of Data Science Strategy/Office of the NIH Director pursuant to OTA-21-009, “Generalist Repository Ecosystem Initiative (GREI)” through Other Transactions Agreement (OTA) Numbers OT2DB000001, OT2DB000002, OT2DB000003, OT2DB000004, OT2DB000005, OT2DB000006, and OT2DB000013.

Invest in Open Infrastructure

Dear Katy Börner, Amy Brand, Kristi Holmes, Daniel Hook, John Inglis, Jones, Arjun Krishnan, Steinn Sigurdsson, Griffin Weber,

I am writing in support of your proposal to the NIH Data Sharing Index Challenge titled "Engaging key stakeholder communities to promote data sharing with an S-Index".

A key objective of Invest in Open Infrastructure (IOI) is to increase the visibility of open infrastructure providers and their value to the ecosystem through targeted research on the benefits, opportunities, and challenges associated with adoption. We work to expand our discovery tools and our mapping of infrastructure providers and needs to new disciplines and communities, including data storage, research data management, data analysis, and digital preservation.

For these reasons, we understand both the benefits and challenges in incentivizing open science and sharing high value datasets. We are pleased to support your team's exciting plans to promote data sharing and data reuse to further drive a culture of openness, transparency, and explainability in science.

I am looking forward to participating in your stakeholder workshops and providing my guidance and expertise.

Best wishes,
Kaitlin

–

Kaitlin Thaney
Executive Director
Invest in Open Infrastructure

27 May 2025

Dear Kristi,

On behalf of Make Data Count, I express our support for your participation in the NIH Data Sharing S-Index Challenge and offer our advice and guidance on your proposal, "Engaging key stakeholder communities to promote data sharing with an S-Index".

Make Data Count is a community-driven initiative focused on building the infrastructure, standards, and cultural change necessary to make research data a primary scholarly output. Our work is anchored on open infrastructure, community ownership, and the development of meaningful data metrics that go beyond traditional publication volume or citation counts to reflect a broader range of dimensions of impact and reuse. We advocate for indicators and metrics that recognize and reward data sharing and data use, aligning with our commitment to transforming the way data is valued in scholarly research.

We see synergy between your work and Make Data Count's goals to establish open, transparent, and community-governed metrics for data reuse and impact. We believe this work can meaningfully contribute to the broader goals of NIH's S-Index Challenge by developing practical, scalable methods to track and incentivize data sharing and reuse in responsible ways.

Make Data Count is open to collaboration with your team on areas of shared interest, including:

- Advising on the development of robust and transparent data metrics
- Sharing insights from our projects on infrastructure for data-usage tracking, and our collaborations with meta-research on data citation practices
- Providing guidance on the use and application of data-usage information collected via Make Data Count workflows
- Contributing to community engagement and alignment on the practices and principles for responsible data metrics
- Participation in your stakeholder workshops and providing our guidance and expertise

Kristi, in addition to serving as an advisor to Make Data Count, you also co-chair the Make Data Count/HELIOS Open Working Group, 'Implementing Data Evaluation in Academia'. This collaborative initiative focuses on developing practical resources to support the implementation of data metrics and data evaluation in institutional processes, particularly in hiring, review, promotion, and tenure decisions. This Working Group can offer another opportunity for collaboration and guidance, given its active work developing frameworks that recognize and reward data sharing and data use, supporting inclusion and recognition of data outputs in academic evaluation processes.

We look forward to exploring how we can work together to advance the recognition of datasets as critical scientific contributions, and as catalysts of research progress.

Best wishes,



Iratxe Puebla
Director, Make Data Count

European Organization for Nuclear Research
Organisation européenne pour la recherche nucléaire

Jose Benito Gonzalez Lopez
Senior Staff
IT Department
CERN
Geneva, CH 1211
Email: jbenito@cern.ch

Kristi Holmes
Northwestern University
Chicago, IL 60611

Katy Börner, Indiana University
Amy Brand, MIT Press
Kristi Holmes, Northwestern University
Daniel Hook, Digital Science
John Inglis, bioRxiv/medRxiv
Chonnettia Jones, Addgene
Arjun Krishnan, University of Colorado Anschutz Medical Campus
Steinn Sigurdsson, arXiv
Griffin Weber, Harvard

Geneva, 28 May 2025

Dear Kristi and colleagues,

On behalf of Zenodo, I am pleased to provide this letter of support and confirm our collaboration in your application to the NIH Data Sharing S-Index Challenge. Your team's efforts to develop meaningful, responsible metrics for data sharing and reuse work in harmony with Zenodo's mission to support FAIR, open, and impactful science across disciplines.

Zenodo, developed under the European OpenAIRE program and operated by CERN, is a trusted generalist repository that provides long-term, open access to research outputs of any type or size. Our underlying platform, InvenioRDM, is a modular, scalable infrastructure that supports customizable metadata, persistent identifiers (e.g., DOIs, ORCIDs, RORs), and integration with services such as GitHub. These capabilities make Zenodo well-suited to serve as a foundation for assessing and enhancing data sharing practices. Through our collaboration with the NIH Generalist Repository Ecosystem Initiative (GREI), Zenodo is actively collaborating with other repositories to improve metrics, metadata standards, interoperability, and user-centered services that support NIH-funded researchers. We are also working on ways to highlight NIH-funded science in Zenodo.

Zenodo is pleased to support your project through active participation in the stakeholder workshop series to co-develop best practices and community-informed strategies for measuring data impact. We bring direct experience as an early implementer of the Make Data Count (MDC) standard, which enables transparent, standardized tracking of dataset views, downloads, and citations, essential information for building a robust S-Index. Our continued collaboration through the Generalist Repository Ecosystem Initiative (GREI) further supports the advancement of shared infrastructure, metric standards, and cross-repository interoperability to improve the discoverability and reuse of NIH-funded data. Additionally, Zenodo's REST API, OAI-PMH capabilities, and open licensing model enable efficient programmatic access and reuse. We also share your interest in applying AI and machine learning to enhance metadata quality and discovery, all areas where the modular architecture of our underlying platform, InvenioRDM, is already driving innovation.

We're excited about your community-centered approach. Please don't hesitate to reach out as your project moves forward.

Yours sincerely,



Jose Benito

Dear Dr Katy Börner (Indiana U), Dr Amy Brand (MIT Press), Dr Kristi Holmes (Northwestern), Dr Daniel Hook (Digital Science), Dr John Inglis (bioRxiv/medRxiv), Dr Chonnettia Jones (Addgene), Dr Arjun Krishnan (U Colorado), Dr Steinn Sigurdsson (arXiv), and Dr Griffin Weber (Harvard),

I am writing in support of your proposal to the NIH Data Sharing Index Challenge titled "Engaging key stakeholder communities to promote data sharing with an S-Index".

DataDotOrg, is a US-based non-profit organisation leading extensive global programs in the data for good and academic data space, particularly in areas of Public Health and Climate and Health. We are long standing supporters of open science, open data and open software. In our data work we partner with key universities, like Harvard University, University of Chicago in the US, Oxford University, London School of Hygiene and Tropical Medicine in the UK, University of Lagos, University of Pretoria in Africa, IIIT Delhi, Ashoka University, National University of Singapore in Asia, Javeriana University, Los Andes University in South America, as well as social impact organisations, governments and intergovernmental organisations from around the world.

For these reasons, we understand both the benefits and challenges in incentivizing open science and sharing high value datasets. We are pleased to support your team's exciting plans to promote data sharing and data reuse to further drive a culture of openness, transparency, and explainability in science.

I am looking forward to participating in your stakeholder workshops and providing my guidance and expertise.

Best wishes,

Danil Mikhailov, PhD

President and CEO
DataDotOrg

19700 Helix Drive
Ashburn, VA 20147
571.209.4000

May 27, 2025

Dear Dr. Jones and Team:

We are writing in support of your proposal to the NIH Data Sharing Index Challenge titled "Engaging key stakeholder communities to promote data sharing with an S-Index".

Janelia Research Campus has been a leader in promoting open science through comprehensive data sharing initiatives that span both large-scale collaborative projects and individual research efforts. We have established institutional guidelines and infrastructure supporting researchers in sharing data according to FAIR principles, recognizing that meaningful data sharing requires both technical excellence and community engagement. Our track record includes disseminating transformative datasets such as the complete Drosophila connectome (Scheffer *et al.* PMID: 32880371) and the comprehensive Mouselight Neuron Browser (Winnubst *et al.* PMID: 31495573), alongside hundreds of smaller datasets through Figshare and specialized repositories. Through this experience, we have identified critical gaps in current data sharing metrics and standards, particularly the need for measures that capture both the quality of shared data and its actual impact on scientific discovery. We believe robust indexing systems like your proposed S-Index are essential for creating sustainable incentives that reward researchers for the substantial effort required to prepare and maintain high-quality, reusable datasets.

For these reasons, we understand the benefits and challenges of incentivizing open science and sharing high-value datasets. We are pleased to express our support for your team's exciting plans to promote data sharing and reuse to further drive a culture of openness, transparency, and explainability in science.

Although the Howard Hughes Medical Institute, Janelia Research Campus does not apply for or accept grant funding and is not participating in the grant application “Engaging key stakeholder communities to promote data sharing with an S-Index*”,* we would be pleased to participate in your stakeholder workshops on an unfunded basis, and are strongly supportive of the effort to develop a robust, meaningful data sharing index.

Best wishes,

Mike Perham, PhD
Director, Janelia Innovations and Open Science

Wyatt Korff, PhD
Senior Director, Project Teams

Northwestern Medicine®
Feinberg School of Medicine

Department of Preventive Medicine
Northwestern University
Feinberg School of Medicine

312.908.7914
312.908.9588 fax

680 North Lake Shore Drive, Suite 1400
Chicago, Illinois 60611

Northwestern

Kristi Holmes, PhD
Northwestern University
Chicago, IL 60611

May 26, 2025

Katy Börner, Indiana University
Amy Brand, MIT Press
Kristi Holmes, Northwestern University
Daniel Hook, Digital Science
John Inglis, bioRxiv/medRxiv
Chonnettia Jones, Addgene
Arjun Krishnan, University of Colorado Anschutz Medical Campus
Steinn Sigurdsson, arXiv
Griffin Weber, Harvard

Dear Kristi and team,

As a researcher and advocate in the fields of research integrity, ethics, and open science, I am pleased to offer my full support for your team's submission to the NIH Data Sharing S-Index Challenge, and to express my strong interest in collaborating on this timely and important initiative.

Ethical research requires more than transparency in methods and findings; it demands responsible stewardship of data, clear attribution of contributions, and inclusive practices that foster public trust in scientific research. Your team's proposal to develop a robust metric for recognizing and incentivizing data sharing addresses a longstanding gap in the research ecosystem. I'm especially encouraged by your focus on open infrastructure and transparency in metric design. These principles are not only foundational to open science, but also to the development of trustworthy artificial intelligence (AI) systems and responsible research governance.

My work lies at the intersection of research ethics, AI, and open scholarship. I lead and contribute to ethics-focused projects within national and international organizations such as the Association of Professional and Practical Ethics, the National Information Standards Organization (NISO), the National Academy of Sciences, and the Global Young Academy. This work is further grounded in my teaching on research ethics and the responsible conduct of research at Northwestern University and other institutions globally. I also serve as associate editor of *Accountability in Research* and as the editor of the *Research Ethics Forum* book series.

Contributor roles and attribution have been central to my scholarship since my PhD dissertation, and I've been fortunate to continue this work in collaboration with you and others since joining Northwestern. Together, we serve on the CRediT Standing Committee at NISO, where we help steward the Contributor Roles Taxonomy (CRediT), a framework that supports proper attribution and accountability and is critical to building incentives for data sharing.

I'm particularly excited by your plan to include stakeholder workshops as part of the project. These workshops are essential from an ethics and integrity standpoint, ensuring that diverse voices such as researchers, infrastructure providers, funders, scholarly organizations and the public can work together to shape a fair, transparent, and accountable metric for recognizing and incentivizing data sharing. Such engagement helps illuminate ethical risks and foster trust in both the qualitative and quantitative components of any metric.

In addition to participating in the workshops, I would be eager to contribute in the following ways:

- Advising on the integration of ethical considerations into the design and implementation of data sharing metrics
- Supporting the alignment of metrics with best practices in AI reproducibility and accountability
- Contributing perspectives on openness, fairness, and the recognition of contributions from different researchers and communities
- Ensuring the metric reflects both openness and the ethical robustness of shared data

This challenge represents an important opportunity to recognize and reward the full spectrum of scholarly contributions, and I believe your team is exceptionally well-positioned to offer a thoughtful and impactful solution. I look forward to supporting your work and contributing however I can.

Sincerely,



Mohammad Hosseini, MA, PhD

Assistant Professor, Preventive Medicine (Biostatistics and Informatics)
Associate Editor, *Accountability in Research*
Northwestern University Feinberg School of Medicine

FASEB
Federation of American Societies for Experimental Biology

Representing Over 110,000 Researchers

6120 Executive Blvd., Suite 230, Rockville, MD 20852 | faseb.org

Katy Börner, Indiana University
Amy Brand, MIT Press
Kristi Holmes, Northwestern University
Daniel Hook, Digital Science
John Inglis, bioRxiv/medRxiv
Chonnettia Jones, Addgene
Arjun Krishnan, University of Colorado Anschutz Medical Campus
Steinn Sigurdsson, arXiv
Griffin Weber, Harvard

May 28, 2025

To whom it may concern:

On behalf of the Federation of American Societies for Experimental Biology (FASEB), we are pleased to offer our strong support for your team's proposal to the NIH Data Sharing (S-Index) Challenge. We also welcome the opportunity to collaborate in this important initiative, which aligns closely with FASEB's mission to advance health and well-being by promoting research and education in the biological and biomedical sciences.

FASEB is a coalition of 22 scientific societies, representing more than 110,000 researchers worldwide across a wide range of disciplines in the biological and biomedical sciences. As a leading voice in the biomedical research community, FASEB advocates for effective policies and practices that support scientific integrity, equitable access to research outputs, and recognition of diverse forms of scholarship—including data sharing and stewardship.

FASEB would be pleased to support your work in several ways, including participating in and helping convene stakeholder workshops to ensure the metric reflects the diverse needs and perspectives of researchers across disciplines and career stages. Additionally, FASEB can contribute to community engagement and communication strategies aimed at building awareness and promoting the adoption of responsible data sharing metrics. Furthermore, FASEB is well-positioned to offer guidance on aligning the proposed metric with broader policy recommendations and infrastructure development priorities within the life sciences.

FASEB looks forward to collaborating with your team to support the development and implementation of a data sharing S-index that truly reflects the values and contributions of the scientific community.

Sincerely,

[redacted]

Frank J. Krause, MBA, CAE
Executive Director and Chief Executive Officer, FASEB

[redacted]

Darla P. Henderson, PhD
Chief Publishing, Open Science, and Research Integrity Officer, FASEB

May 27, 2025

Katy Börner, Indiana University
Amy Brand, MIT Press
Kristi Holmes, Northwestern University
Daniel Hook, Digital Science
John Inglis, bioRxiv/medRxiv
Chonnettia Jones, Addgene
Arjun Krishnan, University of Colorado
Steinn Sigurdsson, arXiv
Griffin Weber, Harvard

RE: Letter of Support and Collaboration for NIH Data Sharing Challenge Proposal

To Whom It May Concern,

As a group of academic health science library leaders deeply engaged in supporting biomedical research, data stewardship, and scholarly communication, we write to express our strong support for your team and your submission to the NIH Data Sharing S-Index Challenge. We also welcome the opportunity to collaborate in this vital initiative to advance equitable and responsible data sharing practices across the health research ecosystem.

Academic health sciences libraries play a critical role in facilitating access to research data, advising on data management and sharing practices, and supporting open science across clinical and biomedical disciplines. The development of a meaningful, scalable metric that recognizes and incentivizes data sharing is of significant importance to our community, particularly in an era where reproducibility, transparency, and equitable access to research outputs are central to scientific integrity and public trust.

Your commitment to stakeholder engagement and inclusive metric design aligns closely with the work of health sciences libraries on our campuses where we frequently play a vital role in partnering with university leadership to advance more responsible approaches to research assessment and reward. Leveraging library expertise in metadata, persistent identifiers, and open infrastructure, libraries contribute to the technical and data foundations necessary for responsible evaluation. At the same time, we support the social and cultural dimensions of change by promoting transparency and recognition of the range of research outputs produced by researchers, such as data, software, and community engagement. Through their expertise and close working relationships with campus researchers, libraries can (and do!) help shape institutional policies and practices that better align with the values of open science and integrity.

We look forward to participating in the stakeholder workshops and can also offer your team insights and guidance across several key areas. Our expertise includes data management, metadata standards, and the use of persistent identifiers within biomedical research. We can inform outreach and engagement efforts with researchers and clinicians to ensure the resulting

metric is appropriate and relevant across domains. Additionally, we are prepared to share insights from institutional data repositories and data literacy initiatives, as well as contribute strategies for integrating the metric into academic research assessment and compliance processes.

We are happy to support this effort.

Sincerely,

[redacted]

Melissa De Santis
Director
Strauss Health Sciences Library
University of Colorado Anschutz Medical Campus
Aurora, CO

[redacted]

Len Levin
Interim Director
Francis A. Countway Library of Medicine
Harvard Medical School
Boston, MA

[redacted]

Kristi Holmes
Director, Galter Health Sciences Library
Northwestern University
Chicago, IL

[redacted]

Jenny Muilenburg
Senior Director
Health Sciences Library
University of Washington
Seattle, WA

[redacted]

Bart Ragon
Director
Claude Moore Health Sciences Library
University of Virginia
Charlottesville, VA

[redacted]

Melissa L. Rethlefsen
Professor; Executive Director
Health Sciences Library and Informatics Center
University of New Mexico
Albuquerque, NM

[redacted]

Gabe Rios
Library Director
Ruth Lilly Medical Library
Indiana University School of Medicine
Indianapolis, IN

[redacted]

Anne Seymour
Associate Dean of the Welch Medical Library
Johns Hopkins University
Baltimore, MD

[redacted]

Donghua Tao
Professor and Associate Dean
Associate University Librarian for Health Sciences
Library of Health Sciences
University of Illinois at Chicago
Chicago, IL

[redacted]

Debra A. Werner
Interim Director, John Crerar Library
University of Chicago
Chicago, IL

[redacted]

Terrie Wheeler
Director of the Samuel J. Wood Library and
C.V. Starr Biomedical Information Center
Weill Cornell Medicine
New York, NY

208. Center for Open Science

I am responding: On behalf of an organization

Name: Maryam Zaringhalam

Name of Organization: Center for Open Science

Type of Organization: Non-Profit Research Organization

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: The Center for Open Science (COS) is a nonprofit organization with a mission to increase the openness, integrity, and trustworthiness of research. Since our founding, we have led several high-profile metascience efforts investigating the reproducibility and replicability of research, including the Reproducibility Project: Psychology, the Reproducibility Project: Cancer Biology, and the Systematizing Confidence in Open Research and Evidence (SCORE) initiative. Across these efforts, we have found that establishing reproducibility and replicability is difficult even under ideal conditions. A failure to reproduce or replicate a finding is not synonymous with misconduct or failure; rather, it is part of how science self-corrects through ongoing inquiry, debate, and refinement. NIH should underscore this nuance in its messaging around rigor and reproducibility, as fears that a lack of repeatability will trigger punitive rather than constructive responses work against the culture change NIH seeks. Nevertheless, NIH has several opportunities to pilot or strengthen programs, policies, and investments to promote rigor and reproducibility.

Reproducibility refers to investigating the same question as a prior study with the same data using the same analysis (https://doi.org/10.31222/osf.io/mqfp4_v1). Reproducibility is a "baseline" form of repeatability — verifying that reported results are repeatable using the same data and analysis. Increasing access to the data, materials, protocols, and analysis underlying a finding is therefore critical to NIH's goals for reproducibility, as we discuss further in our response to "Data, Software, and Model Sharing." Indeed, our SCORE initiative found that papers published in journals with strong data and code sharing policies were more readily reproducible (https://doi.org/10.1038/s41586-026-10251-x). NIH's Data Management and Sharing Policy is a strong step, but NIH has an opportunity to go further by requiring that other elements necessary to reproduce findings are made openly accessible in an appropriate repository and linked to resulting outcomes. These linkages trace findings back to the raw materials and analytic choices that produced discoveries. NIH should invest in infrastructures that enable these linkages across the research lifecycle. For example, COS hosts the Open Science Framework (OSF), an open-source infrastructure that enables researchers to register their study plans and link them to associated data, materials, code, and outcomes in the form of preprints and peer-reviewed articles. OSF thus provides a lifecycle view of a research project, surfacing elements needed to support reproduction, and potentially provide trust markers for findings demonstrated to be reproducible.

Replicability refers to investigating the same question as a prior study with new data. Replicability is a more challenging form of repeatability because one must minimally understand what are the sufficient

conditions to observe a prior finding. Also, it is possible that the prior finding was observed spuriously. As for reproducibility, open sharing of materials, protocols, analyses, and other components of the prior research helps establish replicability by making it easier to scrutinize and reuse the prior research to inform replication efforts. Also, when outcomes differ between original discoveries and replications, transparency of the methods underlying the research supports identification of possible causes. Finally, preregistration of designs and analysis plans supports calibrating confidence in original claims by making transparent actions that may reduce replicability such as selective reporting and p-hacking.

Infrastructures and policies that promote access to research plans and outputs underlying discoveries should also be paired with NIH investments in capacity building for the research community. This can include direct support for training programs that teach skills needed to conduct reproducible and replicable research — including preregistration where appropriate, data and code sharing, and transparent reporting — or incorporating open science practices in required Responsible Conduct of Research training. Critically, reproducibility or replication should not be treated as some boring, rote, or ancillary research activity, but part of the ordinary process by which scientific claims are tested, refined, and generalized (https://doi.org/10.1371/journal.pbio.3000691). Every study extends findings to conditions that differ from those originally examined, so testing whether they hold is foundational to any work that builds on them. Outside of training programs, NIH can further incentivize reproducibility and replication efforts by supporting dedicated replication funding mechanisms and by recognizing replication and reproduction efforts in how it evaluates researchers and research programs.

NIH can also incentivize Registered Reports, a publishing format in which study plans undergo peer review before data are collected, and in-principle acceptance is granted based on the rigor of the question and methods rather than the results. By separating the evaluation of research design from its outcomes, Registered Reports directly reduce publication bias and questionable research practices, such as selective reporting and post hoc hypothesizing, while ensuring that rigorous studies enter the literature regardless of outcome. Uptake of Registered Reports and preregistration offers NIH a concrete, measurable indicator of rigor tied to the quality of the research design rather than the volume or citation count of resulting publications. NIH could pilot Registered Report funding mechanisms, in which peer review of a proposal's design is coupled to a commitment to fund and publish the resulting work, aligning the funding, execution, and publication of research around methodological rigor. Such pilots are useful for building awareness and engagement in scholarly communities that are unfamiliar with the approach — and test the adaptability of the model to new methods and research domains. For example, the Registered Reports model was introduced in a special issue of the journal Social Psychology in 2014 that fostered substantial engagement, refined the model, and promoted adoption by other journals and funders (https://doi.org/10.1027/1864-9335/a000192).

On the question of defining causal research, we note that regardless of how causal claims are categorized, their credibility depends heavily on research design and transparency. Practices such as preregistration and Registered Reports strengthen causal inference by committing researchers to analytic specifications before outcomes are observed, reducing the flexibility that can produce spurious or overstated claims, causal or otherwise.

Any investments NIH makes in reproducibility and rigor should be paired with metascience research to

evaluate the effectiveness of interventions to improve reproducibility and replicability — and to evaluate the costs and benefits of such practices across different disciplines, methodologies, and research contexts. Further, NIH can invest in consistent, sustained replication efforts using a standardized sampling protocol so that replicability can be systematically measured and improvements tracked over time. This ongoing replicability effort should use a standardized sampling method that is determined ahead of time so that the findings are generalizable to the defined population, no single community or methodology receives any more or less scrutiny than deserved, and so that consistent measures can inform the research community with the least amount of bias possible.

2. Data, Software, and Model Sharing: As discussed in our response to "Rigor and Reproducibility," data, software, and model sharing are critical to validating research results and enabling reuse. COS recommends that NIH extend this goal beyond data, software, and models to the protocols, analysis plans, and materials used to generate a research claim — the elements needed to verify how a finding was produced and to build on it (https://doi.org/10.7554/eLife.67995). These outputs are captured in the Transparency and Openness Promotion (TOP) Guidelines, a policy framework led by COS and the TOP Advisory Board that provides recommendations for journals, research institutions, and funders to increase the verifiability of empirical research claims (https://doi.org/10.1186/s41073-026-00223-0). TOP offers a tiered and modular structure for implementing open science policies at three levels of stringency: (1) disclosure that the practice occurred, (2) sharing and citing the open research output, and (3) certification by an independent third party that best practices were followed. COS is continuing to develop implementation guidance for the TOP Guidelines aimed at various communities, including funders, and welcomes engagement on translating this framework into practice.

TOP's broad applicability creates a structured framework for conversations around policies and incentives across research-performing institutions, publishers, and funders — enabling alignment across these sectors so researchers can focus on advancing robust research sharing practices. In addition, policymakers can leverage TOP's modularity to adopt the level of stringency most appropriate to a given field, practice, and the community's readiness to implement it. This flexibility lets NIH recognize and implement policies around sharing in a way that reflects disciplinary differences rather than applying a uniform requirement across all research it funds. COS recommends NIH leverage TOP as a convening tool to align policies and incentives across the disciplinary communities it supports, including by engaging scientific societies and leveraging NIH's own intramural research program as a testbed to pilot incentives and research assessment reform.

This practice-based approach to assessment is increasingly shared across the research system. A recent multi-interest-holder consensus effort, which included COS, identified six core practices that research-performing organizations can monitor and reward — including TOP practices of data, code, and material sharing — as a means of realigning incentives away from publication counts and journal prestige toward research credibility and transparency. As a funder and a research-performing organization, NIH is well positioned to reinforce this movement by aligning its own assessment and reporting expectations with these practices, sending a consistent signal across the institutions and researchers it supports.

NIH should also increase engagement with additional community-driven efforts aimed at incentivizing and rewarding open science and research sharing activities, including the Declaration on Research Assessment, the Coalition for Advancing Research Assessment, Make Data Count, and the Roundtable on

Aligning Incentives for Open Scholarship. These long-standing efforts have a wealth of experiences NIH can draw upon as it develops its own policy and rewards programs. NIH can also draw on examples of funding programs within the government that have made efforts to incentivize meaningful research assessment reform within institutions. For example, the National Science Foundation's Accelerating Research Translation (ART) program, though focused on research translation, conditions its awards on recipient institutions revising their promotion and tenure criteria to reward a broader range of scholarly contributions on par with traditional research outputs. NIH could adapt this model — using funding conditions or incentives to encourage the institutions it supports to recognize and reward rigorous, open, and repeatable research practices in how they assess researchers.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Research that is open, verifiable, and subject to correction is a key contributor for earning and maintaining public trust. The comments we have shared on rigor, reproducibility, and sharing of research outputs are thus foundational to that trust. In assessing public impact, NIH should recognize that practices strengthening the credibility of the research record are themselves contributions to public impact. While producing trustworthy research is not sufficient on its own to secure public trust, practices to promote the trustworthiness of research are within the research community's control — and well within NIH's ability to incentivize. COS and colleagues have advanced a framework tying trustworthiness to measurable indicators, offering one foundation for NIH (https://doi.org/10.1073/pnas.2536736123).

At the same time, NIH should be cautious of overly simplistic measures of public impact — or impact more broadly. Just as publication counts and journal prestige are gameable proxies that can offer the appearance of quality in place of genuinely rigorous research, indicators that reward attention rather than substance, including media mentions, altmetrics, or other measures of visibility, risk rewarding researchers who promote their work effectively over those who do the most rigorous work. NIH should instead prioritize indicators that reflect verifiable contribution and demonstrable use. These indicators should be: driven and tested by the community; transparent and auditable so they are understandable by the community and broader audiences; open, non-commercial, and non-proprietary; contextualized and complemented with narrative descriptions where needed; and iterative and subject to evaluation and change over time.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/COS-Response-to-NIH-Measuring-and-Rewarding-Scientific-Impact-RFI.pdf

Description: COS Response to NIH RFI on Measuring and Rewarding Scientific Impact with citations

The Center for Open Science's Response to NIH Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Submitted on behalf of the Center for Open Science (COS) by Maryam Zaringhalam, Senior Director of Policy, David Mellor, Senior Policy Analyst, and Brian Nosek, Executive Director.

We are writing today on behalf of the Center for Open Science (COS), a non-profit organization working to increase openness, integrity, and reproducibility of research. To meet that objective, COS employs a systems strategy involving technology to make open practices possible and easy, community building to make open practices normative, and stakeholder engagement to make open practices rewarding and required. COS enacts this Theory of Change with mutually reinforcing activities: providing infrastructure through the Open Science Framework (OSF), enabling researchers to register study plans, publish preprints through community-run services, and link externally to related data, materials, code, publications, and other results; conducting research on research (also known as metascience) to understand contributors to repeatability, efficiency, and integrity; and advocating for policies to support open science practices and culture change. COS appreciates the opportunity to draw from this work to provide comments on NIH's approach to measuring and rewarding scientific impact.

Rigor and Reproducibility

> *Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings. NIH also seeks comment on rigor, including with regard to defining causal research.*

The Center for Open Science (COS) is a nonprofit organization with a mission to increase the openness, integrity, and trustworthiness of research. Since our founding, we have led several high-profile metascience efforts investigating the reproducibility and replicability of research, including the Reproducibility Project: Psychology, the Reproducibility Project: Cancer Biology, and the Systematizing Confidence in Open Research and Evidence (SCORE) initiative. Across these efforts, we have found that establishing reproducibility and replicability is difficult even under ideal conditions. A failure to reproduce or replicate a finding is not synonymous with misconduct or failure; rather, it is part of how science

self-corrects through ongoing inquiry, debate, and refinement. NIH should underscore this nuance in its messaging around rigor and reproducibility, as fears that a lack of repeatability will trigger punitive rather than constructive responses work against the culture change NIH seeks. Nevertheless, NIH has several opportunities to pilot or strengthen programs, policies, and investments to promote rigor and reproducibility.

Reproducibility refers to investigating the same question as a prior study with the same data using the same analysis.[1] Reproducibility is a "baseline" form of repeatability — verifying that reported results are repeatable using the same data and analysis. Increasing access to the data, materials, protocols, and analysis underlying a finding is therefore critical to NIH's goals for reproducibility, as we discuss further in our response to "Data, Software, and Model Sharing." Indeed, our SCORE initiative found that papers published in journals with strong data and code sharing policies were more readily reproducible.[2] NIH's Data Management and Sharing Policy is a strong step, but NIH has an opportunity to go further by requiring that other elements necessary to reproduce findings are made openly accessible in an appropriate repository and linked to resulting outcomes. These linkages trace findings back to the raw materials and analytic choices that produced discoveries. NIH should invest in infrastructures that enable these linkages across the research lifecycle. For example, COS hosts the Open Science Framework (OSF), an open-source infrastructure that enables researchers to register their study plans and link them to associated data, materials, code, and outcomes in the form of preprints and peer-reviewed articles. OSF thus provides a lifecycle view of a research project, surfacing elements needed to support reproduction, and potentially provide trust markers for findings demonstrated to be reproducible.

Replicability refers to investigating the same question as a prior study with new data. Replicability is a more challenging form of repeatability because one must minimally understand what are the sufficient conditions to observe a prior finding. Also, it is possible that the prior finding was observed spuriously. As for reproducibility, open sharing of materials, protocols, analyses, and other components of the prior research helps establish

[1] Nosek, B. A., Errington, T. M., Haber, N., Stankov, T., & Tyner, A. H. (2025, December 2). A Brief Glossary of Terms about Repeatability: Replicability, Robustness, and Reproducibility. https://doi.org/10.31222/osf.io/mqfp4_v1

[2] Brodeur, A., Mikola, D., Cook, N., Fiala, L., Brailey, T., Briggs, R., De Gendre, A., Dupraz, Y., Gabani, J., Gauriot, R., Haddad, J., Lima, G., Ankel-Peters, J., Dreber, A., Campbell, D., Kattan, L., Marino Fages, D., Mierisch, F., Sun, P., … Zhong, Y. (2026). Reproducibility and robustness of economics and political science research. *Nature*, *652*(8108), 151–156. https://pure.strath.ac.uk/ws/portalfiles/portal/332550176/Brodeur-etal-Nature-2026-Reproducibility-and-robustness-of-economics-and-political-science-research.pdf. https://doi.org/10.1038/s41586-026-10251-x

replicability by making it easier to scrutinize and reuse the prior research to inform replication efforts. Also, when outcomes differ between original discoveries and replications, transparency of the methods underlying the research supports identification of possible causes. Finally, preregistration of designs and analysis plans supports calibrating confidence in original claims by making transparent actions that may reduce replicability such as selective reporting and p-hacking.

Infrastructures and policies that promote access to research plans and outputs underlying discoveries should also be paired with NIH investments in capacity building for the research community. This can include direct support for training programs that teach skills needed to conduct reproducible and replicable research — including preregistration where appropriate, data and code sharing, and transparent reporting — or incorporating open science practices in required Responsible Conduct of Research training. Critically, reproducibility or replication should not be treated as some boring, rote, or ancillary research activity, but part of the ordinary process by which scientific claims are tested, refined, and generalized.[3] Every study extends findings to conditions that differ from those originally examined, so testing whether they hold is foundational to any work that builds on them. Outside of training programs, NIH can further incentivize reproducibility and replication efforts by supporting dedicated replication funding mechanisms and by recognizing replication and reproduction efforts in how it evaluates researchers and research programs.

NIH can also incentivize Registered Reports, a publishing format in which study plans undergo peer review before data are collected, and in-principle acceptance is granted based on the rigor of the question and methods rather than the results. By separating the evaluation of research design from its outcomes, Registered Reports directly reduce publication bias and questionable research practices, such as selective reporting and post hoc hypothesizing, while ensuring that rigorous studies enter the literature regardless of outcome. Uptake of Registered Reports and preregistration offers NIH a concrete, measurable indicator of rigor tied to the quality of the research design rather than the volume or citation count of resulting publications. NIH could pilot Registered Report funding mechanisms, in which peer review of a proposal's design is coupled to a commitment to fund and publish the resulting work, aligning the funding, execution, and publication of research around methodological rigor. Such pilots are useful for building awareness and engagement in scholarly communities that are unfamiliar with the approach — and test the

[3] Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

adaptability of the model to new methods and research domains. For example, the Registered Reports model was introduced in a special issue of the journal *Social Psychology* in 2014 that fostered substantial engagement, refined the model, and promoted adoption by other journals and funders.[4]

On the question of defining causal research, we note that regardless of how causal claims are categorized, their credibility depends heavily on research design and transparency. Practices such as preregistration and Registered Reports strengthen causal inference by committing researchers to analytic specifications before outcomes are observed, reducing the flexibility that can produce spurious or overstated claims, causal or otherwise.

Any investments NIH makes in reproducibility and rigor should be paired with metascience research to evaluate the effectiveness of interventions to improve reproducibility and replicability — and to evaluate the costs and benefits of such practices across different disciplines, methodologies, and research contexts. Further, NIH can invest in consistent, sustained replication efforts using a standardized sampling protocol so that replicability can be systematically measured and improvements tracked over time. This ongoing replicability effort should use a standardized sampling method that is determined ahead of time so that the findings are generalizable to the defined population, no single community or methodology receives any more or less scrutiny than deserved, and so that consistent measures can inform the research community with the least amount of bias possible.

Data, Software, and Model Sharing

> *Sharing scientific data and tools accelerates biomedical research discovery, in part, by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies. NIH seeks to incentivize and reward these activities to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.*

As discussed in our response to "Rigor and Reproducibility," data, software, and model sharing are critical to validating research results and enabling reuse. COS recommends that NIH extend this goal beyond data, software, and models to the protocols, analysis plans, and materials used to generate a research claim — the elements needed to verify how a

[4] Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

finding was produced and to build on it.[5] These outputs are captured in the Transparency and Openness Promotion (TOP) Guidelines, a policy framework led by COS and the TOP Advisory Board that provides recommendations for journals, research institutions, and funders to increase the verifiability of empirical research claims.[6] TOP offers a tiered and modular structure for implementing open science policies at three levels of stringency: (1) disclosure that the practice occurred, (2) sharing and citing the open research output, and (3) certification by an independent third party that best practices were followed. COS is continuing to develop implementation guidance for the TOP Guidelines aimed at various communities, including funders, and welcomes engagement on translating this framework into practice.

TOP's broad applicability creates a structured framework for conversations around policies and incentives across research-performing institutions, publishers, and funders — enabling alignment across these sectors so researchers can focus on advancing robust research sharing practices. In addition, policymakers can leverage TOP's modularity to adopt the level of stringency most appropriate to a given field, practice, and the community's readiness to implement it. This flexibility lets NIH recognize and implement policies around sharing in a way that reflects disciplinary differences rather than applying a uniform requirement across all research it funds. COS recommends NIH leverage TOP as a convening tool to align policies and incentives across the disciplinary communities it supports, including by engaging scientific societies and leveraging NIH's own intramural research program as a testbed to pilot incentives and research assessment reform.

This practice-based approach to assessment is increasingly shared across the research system. A recent multi-interest-holder consensus effort, which included COS, identified six core practices that research-performing organizations can monitor and reward — including TOP practices of data, code, and material sharing — as a means of realigning incentives away from publication counts and journal prestige toward research credibility and transparency.[7] As a funder and a research-performing organization, NIH is well positioned

[5] Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*, *10*, e67995. https://doi.org/10.7554/eLife.67995

[6] Grant, S., Corker, K. S., Mellor, D., Stewart, S. L. K., Mayo-Wilson, E., Cashin, A. G., Lagisz, M., Moher, D., Umpierre, D., Barbour, V., Buck, S., Collins, G. S., Hazlett, H. F., Hrynaszkiewicz, I., Lee, C. J., Parker, T. H., Rethlefsen, M. L., Toomey, E., Nosek, B. A., & the TOP Advisory Board. (2026). TOP 2025: An update to the Transparency and Openness Promotion Guidelines. *Research Integrity and Peer Review*, *11*(1), 40. https://doi.org/10.1186/s41073-026-00223-0

[7] Moher, D., Cobey, K. D., Armond, A. C. V., Brown, A. N., Cenci, M. S., Chan, L., Gambelli, L., Halas, M., Fox, S. H., Hoffman, M., Hrynaszkiewicz, I., Flake, J. K., Jalil, B., Kip, M. J., Lalu, M. M., Macleod, M. R., Manns, B., Martone, M. E., Naudet, F., … Mellor, D. T. (2026). *Responsible Research Assessment of*

to reinforce this movement by aligning its own assessment and reporting expectations with these practices, sending a consistent signal across the institutions and researchers it supports.

NIH should also increase engagement with additional community-driven efforts aimed at incentivizing and rewarding open science and research sharing activities, including the Declaration on Research Assessment, the Coalition for Advancing Research Assessment, Make Data Count, and the Roundtable on Aligning Incentives for Open Scholarship. These long-standing efforts have a wealth of experiences NIH can draw upon as it develops its own policy and rewards programs. NIH can also draw on examples of funding programs within the government that have made efforts to incentivize meaningful research assessment reform within institutions. For example, the National Science Foundation's Accelerating Research Translation (ART) program, though focused on research translation, conditions its awards on recipient institutions revising their promotion and tenure criteria to reward a broader range of scholarly contributions on par with traditional research outputs. NIH could adapt this model — using funding conditions or incentives to encourage the institutions it supports to recognize and reward rigorous, open, and repeatable research practices in how they assess researchers.

Public Impact

> *NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.*

Research that is open, verifiable, and subject to correction is a key contributor for earning and maintaining public trust. The comments we have shared on rigor, reproducibility, and sharing of research outputs are thus foundational to that trust. In assessing public impact, NIH should recognize that practices strengthening the credibility of the research record are themselves contributions to public impact. While producing trustworthy research is not sufficient on its own to secure public trust, practices to promote the trustworthiness of research are within the research community's control — and well within NIH's ability to incentivize. COS and colleagues have advanced a framework tying trustworthiness to measurable indicators, offering one foundation for NIH.[8]

Faculty: Maximizing quality, transparency, and trustworthiness of scientific research in Research-Performing Organizations. MetaArXiv. https://doi.org/10.31222/osf.io/4qmua_v2

[8] Nosek, B. A., Allison, D. B., Jamieson, K. H., McNutt, M., Nielsen, A. B., & Wolf, S. M. (2026). A framework for assessing the trustworthiness of scientific research findings. *Proceedings of the National Academy of Sciences*, *123*(6), e2536736123. https://doi.org/10.1073/pnas.2536736123

At the same time, NIH should be cautious of overly simplistic measures of public impact — or impact more broadly. Just as publication counts and journal prestige are gameable proxies that can offer the appearance of quality in place of genuinely rigorous research, indicators that reward attention rather than substance, including media mentions, altmetrics, or other measures of visibility, risk rewarding researchers who promote their work effectively over those who do the most rigorous work. NIH should instead prioritize indicators that reflect verifiable contribution and demonstrable use. These indicators should be: driven and tested by the community; transparent and auditable so they are understandable by the community and broader audiences; open, non-commercial, and non-proprietary; contextualized and complemented with narrative descriptions where needed; and iterative and subject to evaluation and change over time.

209. Cherie Marvel

I am responding: On behalf of myself

Name: Cherie Marvel

Name of Organization: Johns Hopkins University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: While I agree that reproducibility and replicability are important aspects of scientific rigor, these aspect should not become prohibitive. By this I mean that some studies, especially those involving human cognition/behavior/physiology, etc., can be difficult to replicate because people are heterogeneous. One way to overcome this is by using extremely large sample sizes, but due to financial constraints, this is not always possible. A requirement of/emphasis on the replication of such studies, and even more so in patient-oriented studies, may disincentivize researchers from attempting to take the first step in this line of research (i.e., by deciding not to submit such proposals at all). Instead, a special RGP category could be useful that acknowledges the difficulty in replication for such human patient studies when a first step in a new direction is desired. Early results could be useful once published, with other investigators able to draw upon those findings to attempt reproducibility, replication, and extension of the original report.

2. Data, Software, and Model Sharing: The best way to uphold scientific integrity is through transparency, especially transparency of methods. Having a clear understanding of how a study was conducted allows independent researchers to assess the validity of the scientific approach. It may be the case that a small aspect of the protocol can make a big difference in results. Thus, in addition to sharing datasets and tools, researchers could be incentivized to share details of the methods (the process) as well. Typically, this information is glossed over in reports, especially when there are word limits. However, researchers could be encouraged to include method details in a repository with the data and tools. For example, such information could include the screening form used to determine patient eligibility, locally developed questionnaires, the MRI exam card, or the even a study visit protocol checklist. When this level of information is shared openly, it allows other researchers to critique the design and attempt to replicate the process in the most accurate way.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

210. Open Research Funders Group

I am responding: On behalf of an organization

Name: Emily Ford

Name of Organization: Open Research Funders Group

Type of Organization: Other

Type of Organization - Other: Coalition of non-governmental research funders

Role: Organizational Official

Date Submitted: 8/18/2026

1. Rigor and Reproducibility: See attached

2. Data, Software, and Model Sharing: See attached

3. Training and Mentorship: N/A

4. Collaboration: See attached

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/For-submission-NIH-RFI-on-Measuring-and-Rewarding-Scientific-Impact-ORFG.pdf

Description: Narrative response from the ORFG addressing rigor and reproducibility, data and code sharing, and colalboration.

To: National Institutes of Health

From: Open Research Funders Group (ORFG)

RE: Response to RFI on Measuring and Rewarding Scientific Impact [NOT-OD-26-087]

Date: 19 August 2026

Please consider the attached as a formal response to the NIH RFI on Measuring and Rewarding Scientific Impact from the Open Research Funders Group.

Open Research Funders Group (ORFG) is a coalition of private philanthropic funders committed to the open sharing of research outputs. Collectively, ORFG members hold assets in excess of $152 billion, with total annual grantmaking in the $11 billion range. We believe openness is better for the American public, better for research, better for global society, and better for philanthropy. Open research accelerates the pace of discovery, reduces information-sharing gaps, encourages innovation, and promotes reproducibility.

As members of the private philanthropic community, we strive to support the rapid and wide dissemination of scientific outputs that serve our society. This community uses a variety of strategies and approaches to increase scientific discovery and continuously works to improve the scientific ecosystem. We have knowledge and expertise in funding scientific outputs – such as preprints, datasets, code, and tangible materials – as well as assessing their impacts.

Much like the NIH, we care deeply about rewarding and incentivizing rigorous research, and we share the view that current metrics do not work as intended.[1]

Current incentive structures fundamentally misalign with the agency's core mission. Researchers are generally rewarded for publication volume, journal prestige, and citation counts rather than for rigorous, reproducible, and collaborative science. This misalignment is the central barrier to the knowledge generation that enhances health and reduces illness. To serve the American public, we encourage the NIH to revise its funding mechanisms and review processes to enhance its evaluation processes by measuring and

[1] Declaration on Research Assessment (DORA), *Guidance on the Responsible Use of Quantitative Indicators in Research Assessment*, September 26, 2024, https://doi.org/10.5281/zenodo.13844873.

rewarding behaviors that strengthen the scientific knowledge base: transparency, open sharing, and investment in research infrastructure.

There are three main areas listed in this RFI that are of utmost importance to ORFG:

1. Rigor and reproducibility
2. Data, software, and model sharing; and
3. Collaboration.

The most effective way to ensure that NIH can better align its measures of research quality with its core mission, and generate knowledge that enhances health, lengthens life, and reduces illness and disability, is to **institute open sharing practices and behaviors, and reward and incentivize these behaviors in new and meaningful ways.** This can and should be achieved using a mix of several levers: policy, rules, workflows, and experimentation with new and innovative assessment and evaluation methods, including qualitative research assessments of individual contributions, and granular outputs.

However, not all approaches to open sharing and incentivization are as impactful and effective as they need to be, and are executed in ways that create unintended and deleterious consequences.

The remainder of this memo outlines ORFG's perspective on how open output sharing and incentivization will best serve the NIH and its mission to serve the American public, and minimize unintended consequences.

Rigor and Reproducibility

Reproducibility at scale requires shared standards and infrastructure, not just individual transparency. Openly sharing research outputs such as preprints, datasets, registered protocols, and code enables the transparent, collaborative peer review and validation necessary for rigorous science. However, this sharing must be supported by community-governed infrastructure (preprint servers, data repositories, registries) and standardized metadata that allows for discovery, interoperability, and reuse. A recent study points to open peer review as a useful research quality indicator; those authors choosing open peer review are more likely to also choose open science practices.[2] While openly sharing alone does not indicate rigor or reproducibility, it enables an environment where rigor and reproducibility can be examined by a large community or researchers. This is why the publish, review, curate model of research sharing is gaining traction in researcher communities.

[2] Adrian Barnett and Matt Spick, "Who Chooses Open Peer Review and Is It an Indicator of Article Quality? An Observational Study of PLOS Journals.," 2b7z5_v1, preprint, MetaArXiv, April 10, 2026, https://osf.io/preprints/metaarxiv/2b7z5_v1/.

Openness coupled with robust infrastructure yields quality reuse and increases/amplifies researcher efforts. Shared outputs must include robust, standardized metadata that enables provenance tracking, machine interoperability, and discovery. This is especially critical as machine learning increasingly drives biomedical research: metadata quality directly enables validation of computational models and responsible AI development.

This metadata is becoming increasingly important as researchers adopt incremental sharing practices known as continuous or modular science. Instead of sharing all results (including data, narrative, code, models, and tangible materials) at the *conclusion* of a study, researchers are sharing these incremental outputs as soon as they are practicably usable by others. One schema that supports this approach is Open Exchange Architecture (OXA). Investment in infrastructures such as schema as well as servers and discovery systems, are critical to the ability for research to be disseminated and reproduced.

Current quantitative incentive structures such as impact factor, h-index, and publication counts are both simplistic and misaligned with the NIH's mission. They measure reputation and citation volume rather than scientific rigor, reuse, or real-world impact. Worse, they perpetuate dependence on for-profit publishers and commercial entities, penalizing researchers for transparent, open work. To align incentives with mission, NIH should adopt non-proprietary, community-governed indicators that measure and reward: (1) preprint and open data publication timelines, (2) participation in open peer review platforms (PREreview, Review Commons), (3) adoption of appropriate transparency standards for the research at hand (OSF registration, registered reports), and (4) replication and replication-focused infrastructure contributions. These shifts should be embedded in grant review rubrics and funding decisions, not merely encouraged.

To reiterate, these behaviors alone do not assure rigor and reproducibility. However, enabling and incentivizing open behaviors creates the condition for the community to be able to evaluate rigor and reproducibility of studies. (It would be very difficult to reproduce a study, where methods, data, and the like are not openly available.) Public peer review of preprints such as those that occur at PREreview, are exemplary of how open enables robust community review and evaluation of rigor.

Data, Software, and Model Sharing

Sharing data, software, and models without infrastructure, standards, or purpose yields limited reuse and creates burden on researchers. Instead, NIH should incentivize and reward sharing that meets community standards: FAIR data principles, robust metadata, open licensing, and demonstrated downstream use. As computational models increasingly drive biomedical research, open licensing of trained models—enabling external validation and collaborative refinement—is essential to reproducibility and scientific integrity.

ORFG recently released a report on *ORFG Funder Policy Landscape Report and Recommendations,* which includes an analysis of data, software, and model sharing practices of private research funders.[3] Among its recommendations is the expansion of data management plan requirements and the need for funders to support existing data principles and practices. Included in this is the use of FAIR data practices and principles that ensure shared data is findable, accessible, interoperable, and reusable.[4]

As an example, Aligning Science Across Parkinson's, discusses FAIR data principles in its open science handbook[5]:

> "Data must be deposited in their rawest reasonable form (e.g., the data have not been cleaned or preprocessed) in a discipline-specific, community-recognized repository that follows FAIR principles. We acknowledge that, depending on the type of data and the repository used, some level of preprocessing or curation may be required before sharing."[6]

Additionally, all shared computational models and software should be openly licensed (e.g., MIT, Apache 2.0) to enable validation, reuse, and adaptation. As researchers increasingly use machine learning models in biomedical discovery, proprietary or closed models limit reproducibility and scientific accountability. NIH should require that funded computational models and ML-derived predictions include: (a) training data specifications, (b) licensing terms that permit validation and derivative use, and (c) metadata documenting model assumptions and limitations. Recognizing data and software contributions equally with publications—through CV sections, grant reviewer rubrics, and annual report metrics—will incentivize quality sharing practices.

Software sharing should follow similar principles: deposit on trusted, community-governed platforms (e.g. Zenodo); use standard open licenses (MIT, Apache 2.0); include documentation of dependencies and computational requirements; and track download/reuse statistics. NIH should measure impact by active reuse such as downloads, forks, and adaptations as one indicator of scientific impact.

In short: data, software, and model sharing should be incentivized based on adherence to community standards (FAIR, open licensing, etc.) and measurable downstream reuse is one indicator of impact. NIH funding mechanisms should track and reward both data reuse and software adoption, recognizing these as primary research contributions alongside publications.

---

[3] Open Research Funders Group, *ORFG Funder Policy Landscape Report and Recommendations: 2026* (Zenodo, 2026), https://doi.org/10.5281/zenodo.20451484.
[4] "FAIR Principles," *GO FAIR*, n.d., accessed February 17, 2026, https://www.go-fair.org/fair-principles/.
[5] Aligning Science Across Parkinson's, "ASAP Open Science Policy Handbook," Zenodo, December 17, 2025, https://doi.org/10.5281/zenodo.17968591.
[6] Aligning Science Across Parkinson's, "ASAP Open Science Policy Handbook."

Some funders, such as The Michael J. Fox Foundation for Parkinson's Research and the Wellcome Trust have implemented similar policies requiring open licensing and deposit of data and software in trusted repositories. Moreover, national funders in the UK, Netherlands, and the EU all support robust data sharing practices. NIH should align with and amplify these efforts.

Collaboration

Team-based science requires new approaches to measure individual contribution. Current publication-based metrics obscure who did what and penalize researchers (statisticians, bioinformaticians, data stewards) whose contributions are essential to rigor but invisible in authorship. Modular outputs—preprints, datasets, code repositories, registered protocols—help make individual contributions transparent and enable earlier visibility into contributor roles. NIH should incentivize collaboration in its awardee assessment and evaluation by recognizing infrastructure contributions (maintaining repositories, creating standards, curating datasets) as primary research contributions in grant review and CV assessment.

Moreover, when considering approaches for assessing and recognizing contributions to team-based science, looking to the open science and open research communities will be key. For example, the Declaration on Research Assessment (DORA) recently released A Practical Guide to Implementing Responsible Research Assessment at Research Funding Organizations stresses the need for research funders themselves to engage in alignment and collaboration in their research assessment approaches and incentives structures. Implementing these changes requires concrete workflow modifications. NIH should: (1) expand SciENcv to allow narrative descriptions of collaboration and team contributions alongside traditional CV sections; (2) allow grant applicants to list modular outputs (datasets, code, protocols, etc.) separately from bundled "publications" as indicators of their scientific contributions; and (3) train grant reviewers to utilize team-based contributions as one indicator of scientific impact.

Narrative CVs—already adopted by funders in the Netherlands, Ireland, Canada, and New Zealand—allow researchers to explain their approach to collaboration, team contributions, and research philosophy rather than relying on publication counts. The NIH has similarly adopted the narrative CV. An additional step to incorporate narrative CVs at NIH would be to modify the SciENcv to support this format at the grant application stage, enabling researchers to highlight modular contributions and collaborative roles before final publication.

As with the other areas addressed by this RFI response, community governed and created metadata and infrastructures enable team science. The CRediT taxonomy allows for

contributor roles on research outputs to be more transparent. The Open Exchange Architecture (OXA) enables transparent attribution across modular outputs, allowing team contributions to be visible before final publication. Combined with the CRediT taxonomy for role specification, OXA creates accountability and visibility into team science at every stage of research, not just at publication.

Finally, NIH should align its assessment and incentive policies with other governmental funders advancing research reform. Initiatives such as the Coalition for Advancing Research Assessment (CoARA,) the Research on Research Institute (RoRI), Declaration on Research Assessment (DORA), and the Global Research Council (GRC) provide templates for policy coordination. This alignment strengthens the collective signal to researchers that hiring, promotion, and grant decisions across funders now reward open, collaborative, team-based science—making individual institutions' adoption easier and more consequential.

Conclusion

To conclude: ORFG urges NIH to align its funding mechanisms, review processes, and workforce policies with a science of rigor, not reputation. This requires three coordinated shifts:

1. Infrastructure Investment. Support community-governed, non-proprietary infrastructure including but not limited to repositories, preprint servers, registries, and schema that enable open standards and shared metadata across the research enterprise.
2. Responsible Indicators. Adopt the use of qualitative research indicators and assessments. Incentivize open science behaviors that enable scientific rigor and reproducibility. Some of these indicators may replace journal-based metrics. Indicators to consider for inclusion are: open peer review participation, FAIR data practices, and the ability to communicate individual and modular contributions to research.
3. Incentive Alignment. Embed these indicators in grant review rubrics, proposal templates (SciENcv), and annual reporting. Recognize and assess individual contributions to teams, considering their modular outputs, infrastructure contributions, and team roles equally with journal publications.

These changes reflect modern science's collaborative, transparency-dependent nature. They better align research assessment with the NIH's core mission: advancing knowledge that enhances health.

211. Elsevier

I am responding: On behalf of an organization

Name: Abi Hilditch

Name of Organization: Elsevier

Type of Organization: Publisher

Role: Other

Role - Other: Policy and Industry Relations Manager

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Elsevier strongly supports NIH's goal to strengthen replication and reproducibility of biomedical research, a key tenet of Gold Standard Science and the foundation of rigorous research. We believe that stakeholders including funders, institutions, publishers, and researchers must collaborate to build a research system that rewards and integrates reproducibility practices.

Prerequisites for reproducibility are full transparency in how studies and experiments were undertaken, the methods and protocols involved, and the results obtained. Publishers play an important role by making continuous and substantial investments in infrastructure, processes and tooling to support transparency across these areas. For example, Elsevier prompts authors to share data, report openly on methods, and experimentation with approaches to disseminate research findings including negative and null results. Publishers, including Elsevier, additionally operationalize trust and transparency in the publishing process by supporting authors to consistently report funding sources, conflicts of interest, and ethics approvals; linking research articles to related data, code, and metadata that place the research in context; and maintaining the trusted version of record along with a clear, public trail of any subsequent corrections or updates.

NIH and publisher investments in fraud prevention and detection are needed alongside replicability and reproducibility to truly guarantee rigor in research. Publishers aim to prevent fraudulent articles from being published by using sophisticated, data-driven technology at all stages of the submission, editorial, and peer review processes to screen for integrity and ethics issues.

These investments are made possible by sustainable publishing models which enable publishers to recoup their costs. Sustainable models mean that publishers can invest in services which ensure the quality, integrity, discoverability, and accessibility of research in perpetuity. We encourage NIH to consider the need for sustainability in the scholarly communication system that underpins reproducibility and rigor in research, when developing its approach to impact measurement.

Building on this knowledge and experience, we propose the following potential indicators to measure rigor and reproducibility : preregistration, published protocols, transparent methods, materials

availability, statistical/code review, correction/versioning practices, replication-enabling outputs such as data and replication studies. Stakeholders from across the research ecosystem, including publishers, will need to collaborate to develop measures and align on successful indicators and incentives, which should avoid unintended consequences, such as those arising from penalising researchers for corrections.

We encourage NIH to see publishers as expert partners in these areas, and to recognize the importance of supporting existing research communication, research measurement, and research documentation systems that enable replication and reproducibility.

2. Data, Software, and Model Sharing: Elsevier agrees that sharing research outputs such as data, methods, and models is an integral part of the research process and extends (re)use of research, driving efficiencies, boosting the potential impact of research, and fostering innovation within academia and industry. We encourage NIH to explore the following examples of best practices when developing incentives for data sharing:

• The Force 11 FAIR principles for data sharing: supporting Findable, Accessible, Interoperable, Reusable data. These principles help optimize the use of data and methods across stakeholders.
• Researchers and institutions should be encouraged to store, share, discover and effectively (re)use data, following the principle "as open as possible, as closed as necessary." Federal agencies already have requirements in place to encourage this. Publishers provide further support by building prompts into the workflows for article submissions for authors to share links to datasets and include information in a repository of their choice.
• At Elsevier, we additionally make available Mendeley Data, an NIH Generalist Repository Ecosystem Initiative (GREI)-supported open and free generalist repository. We strongly support this initiative and would like to continue to partner with NIH on this effort.
• The Center for Open Science (COS) Transparency and Openness Promotion (TOP) developed community based guidelines to promote transparency, open sharing, and reproducibility.
• The Data Sharing Index (S-index) Challenge, led by the National Eye Institute (NEI) and NIH, which aims to incentivize the creation, development, and validation of a quantitative data sharing index.

Based on our experience of supporting and promoting data sharing, potential indicators for measuring data, software and model sharing include: availability of FAIR deposits in generalist and specialised repositories with DOIs, downloads, reuse signals, documentation quality, steps-to-reproduce fields, and citations in articles which provide vital contextual information to aid reproducibility. When considering policies and incentives, NIH should recognize the need to protect sensitive data, acknowledge domain-specific differences, and avoid rewarding volume of deposits over quality and usability.

Elsevier would welcome the opportunity to partner with NIH to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.

FAIR Principles: https://force11.org/info/the-fair-data-principles/
Mendeley Data: https://data.mendeley.com/
The Center for Open Science (COS) Transparency and Openness Promotion Guidelines: https://www.cos.io/initiatives/top-guidelines

The Data Sharing Index (S-index) Challenge: https://www.nih.gov/challenges/nih-data-sharing-index-s-index-challenge

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Demonstrating the public impact of federally funded research is key to strengthening public trust in science. Below, we provide examples of how Elsevier is supporting federal agencies to assess public impact:

• NSF Technology, Innovation and Partnerships (TIP) Investment Explorer is an interactive research intelligence portal and map developed by Elsevier in collaboration with the U.S. National Science Foundation's TIP Directorate. It visualizes NSF TIP awards across institutions, states, and technology domains, using Elsevier's Pure and InsightGraph platforms to integrate award data with broader research information. By linking grants to organizations, research outputs, and innovation areas, the TIP Investment Explorer helps stakeholders track where federally funded research investments are being made and how they contribute to national priorities such as AI, quantum science, biotechnology, and advanced manufacturing. This tool supports monitoring benefits and impacts by showing not just the distribution of funding, but also downstream outcomes, such as creation of jobs and start-ups, and follow-on funding. This provides funders, policymakers, and institutions with evidence of how public research dollars drive innovation, collaboration, economic development, and societal progress.
• Elsevier has also developed the US research and innovation portal, which explores the latest trends in US research excellence, highlighting data on high-quality outputs in five key critical and emerging technology areas identified by the US federal government, including biotechnology. This analysis shows how publication data can be used beyond citations to support analysis of research collaboration networks as well as identification of talented individuals and new potential partnerships, which further aligns with NIH's intention to assess and recognise contributions to team-based science.

Building on our extensive experience in this area, we believe NIH should consider measuring the following specific indicators for public impact: clinical guideline influence, policy mentions, media and podcast attention, public engagement, accessibility, patient/community benefit, partner-attested impact, economic activity, job and start-up creation, and evidence of trust-building or adoption. Given the sensitive nature of public impact of research, we recommend that NIH set-up appropriate safeguards to avoid overstating causality of research outcomes and use contribution analysis and confidence levels when health or societal outcomes involve many actors.

Finally, we encourage NIH to consider the infrastructure which allows research to have public impact. To achieve impact, research must be disseminated and accessible, with processes in place to ensure quality, integrity, discoverability and preservation of the version of record in perpetuity. Publishers make significant investments to support a trusted scholarly communications system, which ensures that articles, data and research outputs can continue to have long-term public impact.

NSF Technology, Innovation and Partnerships (TIP) Investment Explorer: https://nsfmap.services.elsevierpure.com/map
Elsevier US research and innovation portal: https://elsevier.shorthandstories.com/us-research-and-innovation-performance/index.html

8. Other Areas: Elsevier welcomes the opportunity to provide feedback to the National Institutes of Health's request for information on Measuring and Rewarding Scientific Impact. For many years, Elsevier has supported the careful use of metrics and indicators in the evaluation of research and has advocated for a combined qualitative and quantitative approach, including support for initiatives such as the San Francisco Declaration on Reforming Research Assessment (DORA). In line with this, we have endeavoured to share specific indicator suggestions in the areas of Rigor and Reproducibility, Data, Software and Model Sharing, and Public Impact while highlighting how publishers support across these areas.

To ensure an aligned approach to measuring and rewarding scientific impact across priority topics, we further recommend that NIH develops an overarching Impact and Rigor Evidence Framework. This should take a transparent, multidimensional approach that uses a basket of validated indicators, qualitative evidence, expert judgment, and clear guardrails. The framework should be designed for portfolio learning and incentive alignment, not for reductive ranking of individuals or automated funding decisions.

NIH should build the framework around principles that keep metrics useful, interpretable, fair, and reusable. Each indicator should have a documented purpose, definition, data source, update cadence, normalization method, limitations, and intended interpretation. Each output should document where information originated and how it contributed to the indicator to support transparency and enable auditing. No single indicator should be treated as a proxy for scientific quality or public impact.
Finally, any framework should be governed as a reusable evidence asset to support future NIH reporting and evaluation needs, subject to necessary and appropriate review, evolution and updates, rather than a one-time reporting exercise.

We believe NIH's aim to incentivize and reward rigorous research that delivers value to the American public is laudable, and we support the responsible use of indicators of success that reflect the collaborative and mission-driven nature of modern biomedical science. We look forward to collaborating with NIH as it further develops this work.

212. American Statistical Association

I am responding: On behalf of an organization

Name: Steve Pierson

Name of Organization: American Statistical Association

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: See attached for formatted version of the following:
1.1 The missing instrument: indicators for analytic replicability
We are highly encouraged by NIH's renewed focus on transparency, reproducibility, and replication shown in the Request for Information on Measuring and Rewarding Scientific Impact, the Initiative to promote Strengthening Replication and Reproducibility of NIH-funded Research, the Gold Standard implementation plan, and the recent Highlighted Topic on Enhancing Scientific Rigor, Transparency and Replicability. These documents name reproducibility and replicability as distinct goals, correctly, and the Initiative has instruments aimed at each: for the former, longstanding investments in data sharing, standardization, and public access to findings; for the latter, funded replication research through the Common Fund, a Replication Prize, and a stated intent to engage researchers in identifying areas ripe for replication.
These are steps in the right direction. However, we have yet to observe incentives targeting the gap between reproducibility and replicability: whether the reported inference is valid given how the analysis proceeded. Statisticians and biostatisticians have a useful and unusual vantage point on this gap: we are often among the first to recognize why a finding may not replicate, while not always having the standing or involvement needed to address the problem.
This matters because replication studies detect errors, but do not prevent them. They measure the failure rate after the fact, one finding at a time, at a high cost per finding. A study can be perfectly reproducible — fully containerized, one-click re-runnable — and still report inference that is unlikely to replicate, and the replication study will discover this several years and several hundred thousand dollars later.
A common route to this is inference on a model that was chosen using the same data used for inference: unadjusted post-selection inference (UPSI). It is one instance of what Simmons, Nelson, and Simonsohn (2011) termed researcher degrees of freedom: undisclosed flexibility in analytic choices, sufficient on its own to produce significant findings without any intent to deceive. In a satirical blog post, Perry Hackman used simulations with pure-noise outcomes and a modest set of candidate subgroup interactions to show that an ostensibly rigorous method, forward stepwise selection with information criteria followed by standard statistical inference (dubbed UPSI), can produce at least one "significant" effect in 90+% of simulated trials (Peterson, 2025b). Nothing in that pipeline is clearly fraudulent; the analysis plan reads as rigorous, and the resulting code can be shared, containerized, and re-run perfectly.
Recommendation: The Initiative's replication instruments would return more per dollar if paired with a cheap upstream instrument that flags the findings most likely to need them. In analysis plans, progress

reports, and publication reporting, a short structured statement answering the following could be incentivized:

Was the reported model, subgroup, outcome scale, or covariate set chosen using the same data used to compute the reported inference?

If so, how were the inferences adjusted for selection? (sample splitting; selective or conditional inference; bootstrap or model-averaged uncertainty; explicit relabeling as exploratory). If not, where was the full analysis specification verifiably pre-registered?

How many candidate specifications were considered, including those examined and discarded?

Based on the above, results should be taken as on a spectrum between confirmatory (to provide stronger evidence for prespecified claims) and exploratory (to identify patterns and generate hypotheses).

Another recommendation is to commission periodic targeted statistical review of a random sample of supported publications, scored by trained statistical reviewers, reporting the proportion that (a) report selection-conditional inference without adjustment, (b) report subgroup findings without pre-specified subgroup structure, and (c) contain code that does not execute. Tracking that proportion over time measures whether the Initiative is working.

1.2 Candidate pools and model complexity are growing faster than the standards applied

Beneath many variable-selection procedures is an implicit assumption of covariate equipoise: that every candidate predictor is equally worthy of entering the model. Stepwise procedures, popular information criteria searches, and standard lasso implementations can all behave this way, and the assumption becomes problematic in settings where NIH is investing heavily.

Independent groups have converged on the same principle from different directions. Yu, Bien, and Tibshirani (2019) articulate a reluctant interaction selection principle — prefer a main effect over an interaction if all else is equal — and Tay and Tibshirani (2020) extend this method to a reluctant non-linear selection principle, preferring a linear term over a non-linear one on the same grounds. In his standard reference text, Harrell (2015) advises pre-specifying a small set of plausible interactions rather than testing all of them. Peterson and Cavanaugh (2022) advise that human-understandable "glass-box" models with meaningful features should be preferred to more complex ones precisely because they are easier to interrogate and replicate. Different penalties, different algorithms, same underlying claim: complexity should have to earn its way into a model, and treating every candidate term as equally deserving is what allows spurious complexity through. NIH need not adjudicate among these methods, but it should seek to require the reporting that makes the problem visible.

Transparent modeling methods matter because replication requires something to replicate: a claim specific enough that an independent team can test it on new data and report whether it held. Transparent statistical models often make such claims directly available. Waist circumference is associated with one-year HDL, $\beta = -0.04$ (95% CI -0.07 to -0.01) is a statement another group can go and check in a different dataset, and the check has a clear objective.

On the other hand, black-box predictive models, such as neural networks, XGBoost, random forests, and statistical models with high-order interaction terms, instead primarily yield prediction functions. A second team can apply the deposited model to new data and test whether predictive performance holds up. That is a meaningful form of external validation, but it is different from replicating an interpretable scientific claim. When performance drops, the failure may be difficult to localize: a transparent model can reveal which estimated relationship moved and in which direction, while an opaque model may reveal only that predictive performance deteriorated, and, if subgroup metrics are tracked, for whom.

This has a direct implication for the Initiative. Predictive modeling is an increasingly prominent category of NIH-supported research (e.g., Arshi et al. 2025, Le et al. 2024), and it is precisely the category on which replication instruments have the least purchase. Transparency is therefore not only a matter of trust, fairness, or regulatory compliance, though it is those too. It also determines how specifically a scientific claim can be interrogated and replicated.

As datasets grow larger and contain newer, more high-dimensional modalities, new statistical methodology will be critical. For instance, with only twenty candidate predictors, to capture potential interactions, one would need to sift through 190 pairwise interactions, 1,140 three-way interactions, and over a million candidate models across all orders. Further, when modalities differ in their dimension and overall signal, they quickly run into practical challenges. For example, in a lung adenocarcinoma survival analysis with 7 clinical covariates and 22,283 gene expression measurements, assuming covariate equipoise and penalizing both groups equally encodes a structural disadvantage for clinically well-motivated variables, which get de-selected due to the sheer difference in dimensionality of the two modalities. With NIH's data-integration priorities, as studies become multimodal, consisting of genomics plus clinical plus imaging plus wearable plus environmental and more, the need for novel methods that can accommodate such data is critical; important methodological challenges remain at the scale and complexity of these data.

Recommendations: First, NIH should be wary of accepting post-hoc explanation of an opaque model (such as a locally interpretable model, a SHAP-based variable importance metric, etc.) as a reliable, replicable scientific finding; the explainer is a separate model that may or may not describe what the first one does, and should not be treated as equivalent to a directly interpretable scientific claim. Second, the accuracy cost of transparency is an empirical question that has been studied. A systematic review of 71 studies developing clinical prediction models found no performance benefit of machine learning over logistic regression (Christodoulou et al., 2019). Benchmarking work supports this outside the clinical setting as well: across 110 datasets from the Penn Machine Learning Benchmarks database, a transparent regression approach was within 5% of the best-performing method on 79% of classification and 62% of regression datasets, with random forests and gradient boosting showing no statistically distinguishable advantage (Peterson et al., 2024). Where transparent methods fell short, the reasons were diagnosable and specific and not reflective of a general law that complexity buys accuracy (specific examples can be found in Peterson et al., 2024, Section 3.3.2).

The implication is not that black-box methods should be discouraged, but that the burden of justification runs backward. Investigators are rarely asked to show that their model's opacity bought anything. For NIH-funded models intended to inform clinical decisions, a low-cost expectation is that a transparent baseline be fit and reported alongside the complex one, so the cost of interpretability is measured in that application rather than assumed. Where the gap is negligible, as is often the case, the transparent model is the better scientific product, because it can be interrogated and corrected.

1.3 Adequacy at the design stage: making evidentiary expectations auditable

These analysis-stage safeguards are necessary, but inference is only as strong as the data generated by the study design. The design determines which populations, outcomes, contrasts, and sources of variation are observed and therefore which scientific claims the resulting data can support. One of the earliest places this becomes explicit is the sample-size or design justification, where investigators state what they expect the proposed design to be capable of learning.

This makes the justification useful for more than determining a number. It provides a prospective, auditable statement of what adequate evidence was expected to look like. NIH can then ask whether the

realized study provided that evidence and, if not, whether the departures and their implications were documented. A study need not unfold exactly as anticipated. Recruitment, measurement properties, effect sizes, intraclass correlations, or other design features must often be learned during the course of a study, at which point they become scientific information in their own right. Reporting them allows planning assumptions to be updated across studies rather than repeatedly regenerated from incomplete information.

For studies relying on existing or linked data, realized data quality may also differ from what was assumed at design, with implications for which claims the data can support. Rigor requires making those departures visible and reconsidering the claims they permit, rather than retrospectively treating the original justification as though nothing changed.

Gold Standard Science creates a policy demand for determining and justifying adequate sample sizes, but no common definition of “adequate” accompanies it. Raising conventional thresholds does not settle the problem. A study can have 90% power under poorly supported assumptions, for an effect of little scientific consequence, using an inefficient design, or with an imprecisely measured outcome. Adequacy is instead a fit-for-purpose judgment: whether the design, measurement, data quality, and analysis are sufficient for the claims the study is intended to support.

NIH guidance already reflects this principle. Under the current Parent R01 review framework, reviewers assess whether sample size is sufficient and well justified in relation to the research questions, outcomes, and design (NIH, 2024). NCCIH feasibility trials instead require quantitative feasibility or acceptability benchmarks and justification that the sample is sufficient to evaluate them; testing efficacy or “preliminary efficacy” is explicitly not the purpose (NCCIH, 2024a). By contrast, NCCIH efficacy, effectiveness, and pragmatic trials require power calculations linked to endpoints and hypotheses, at least 90% power for the primary outcome, adjustment for multiple primary outcomes, and a clinically meaningful effect size (NCCIH, 2024b). NIH therefore already treats adequacy as purpose-dependent; the opportunity is to make that principle explicit, consistent, and auditable.

The same principle extends beyond these examples. Exploratory secondary analyses, transportability or triangulation studies, resource construction, and methods development may require adequate precision, population overlap, analytic scope, ability to assess discordance, or support for future studies rather than power against a single alternative. Requiring a power calculation regardless can produce the calculation without the justification. This is not a lower bar for exploratory work but a different one: identify the intended claim, the criterion of adequacy it requires, and the corresponding limits on inference.

Design information also interacts with the candidate analysis space. Limited information combined with broad analytic flexibility can generate apparently strong but unstable findings, while each source of risk is largely invisible to instruments aimed only at the other.

Declaring intended claims and criteria of adequacy at the design stage gives the disclosure in §1.1 something to be checked against. A confirmatory aim that later relies on substantial data-dependent selection, or a secondary outcome later presented as a primary claim, has changed evidentiary status during execution. That is not necessarily a scientific failure, provided that its path from "design" to "discussion" is visible and reflected in the claims made.

Recommendations:

First, require fit-for-purpose definitions of adequacy. Applications should identify the claims or objectives the study is intended to support, and the criterion for an adequate sample size or design should reflect that purpose. When an effect size is used to justify adequacy, its magnitude should be grounded where

possible in empirical evidence and interpreted on a scale meaningful to the scientific question. Generic standardized effect-size conventions such as “small,” “medium,” and “large” should not substitute for separately considering the magnitude of effect that would matter and the variability anticipated in the study population. Not every aim of a complex study requires a separate, fully developed power analysis. However, the study should identify the minimum set of claims or objectives it is designed to support with strong evidentiary justification and demonstrate that the proposed design is adequate for those purposes.

Second, treat sample-size justification as an integral part of study design rather than a calculation performed after the design is largely fixed. The amount of information a study can provide depends not only on the number of observations but on interdependent choices about the population recruited, comparisons made, measurements collected, timing and intensity of follow-up, analytic strategy, and other features that can affect the credibility of conclusions. Those choices must also be made within constraints on participant burden, operational feasibility, time, and budget. Improving one dimension may worsen another. Sample-size justification should therefore be an iterative, collaborative process in which statistical, substantive, operational, and other relevant experts identify the major threats to the study’s success, examine how alternative design choices redistribute those threats, and determine how limited resources can be used to produce the strongest evidence for the study’s purpose. This provides a concrete, early example of collaboration that NIH could recognize and measure: integrating statistical, scientific, operational, and other relevant expertise into consequential study-design decisions.

Third, make sample-size and design justification transparent, auditable, and cumulative. Power analyses and other sample-size calculations should be reproducible from the statistical analysis plan, with the method, inputs, assumptions, and software or tools documented sufficiently for an independent reviewer to recreate and vary the calculation. Studies should subsequently compare planning assumptions with what was realized, including recruitment and retention, variability, clustering, measurement performance, and other design quantities. Meaningful departures should be reported along with their implications for the evidence the study can support, rather than treated only as implementation failures. Recording what was expected, what occurred, and how those differences affected the study would allow empirical knowledge about design assumptions to accumulate across studies, so that future investigators can ground their assumptions in better evidence. NIH could track the proportion of funded studies that make their design justifications reproducible and subsequently report comparisons between key planning assumptions and realized study conditions.

1.4 Identifying findings ripe for replication

The Initiative states that NIH will engage researchers in selecting research areas ripe for replication and reproducibility studies. The indicators proposed above provide a practical way to do so.

Statistical fragility can triage more effectively than citation count alone. A tractable prioritization score could combine downstream influence (citations, guideline adoption, use as the basis for subsequent trials) with fragility markers extractable from the published record: limited effective information relative to analytic complexity; post-hoc subgroup findings; adaptive analytic choices; estimates where the magnitude just passes a significance threshold; single-site designs; and material departures from the design assumptions or adequacy criteria established at the outset. Findings that are both highly influential and structurally fragile are where replication dollars buy the most information.

The proposed indicators make these markers much easier to identify prospectively. Section 1.1 records whether inference was conditioned on data-driven selection; §1.3 records what the design was expected to support and whether realized information or study objectives materially changed. Issues like a lower-

than-anticipated effective sample size, unexpectedly high or low intraclass correlation, or a claim that moved from exploratory to confirmatory status are not themselves a scientific failure. But when such changes materially affect the strength of a consequential finding, they provide a rational signal for replication priority.

1.5 Learning from discordance: reconciliation and transportability

Discordance in the existing evidence is itself informative. Studies that appear to address the same question may differ in estimand, population, measurement, design, or analytic assumptions. Triangulation uses this principle: agreement across approaches with different sources of bias can strengthen causal inference, while disagreement can reveal which assumptions or design features matter (Lawlor et al., 2016). Rather than defaulting to another de novo study, NIH could support structured reconciliation of discordant evidence, harmonizing these dimensions where possible and identifying the likely sources of remaining disagreement.

The same idea is more powerful when planned prospectively. Where transportability matters, a second cohort, site, health system, or data source can apply the same estimand, outcome definition, and prespecified analysis. This narrows the dimensions that vary and provides direct evidence about whether a result transports across populations or settings. Methods for generalizability and transportability already formalize this problem as extending causal inferences from a study population to a defined target population (Dahabreh et al., 2021).

One element of a definition of causal research is explicit definitions of the theoretical and empirical estimands and assumptions under which the design and analysis identify that quantity. Randomization alone does not remove this requirement: PCORI-funded work on post-randomization selection in cluster trials shows how treatment-dependent identification, recruitment, or noncompliance can change the population represented by the observed data and the causal effect that can be identified (Li et al., 2024). Recent causal frameworks similarly emphasize the target estimand, target population, target trial, and target validity (Lu et al., 2024). Multiple data sources do not establish causality by themselves, but they can show whether a causal conclusion depends on the population, measurement, or contextual factors from which it was obtained.

Recommendations: Fund reconciliation of consequential discordant evidence as a scientific activity, with harmonized comparisons and identified sources of disagreement as deliverables rather than another narrative review. Where claims are intended to extend beyond a single source population, allow and incentivize budgets for harmonization and prespecified analysis in a second data source. The work required to make those sources comparable should itself be recognized as part of the scientific contribution.

1.6 A model that already works

NIH need not build a reproducibility review from scratch. Since 2016, the Journal of the American Statistical Association (JASA) has operated a reproducibility initiative establishing minimum criteria for code, data, and workflow, and created the Associate Editor of Reproducibility role to implement them; it has since expanded from Applications and Case Studies to all original research (Wrobel et al., 2024). Public materials include a reviewer and author guide, a pre-submission checklist, a standardized artifact declaration form, and a template repository.

Directly transferable elements:

A standardized artifact declaration, submitted with progress reports or at closeout, stating what is reproducible, by which script, from which data, and where scope is limited and why.

A designated, credited reproducibility reviewer role. The single most important design feature of the

JASA program is that this is a named position with real editorial standing, not volunteer labor. NIH could support analogous roles in biostatistics cores as an allowable and expected budget line.
A recognition mechanism. JASA gives an award for materials exceeding the minimum. NIH's Replication Prize already establishes the model; the same logic could extend to reproducibility artifacts, recognized in award reporting and citable in biosketches.
Master-script executability: relative paths only, a wrapper script that reproduces results without modification, transparent run order, seeds set where exact reproduction is expected, versioned software documented and cited. Low-burden if adopted at project start rather than retrofitted.

2. Data, Software, and Model Sharing: Extend sharing from data to analytic artifacts. A dataset without the analysis is insufficient for validation. Deposited artifacts should include the executable pipeline, the computational environment (container or lockfile), seeds, and the selection procedure itself, assuming the model choice was data-dependent, not merely the final chosen model. Reporting only the winning specification discards what is needed to evaluate the inference.

Adopt synthetic facsimile data as the standard answer to restricted access. Where real data cannot be shared, JASA (for instance) requires a synthetic or facsimile dataset in the same format so code can at least be executed and inspected. This is a tested solution to NIH's hardest sharing problem and could be made an explicit, fundable expectation rather than an ad hoc accommodation.

Measure executed reproductions, not availability statements. The proportion of awards with a data availability statement is helpful but less meaningful as a quality signal. The meaningful metric is the proportion of deposited artifacts a qualified third party can execute end-to-end in a clean environment. NIH could seed this by funding a small number of institutional or consortium reproduction services and reporting aggregate success rates.

Fund data resources, not just data access. Valuable clinical and administrative data often require substantial legal, technical, and scientific work before they are usable beyond the institution that generated them. Curation, linkage, harmonization, missing-data assessment, and measurement-quality evaluation are part of evidence production, not background plumbing. NIH should fund and recognize this work, including contributions from the institutions and personnel that create usable data resources, and expect data quality to be assessed relative to specified use cases rather than treated as an intrinsic property of a dataset. PCORI's Patient-Centered Outcomes Data Repository provides one model: methodological projects can deposit documented, reusable study data expressly to support secondary analysis, reproduction, and further methods research (Li, 2026).

Integrating modalities is a methodological problem requiring renewed investment and awareness to protect the scientific record. Making heterogeneous data available together does not ensure that existing tools will handle them appropriately. Differences in measurement, missingness, dimensionality, and relevance to the scientific question across modalities require methodological development alongside infrastructure investment. As argued in §1.2, funding data integration without funding the methods to analyze integrated data legitimately risks producing a great deal of accessible data and a great many non-replicable findings drawn from it.

Fund software maintenance. Methods software is infrastructure that decays without support.

Statisticians and data scientists maintain packages on CRAN that are used well beyond the projects that produced them, with no funding attached to that maintenance — which is the ordinary arrangement rather than an unusual one. NIH should permit and expect maintenance budgets, count software citations as scholarly output, and expect critical software to be cited in references alongside papers so its use is traceable.

3. Training and Mentorship: Training impact should be measured as development, not placement alone. Career outcomes remain useful, but they depend on trainees' starting goals, opportunities, and the changing research labor market. NIH could instead pair placement outcomes with brief baseline and follow-up measures of trainees' skills, confidence, research responsibilities, and career plans. Useful indicators include the number of trainees supported; gains in competencies such as study design, data analysis, communication, collaboration, and reproducible practice; and whether trainees gained responsibilities or opportunities they did not have at entry.

Access to appropriate mentorship is itself measurable. Trainees on non-methodological projects should have documented access to statistical and other relevant quantitative mentorship with protected effort, and training plans should treat reproducible workflows, including version control, environment management, and executable analysis as basic research competencies rather than specialty skills.

Suggested indicators: baseline and follow-up assessments of trainee competencies and career plans; trainee placement interpreted relative to stated goals at entry; access to appropriate methodological mentorship; and inclusion of reproducible-research competencies in training plans.

4. Collaboration: See attached for formatted version of the following:

Recognize methodological and reproducibility service as a scored contribution. Reproducibility review, cross-lab statistical mentoring, maintenance of shared analytic infrastructure, and other methodological service are largely invisible in conventional productivity metrics. Making these contributions visible in biosketches and review criteria is a low-cost way to recognize work that directly supports team science and research rigor. Independent methodological service, including statistical service on DSMBs and other monitoring bodies, should likewise be recognized as a substantive contribution even when it does not result in co-authorship.
Collaborative statistical contribution is otherwise typically recorded as either co-authorship or an acknowledgment. Better indicators would capture both the role and structure of the collaboration, including the following:
Named quantitative collaborators with meaningful, specified effort, ideally distinguishable from nominal statistical support at token (e.g., supervisory or consulting) levels.
Contribution taxonomy (e.g., CRediT) at publication, naming methodology, software, formal analysis, and data curation roles.
Collaborative products beyond publications, including jointly submitted grants, workshops or meetings developed, shared data or software resources, and trainees taught or mentored across disciplines.
Whether the analysis plan was authored or co-authored by, and final analytic decisions authorized by, a named quantitative scientist with substantive responsibility for the analysis.
Collaboration should reflect more than the number of people a scientist has worked with, disciplines spanned, or projects contributed to. A quantitative scientist who develops a deep, sustained

collaboration in one scientific area may contribute as much as one working across many fields. At the program or institutional level, NIH could instead characterize the reach and durability of collaborative networks, including repeated collaboration across projects and connections among otherwise separate scientific groups.

Analytic independence should be an explicit expectation. Individually reasonable analysis decisions, such as which covariates to adjust for, how to restrict a population, which subgroups to consider, how to handle outliers, etc., can cumulatively add up to an undocumented selection procedure, especially if there is pressure on such decisions in the direction of increased significance. Gelman and Loken (2014) call this the garden of forking paths: multiplicity exists even when only one analysis was ever run, because the path taken depended on the data. A quantitative collaborator supported only at token effort may lack the standing to resist result-driven analytic choices or to advocate independently for the ethical and judicious analysis.

Such independence should be supported by expecting named quantitative experts to participate meaningfully on research teams with substantive, not token, effort. Further, such independence is potentially measurable: whether the quantitative scientist holds authorship rather than acknowledgment, giving them standing to insist on a caveat or decline; whether their effort is protected on the award rather than discretionary; and, for high-stakes observational analyses, whether independent statistical review occurred, on the model of the independence already expected of trial statisticians.

5. Entrepreneurship and Translation: Translation measured through patents, licenses, and company formation systematically undercounts methodological translation. A method that moves from paper to open-source implementation to routine use in NIH-funded studies, clinical guidelines, regulatory submissions, or policy analysis is translated in the sense NIH cares about, and the process is measurable: downstream software citation, dependency counts in other NIH-supported analyses, adoption in guidelines and regulatory review.

NIH should include methods-to-practice pathways in translation indicators and support the unglamorous middle step between invention and adoption: testing, documentation, maintenance, training, and integration of methods into usable software and workflows. These activities currently fall through the cracks between methodological and translational funding.

6. Foundational Scientific Exploration: Valid inference after data-dependent selection is, at best, only partially solved. General solutions face fundamental limits; partial solutions exist but remain underused, in part because accounting honestly for selection often produces wider uncertainty, and wider uncertainty can be harder to publish.

Adjacent problems are similarly open and consequential: quantifying uncertainty for models produced by regularized selection; setting evidentiary thresholds that scale with the size and structure of a candidate space; establishing theoretical guarantees when covariate groups grow at different rates; and building transparent methods that remain competitive for outcome distributions where they currently underperform, such as zero-inflated and heavily skewed clinical endpoints.

This is precisely the foundational work the Initiative should fund: methodological, high-risk, without necessarily having an immediate clinical output, and directly relevant to whether findings across the rest

of the portfolio replicate. As biomedical data and analytic choice expand, investment in replication alone addresses failures downstream; foundational statistical research can prevent some of them upstream.

7. Public Impact: Public impact should be evaluated along an impact pathway rather than inferred from citation counts alone. Evaluation frameworks distinguish intermediate outcomes such as reach, adoption, and sustained use from downstream changes in population health (Glasgow et al., 1999). NIH could similarly assess whether research produces trustworthy findings that reach intended populations, enter practice or policy, and are sustained over time.

Population outcomes are harder to attribute because the population is ultimately uncontrolled. Before asking whether research changed health outcomes, NIH often needs to know whether the resulting intervention, diagnostic, screening strategy, or clinical practice actually reached the population. NIH should incentivize studies linking changes in disease profiles and outcomes to changes in management and uptake, supported by data on where and among whom products and practices are used. This may require linking surveillance and health-system data with commercial, claims, registry, or other sources that capture diffusion beyond traditional research datasets.

Trustworthy evidence is a prerequisite for meaningful public impact. The replication rate of influential NIH-supported findings, tracked over time, would therefore provide an expensive but unusually direct indicator of whether the knowledge being propagated into research, practice, and policy remains supported when tested again.
Two additional indicators worth developing are the estimated resources committed downstream to findings that later failed to replicate, which NIH is uniquely positioned to assess; and the inspectability of clinical decision tools derived from NIH-funded research, including whether clinicians and patients can understand and contest the basis of the model’s outputs.

Finally, not all public impact should be reduced to a count. Periodic expert evaluation of influential NIH-supported bodies of work could assess whether and how they changed scientific understanding, practice, policy, or population health over time.

8. Other Areas: See attached for formatted version of the attached

8. Other Area: Generative AI, Agents, & Accountability
Generative AI and modern agentic analytic tools interact with every other category, and current trajectories run against the goals of this initiative.
AI-assisted analysis can quickly become an unrecorded selection procedure. An agent that iterates through dozens or hundreds of specifications and surfaces the one that "works best" is performing model selection at a scale exceeding what any analysis plan can reasonably specify in advance. The path taken by an LLM agent is rarely, if ever, reproducible from the prompt alone. If human-led data analysis represents Gelman's "garden of forking paths," an AI agent turns that garden into a random forest. Mandating a complete audit trail of this search process is technically possible. However, with modern "chain of thought" LLMs, the massive volume of intermediate and erratic "thinking" steps renders any such record highly convoluted. An AI-assisted scientist must then decide among several alternatives. One option is to diligently store a large, complex record of these AI-dictated choices, and to publish this record alongside their research. Another option is to narratively summarize the AI-enabled search, which

asks readers and funders alike to trust that whatever decisions were made, justified or not, were appropriate. Yet another option is to proceed as though the selected path was the "right" one all along. Every incentive pushes the scientist toward this final path of least resistance, which also carries the worst consequences for replicability. Consequently, as AI-assisted analysis becomes normalized, rejected specifications and the rationale for moving among them will disappear — not because scientists intend to conceal them, but because we lack both the norms of analytic accountability and the practical, rigorous tools to support them. This is multiplicity at machine speed, and it is poised to undermine replicability if left unchecked.

Recommendation: prevent multiplicity at the used for analytic work should make this information exportable rather than requiring investigators to reconstruct it after the fact. NIH could support development of provenance-capture tooling as research infrastructure.

Recommendation: retain locatable human accountability. A named individual should be able to explain and defend every analytic choice. NIH could require an attestation, parallel to existing authorship and conflict attestations, that a responsible analyst reviewed and understood AI-generated analytic code and decisions. AI may propose analytic choices, but a named human must adopt them and be able to state why each was selected over considered alternatives. This shifts analytic authority from AI dictation, which cannot be held to account for its mistakes, back to human judgment.

Recommendation: promote team science through shared analytic artifacts. Collaboration requires a shared, inspectable artifact to collaborate on. If each member works in a private AI-assisted workflow producing unshared code and unlogged decisions, team science becomes several parallel solo efforts with a joint author list (or worse, reverts to solo work with an AI). Shared repositories, executable master scripts, versioned environments, and declared selection procedures are the practical antidote, and NIH should fund and expect them at the team level.

9. Other Area: Unintended Consequences and Feasibility

Checklist theater. Any requirement can become a compliance artifact if checked for existence rather than content; data-availability statements are a cautionary example of the limitations of existence-based metrics. Mitigation: sample-based audits by qualified reviewers, and metrics based on execution rather than deposit.

Chilling exploration. Pre-specification requirements can be misread as prohibiting exploration. The requirement is labeling, not prohibition. Selection disclosure should let exploratory work proceed freely and be reported honestly as exploratory rather than presented as confirmatory.

Over-rewarding confirmation. If replication success becomes a scored metric for individual investigators, the rational response is to study safe questions. This ultimately risks stifled innovation and hinders research exploration. Replication incentives should attach to the enterprise and to replication teams, not to the original investigator's record.

Regressive burden. Container registries, reproducibility reviewers, and synthetic data generation are easier to absorb at well-resourced institutions than at small ones. Mitigation: fund shared infrastructure, allow these costs explicitly in budgets, and scale expectations to award size and complexity rather than applying a uniform standard.

Career-stage asymmetry. Compliance and service expectations impose different burdens across career stages, with early-career investigators often least able to absorb them. Reproducibility service and exemplary artifacts should count concretely in review.

Metric capture generally. The indicators emphasized here, including executed reproductions, audited selection disclosure, candidate-space reporting, and downstream enablement, are relatively resistant to

gaming because satisfying them requires producing much of the underlying scientific behavior they are intended to measure.
Feasibility across contexts: selection disclosure, candidate-space reporting, and artifact declaration are relatively low burden when incorporated prospectively into the workflow, because they record decisions as they occur rather than reconstructing them later. Executable-artifact standards carry moderate costs and scale with award size. Independent reproduction services are more resource-intensive and should be centrally or consortium-funded rather than mandated institution by institution.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Response-to-NIH-Request-for-Information_-Measuring-and-Rewarding-Scientific-Impact.pdf

Description: The attached file has the American Statistical Association's entire set of comments, which are also pasted into the text boxes above but without the formatting (e.g., section titles, bullets).

ASA Response to NIH Request for Information: Measuring and Rewarding Scientific Impact

August 19, 2026

Developed with scientific and technical input from Ryan A. Peterson, PhD, University of Iowa; Elizabeth W. Eisenhauer, PhD; Theresa Kim, PhD, MS; and other ASA members.

Claude Opus 5 (Anthropic) and GPT-5.6 Sol (OpenAI) were used to assist with drafting and editing these comments.

Contributor disclaimer: The views do not necessarily represent the views of contributors' employers or affiliated institutions. Theresa Kim provided scientific and technical input during off-duty hours and without use of government-furnished equipment; her participation does not represent the views of NIH, HHS, or the U.S. Government.

---

1. Rigor and Reproducibility

1.1 The missing instrument: indicators for analytic replicability

We are highly encouraged by NIH's renewed focus on transparency, reproducibility, and replication shown in the Request for Information on Measuring and Rewarding Scientific Impact, the Initiative to promote Strengthening Replication and Reproducibility of NIH-funded Research, the Gold Standard implementation plan, and the recent Highlighted Topic on Enhancing Scientific Rigor, Transparency and Replicability. These documents name reproducibility and replicability as distinct goals, correctly, and the

Initiative has instruments aimed at each: for the former, longstanding investments in data sharing, standardization, and public access to findings; for the latter, funded replication research through the Common Fund, a Replication Prize, and a stated intent to engage researchers in identifying areas ripe for replication.

These are steps in the right direction. However, we have yet to observe incentives targeting the gap between reproducibility and replicability: whether the reported inference is valid given how the analysis proceeded. Statisticians and biostatisticians have a useful and unusual vantage point on this gap: we are often among the first to recognize why a finding may not replicate, while not always having the standing or involvement needed to address the problem.

This matters because replication studies *detect* errors, but do not *prevent* them. They measure the failure rate after the fact, one finding at a time, at a high cost per finding. A study can be perfectly reproducible — fully containerized, one-click re-runnable — and still report inference that is unlikely to replicate, and the replication study will discover this several years and several hundred thousand dollars later.

A common route to this is inference on a model that was chosen using the same data used for inference: **unadjusted post-selection inference (UPSI)**. It is one instance of what Simmons, Nelson, and Simonsohn (2011) termed researcher degrees of freedom: undisclosed flexibility in analytic choices, sufficient on its own to produce significant findings without any intent to deceive. In a satirical blog post, Perry Hackman used simulations with pure-noise outcomes and a modest set of candidate subgroup interactions to show that an ostensibly rigorous method, forward stepwise selection with information criteria followed by standard statistical inference (dubbed UPSI), can produce at least one "significant" effect in 90+% of simulated trials (Peterson, 2025b). Nothing in that pipeline is clearly fraudulent; the analysis plan reads as rigorous, and the resulting code can be shared, containerized, and re-run perfectly.

Recommendation: The Initiative's replication instruments would return more per dollar if paired with a cheap upstream instrument that flags the findings most likely to need them. In analysis plans, progress reports, and publication reporting, a short structured statement answering the following could be incentivized:

- Was the reported model, subgroup, outcome scale, or covariate set chosen using the same data used to compute the reported inference?
- If so, how were the inferences adjusted for selection? (sample splitting; selective or conditional inference; bootstrap or model-averaged uncertainty; explicit relabeling as exploratory). If not, where was the full analysis specification verifiably pre-registered?
- How many candidate specifications were considered, including those examined and discarded?

Based on the above, results should be taken as on a spectrum between confirmatory (to provide stronger evidence for prespecified claims) and exploratory (to identify patterns and generate hypotheses).

Another recommendation is to commission periodic targeted statistical review of a random sample of supported publications, scored by trained statistical reviewers, reporting the proportion that (a) report selection-conditional inference without adjustment, (b) report subgroup findings without pre-specified subgroup structure, and (c) contain code that does not execute. Tracking that proportion over time measures whether the Initiative is working.

1.2 Candidate pools and model complexity are growing faster than the standards applied

Beneath many variable-selection procedures is an implicit assumption of covariate equipoise: that every candidate predictor is equally worthy of entering the model. Stepwise procedures, popular information criteria searches, and standard lasso implementations can all behave this way, and the assumption becomes problematic in settings where NIH is investing heavily.

Independent groups have converged on the same principle from different directions. Yu, Bien, and Tibshirani (2019) articulate a *reluctant interaction selection principle* — prefer a main effect over an interaction if all else is equal — and Tay and Tibshirani (2020) extend this method to a *reluctant non-linear selection principle*, preferring a linear term over a non-linear one on the same grounds. In his standard reference text, Harrell (2015) advises pre-specifying a small set of plausible interactions rather than testing all of them. Peterson and Cavanaugh (2022) advise that human-understandable "glass-box" models with meaningful features should be preferred to more complex ones precisely because they are easier to interrogate and replicate. Different penalties, different algorithms, same underlying claim: complexity should have to earn its way into a model, and treating every candidate term as equally deserving is what allows spurious complexity through. NIH need not adjudicate among these methods, but it should seek to require the reporting that makes the problem visible.

Transparent modeling methods matter because **replication requires something to replicate**: a claim specific enough that an independent team can test it on new data and report whether it held. Transparent statistical models often make such claims directly available. *Waist circumference is associated with one-year HDL, $\beta = -0.04$ (95% CI −0.07 to −0.01)* is a statement another group can go and check in a different dataset, and the check has a clear objective.

On the other hand, black-box predictive models, such as neural networks, XGBoost, random forests, and statistical models with high-order interaction terms, instead primarily yield prediction functions. A second team can apply the deposited model to new data and test whether predictive performance holds up. That is a meaningful form of external validation, but it is different from replicating an interpretable scientific claim. When performance drops, the failure may be difficult to localize: a transparent model can reveal which estimated relationship moved and in which direction, while an opaque model may reveal only that predictive performance deteriorated, and, if subgroup metrics are tracked, for whom.

This has a direct implication for the Initiative. Predictive modeling is an increasingly prominent category of NIH-supported research (e.g., Arshi et al. 2025, Le et al. 2024), and it is precisely the category on

which replication instruments have the least purchase. Transparency is therefore not only a matter of trust, fairness, or regulatory compliance, though it is those too. It also determines how specifically a scientific claim can be interrogated and replicated.

As datasets grow larger and contain newer, more high-dimensional modalities, new statistical methodology will be critical. For instance, with only twenty candidate predictors, to capture potential interactions, one would need to sift through 190 pairwise interactions, 1,140 three-way interactions, and over a million candidate models across all orders. Further, when modalities differ in their dimension and overall signal, they quickly run into practical challenges. For example, in a lung adenocarcinoma survival analysis with 7 clinical covariates and 22,283 gene expression measurements, assuming covariate equipoise and penalizing both groups equally encodes a structural disadvantage for clinically well-motivated variables, which get de-selected due to the sheer difference in dimensionality of the two modalities. With NIH's data-integration priorities, as studies become multimodal, consisting of genomics plus clinical plus imaging plus wearable plus environmental and more, the need for novel methods that can accommodate such data is critical; important methodological challenges remain at the scale and complexity of these data.

Recommendations: First, NIH should be wary of accepting post-hoc explanation of an opaque model (such as a locally interpretable model, a SHAP-based variable importance metric, etc.) as a reliable, replicable scientific finding; the explainer is a separate model that may or may not describe what the first one does, and should not be treated as equivalent to a directly interpretable scientific claim. Second, the accuracy cost of transparency is an empirical question that has been studied. A systematic review of 71 studies developing clinical prediction models found no performance benefit of machine learning over logistic regression (Christodoulou et al., 2019). Benchmarking work supports this outside the clinical setting as well: across 110 datasets from the Penn Machine Learning Benchmarks database, a transparent regression approach was within 5% of the best-performing method on 79% of classification and 62% of regression datasets, with random forests and gradient boosting showing no statistically distinguishable advantage (Peterson et al., 2024). Where transparent methods fell short, the reasons were diagnosable and specific and not reflective of a general law that complexity buys accuracy (specific examples can be found in Peterson et al., 2024, Section 3.3.2).

The implication is not that black-box methods should be discouraged, but that the burden of justification runs backward. **Investigators are rarely asked to show that their model's opacity bought anything**. For NIH-funded models intended to inform clinical decisions, a low-cost expectation is that a transparent baseline be fit and reported alongside the complex one, so the cost of interpretability is measured in that application rather than assumed. Where the gap is negligible, as is often the case, the transparent model is the better scientific product, because it can be interrogated and corrected.

1.3 Adequacy at the design stage: making evidentiary expectations auditable

These analysis-stage safeguards are necessary, but inference is only as strong as the data generated by the study design. The design determines which populations, outcomes, contrasts, and sources of variation are observed and therefore which scientific claims the resulting data can support. One of the earliest places this becomes explicit is the sample-size or design justification, where investigators state what they expect the proposed design to be capable of learning.

This makes the justification useful for more than determining a number. It provides a prospective, auditable statement of what adequate evidence was expected to look like. NIH can then ask whether the realized study provided that evidence and, if not, whether the departures and their implications were documented. A study need not unfold exactly as anticipated. Recruitment, measurement properties, effect sizes, intraclass correlations, or other design features must often be learned during the course of a study, at which point they become scientific information in their own right. Reporting them allows planning assumptions to be updated across studies rather than repeatedly regenerated from incomplete information.

For studies relying on existing or linked data, realized data quality may also differ from what was assumed at design, with implications for which claims the data can support. Rigor requires making those departures visible and reconsidering the claims they permit, rather than retrospectively treating the original justification as though nothing changed.

Gold Standard Science creates a policy demand for determining and justifying adequate sample sizes, but no common definition of “adequate” accompanies it. Raising conventional thresholds does not settle the problem. A study can have 90% power under poorly supported assumptions, for an effect of little scientific consequence, using an inefficient design, or with an imprecisely measured outcome. Adequacy is instead a fit-for-purpose judgment: whether the design, measurement, data quality, and analysis are sufficient for the claims the study is intended to support.

NIH guidance already reflects this principle. Under the current Parent R01 review framework, reviewers assess whether sample size is sufficient and well justified in relation to the research questions, outcomes, and design (NIH, 2024). NCCIH feasibility trials instead require quantitative feasibility or acceptability benchmarks and justification that the sample is sufficient to evaluate them; testing efficacy or “preliminary efficacy” is explicitly not the purpose (NCCIH, 2024a). By contrast, NCCIH efficacy, effectiveness, and pragmatic trials require power calculations linked to endpoints and hypotheses, at least 90% power for the primary outcome, adjustment for multiple primary outcomes, and a clinically meaningful effect size (NCCIH, 2024b). NIH therefore already treats adequacy as purpose-dependent; the opportunity is to make that principle explicit, consistent, and auditable.

The same principle extends beyond these examples. Exploratory secondary analyses, transportability or triangulation studies, resource construction, and methods development may require adequate precision, population overlap, analytic scope, ability to assess discordance, or support for future studies rather

than power against a single alternative. Requiring a power calculation regardless can produce the calculation without the justification. This is not a lower bar for exploratory work but a different one: identify the intended claim, the criterion of adequacy it requires, and the corresponding limits on inference.

Design information also interacts with the candidate analysis space. Limited information combined with broad analytic flexibility can generate apparently strong but unstable findings, while each source of risk is largely invisible to instruments aimed only at the other.

Declaring intended claims and criteria of adequacy at the design stage gives the disclosure in §1.1 something to be checked against. A confirmatory aim that later relies on substantial data-dependent selection, or a secondary outcome later presented as a primary claim, has changed evidentiary status during execution. That is not necessarily a scientific failure, provided that its path from "design" to "discussion" is visible and reflected in the claims made.

Recommendations:

First, require fit-for-purpose definitions of adequacy. Applications should identify the claims or objectives the study is intended to support, and the criterion for an adequate sample size or design should reflect that purpose. When an effect size is used to justify adequacy, its magnitude should be grounded where possible in empirical evidence and interpreted on a scale meaningful to the scientific question. Generic standardized effect-size conventions such as “small,” “medium,” and “large” should not substitute for separately considering the magnitude of effect that would matter and the variability anticipated in the study population. Not every aim of a complex study requires a separate, fully developed power analysis. However, the study should identify the minimum set of claims or objectives it is designed to support with strong evidentiary justification and demonstrate that the proposed design is adequate for those purposes.

Second, treat sample-size justification as an integral part of study design rather than a calculation performed after the design is largely fixed. The amount of information a study can provide depends not only on the number of observations but on interdependent choices about the population recruited, comparisons made, measurements collected, timing and intensity of follow-up, analytic strategy, and other features that can affect the credibility of conclusions. Those choices must also be made within constraints on participant burden, operational feasibility, time, and budget. Improving one dimension may worsen another. Sample-size justification should therefore be an iterative, collaborative process in which statistical, substantive, operational, and other relevant experts identify the major threats to the study’s success, examine how alternative design choices redistribute those threats, and determine how limited resources can be used to produce the strongest evidence for the study’s purpose. This provides a concrete, early example of collaboration that NIH could recognize and measure: integrating statistical, scientific, operational, and other relevant expertise into consequential study-design decisions.

Third, make sample-size and design justification transparent, auditable, and cumulative. Power analyses and other sample-size calculations should be reproducible from the statistical analysis plan, with the

method, inputs, assumptions, and software or tools documented sufficiently for an independent reviewer to recreate and vary the calculation. Studies should subsequently compare planning assumptions with what was realized, including recruitment and retention, variability, clustering, measurement performance, and other design quantities. Meaningful departures should be reported along with their implications for the evidence the study can support, rather than treated only as implementation failures. Recording what was expected, what occurred, and how those differences affected the study would allow empirical knowledge about design assumptions to accumulate across studies, so that future investigators can ground their assumptions in better evidence. NIH could track the proportion of funded studies that make their design justifications reproducible and subsequently report comparisons between key planning assumptions and realized study conditions.

1.4 Identifying findings ripe for replication

The Initiative states that NIH will engage researchers in selecting research areas ripe for replication and reproducibility studies. The indicators proposed above provide a practical way to do so.

Statistical fragility can triage more effectively than citation count alone. A tractable prioritization score could combine downstream influence (citations, guideline adoption, use as the basis for subsequent trials) with fragility markers extractable from the published record: limited effective information relative to analytic complexity; post-hoc subgroup findings; adaptive analytic choices; estimates where the magnitude just passes a significance threshold; single-site designs; and material departures from the design assumptions or adequacy criteria established at the outset. Findings that are both highly influential and structurally fragile are where replication dollars buy the most information.

The proposed indicators make these markers much easier to identify prospectively. Section 1.1 records whether inference was conditioned on data-driven selection; §1.3 records what the design was expected to support and whether realized information or study objectives materially changed. Issues like a lower-than-anticipated effective sample size, unexpectedly high or low intraclass correlation, or a claim that moved from exploratory to confirmatory status are not themselves a scientific failure. But when such changes materially affect the strength of a consequential finding, they provide a rational signal for replication priority.

1.5 Learning from discordance: reconciliation and transportability

Discordance in the existing evidence is itself informative. Studies that appear to address the same question may differ in estimand, population, measurement, design, or analytic assumptions. Triangulation uses this principle: agreement across approaches with different sources of bias can strengthen causal inference, while disagreement can reveal which assumptions or design features matter (Lawlor et al., 2016). Rather than defaulting to another *de novo* study, NIH could support structured reconciliation of discordant evidence, harmonizing these dimensions where possible and identifying the likely sources of remaining disagreement.

The same idea is more powerful when planned prospectively. Where transportability matters, a second cohort, site, health system, or data source can apply the same estimand, outcome definition, and prespecified analysis. This narrows the dimensions that vary and provides direct evidence about whether a result transports across populations or settings. Methods for generalizability and transportability already formalize this problem as extending causal inferences from a study population to a defined target population (Dahabreh et al., 2021).

One element of a definition of causal research is explicit definitions of the theoretical and empirical estimands and assumptions under which the design and analysis identify that quantity. Randomization alone does not remove this requirement: PCORI-funded work on post-randomization selection in cluster trials shows how treatment-dependent identification, recruitment, or noncompliance can change the population represented by the observed data and the causal effect that can be identified (Li et al., 2024).

Recent causal frameworks similarly emphasize the target estimand, target population, target trial, and target validity (Lu et al., 2024). Multiple data sources do not establish causality by themselves, but they can show whether a causal conclusion depends on the population, measurement, or contextual factors from which it was obtained.

Recommendations: Fund reconciliation of consequential discordant evidence as a scientific activity, with harmonized comparisons and identified sources of disagreement as deliverables rather than another narrative review. Where claims are intended to extend beyond a single source population, allow and incentivize budgets for harmonization and prespecified analysis in a second data source. The work required to make those sources comparable should itself be recognized as part of the scientific contribution.

1.6 A model that already works

NIH need not build a reproducibility review from scratch. Since 2016, the Journal of the American Statistical Association (JASA) has operated a reproducibility initiative establishing minimum criteria for code, data, and workflow, and created the Associate Editor of Reproducibility role to implement them; it has since expanded from Applications and Case Studies to all original research (Wrobel et al., 2024). Public materials include a reviewer and author guide, a pre-submission checklist, a standardized artifact declaration form, and a template repository.

Directly transferable elements:

- **A standardized artifact declaration**, submitted with progress reports or at closeout, stating what is reproducible, by which script, from which data, and where scope is limited and why.
- **A designated, credited reproducibility reviewer role.** The single most important design feature of the JASA program is that this is a named position with real editorial standing, not volunteer labor. NIH could support analogous roles in biostatistics cores as an allowable and expected budget line.

- **A recognition mechanism.** JASA gives an award for materials exceeding the minimum. NIH's Replication Prize already establishes the model; the same logic could extend to reproducibility artifacts, recognized in award reporting and citable in biosketches.
- **Master-script executability**: relative paths only, a wrapper script that reproduces results without modification, transparent run order, seeds set where exact reproduction is expected, versioned software documented and cited. Low-burden if adopted at project start rather than retrofitted.

---

2. Data, Software, and Model Sharing

Extend sharing from data to analytic artifacts. A dataset without the analysis is insufficient for validation. Deposited artifacts should include the executable pipeline, the computational environment (container or lockfile), seeds, and the *selection procedure itself,* assuming the model choice was data-dependent, not merely the final chosen model. Reporting only the winning specification discards what is needed to evaluate the inference.

Adopt synthetic facsimile data as the standard answer to restricted access. Where real data cannot be shared, JASA (for instance) requires a synthetic or facsimile dataset in the same format so code can at least be executed and inspected. This is a tested solution to NIH's hardest sharing problem and could be made an explicit, fundable expectation rather than an *ad hoc* accommodation.

Measure executed reproductions, not availability statements. The proportion of awards with a data availability statement is helpful but less meaningful as a quality signal. The meaningful metric is the proportion of deposited artifacts a qualified third party can execute end-to-end in a clean environment. NIH could seed this by funding a small number of institutional or consortium reproduction services and reporting aggregate success rates.

Fund data resources, not just data access. Valuable clinical and administrative data often require substantial legal, technical, and scientific work before they are usable beyond the institution that generated them. Curation, linkage, harmonization, missing-data assessment, and measurement-quality evaluation are part of evidence production, not background plumbing. NIH should fund and recognize this work, including contributions from the institutions and personnel that create usable data resources, and expect data quality to be assessed relative to specified use cases rather than treated as an intrinsic property of a dataset. PCORI's Patient-Centered Outcomes Data Repository provides one model: methodological projects can deposit documented, reusable study data expressly to support secondary analysis, reproduction, and further methods research (Li, 2026).

Integrating modalities is a methodological problem requiring renewed investment and awareness to protect the scientific record. Making heterogeneous data available together does not ensure that existing tools will handle them appropriately. Differences in measurement, missingness, dimensionality, and relevance to the scientific question across modalities require methodological development alongside

infrastructure investment. As argued in §1.2, funding data integration without funding the methods to analyze integrated data legitimately risks producing a great deal of accessible data and a great many non-replicable findings drawn from it.

Fund software maintenance. Methods software is infrastructure that decays without support. Statisticians and data scientists maintain packages on CRAN that are used well beyond the projects that produced them, with no funding attached to that maintenance — which is the ordinary arrangement rather than an unusual one. NIH should permit and expect maintenance budgets, count software citations as scholarly output, and expect critical software to be cited in references alongside papers so its use is traceable.

---

3. Training and Mentorship

Training impact should be measured as development, not placement alone. Career outcomes remain useful, but they depend on trainees' starting goals, opportunities, and the changing research labor market. NIH could instead pair placement outcomes with brief baseline and follow-up measures of trainees' skills, confidence, research responsibilities, and career plans. Useful indicators include the number of trainees supported; gains in competencies such as study design, data analysis, communication, collaboration, and reproducible practice; and whether trainees gained responsibilities or opportunities they did not have at entry.

Access to appropriate mentorship is itself measurable. Trainees on non-methodological projects should have documented access to statistical and other relevant quantitative mentorship with protected effort, and training plans should treat reproducible workflows, including version control, environment management, and executable analysis as basic research competencies rather than specialty skills.

Suggested indicators: baseline and follow-up assessments of trainee competencies and career plans; trainee placement interpreted relative to stated goals at entry; access to appropriate methodological mentorship; and inclusion of reproducible-research competencies in training plans.

---

4. Collaboration

Recognize methodological and reproducibility service as a scored contribution. Reproducibility review, cross-lab statistical mentoring, maintenance of shared analytic infrastructure, and other methodological service are largely invisible in conventional productivity metrics. Making these contributions visible in biosketches and review criteria is a low-cost way to recognize work that directly supports team science and research rigor. Independent methodological service, including statistical service on DSMBs and other

monitoring bodies, should likewise be recognized as a substantive contribution even when it does not result in co-authorship.

Collaborative statistical contribution is otherwise typically recorded as either co-authorship or an acknowledgment. Better indicators would capture both the role and structure of the collaboration, including the following:

- Named quantitative collaborators with **meaningful, specified effort**, ideally distinguishable from nominal statistical support at token (e.g., supervisory or consulting) levels.
- Contribution taxonomy (e.g., CRediT) at publication, naming methodology, software, formal analysis, and data curation roles.
- Collaborative products beyond publications, including jointly submitted grants, workshops or meetings developed, shared data or software resources, and trainees taught or mentored across disciplines.
- Whether the analysis plan was authored or co-authored by, and final analytic decisions authorized by, a named quantitative scientist with substantive responsibility for the analysis.

Collaboration should reflect more than the number of people a scientist has worked with, disciplines spanned, or projects contributed to. A quantitative scientist who develops a deep, sustained collaboration in one scientific area may contribute as much as one working across many fields. At the program or institutional level, NIH could instead characterize the reach and durability of collaborative networks, including repeated collaboration across projects and connections among otherwise separate scientific groups.

Analytic independence should be an explicit expectation. Individually reasonable analysis decisions, such as which covariates to adjust for, how to restrict a population, which subgroups to consider, how to handle outliers, etc., can cumulatively add up to an undocumented selection procedure, especially if there is pressure on such decisions in the direction of increased significance. Gelman and Loken (2014) call this the garden of forking paths: multiplicity exists even when only one analysis was ever run, because the path taken depended on the data. A quantitative collaborator supported only at token effort may lack the standing to resist result-driven analytic choices or to advocate independently for the ethical and judicious analysis.

Such independence should be supported by expecting named quantitative experts to participate meaningfully on research teams with substantive, not token, effort. Further, such independence is potentially measurable: whether the quantitative scientist holds authorship rather than acknowledgment, giving them standing to insist on a caveat or decline; whether their effort is protected on the award rather than discretionary; and, for high-stakes observational analyses, whether independent statistical review occurred, on the model of the independence already expected of trial statisticians.

5. Entrepreneurship and Translation

Translation measured through patents, licenses, and company formation systematically undercounts methodological translation. A method that moves from paper to open-source implementation to routine use in NIH-funded studies, clinical guidelines, regulatory submissions, or policy analysis is translated in the sense NIH cares about, and the process is measurable: downstream software citation, dependency counts in other NIH-supported analyses, adoption in guidelines and regulatory review.

NIH should include **methods-to-practice pathways** in translation indicators and support the unglamorous middle step between invention and adoption: testing, documentation, maintenance, training, and integration of methods into usable software and workflows. These activities currently fall through the cracks between methodological and translational funding.

---

6. Foundational Scientific Exploration

Valid inference after data-dependent selection is, at best, only partially solved. General solutions face fundamental limits; partial solutions exist but remain underused, in part because accounting honestly for selection often produces wider uncertainty, and wider uncertainty can be harder to publish.

Adjacent problems are similarly open and consequential: quantifying uncertainty for models produced by regularized selection; setting evidentiary thresholds that scale with the size and structure of a candidate space; establishing theoretical guarantees when covariate groups grow at different rates; and building transparent methods that remain competitive for outcome distributions where they currently underperform, such as zero-inflated and heavily skewed clinical endpoints.

This is precisely the foundational work the Initiative should fund: methodological, high-risk, without necessarily having an immediate clinical output, and directly relevant to whether findings across the rest of the portfolio replicate. As biomedical data and analytic choice expand, investment in replication alone addresses failures downstream; foundational statistical research can prevent some of them upstream.

---

7. Public Impact

Public impact should be evaluated along an impact pathway rather than inferred from citation counts alone. Evaluation frameworks distinguish intermediate outcomes such as reach, adoption, and sustained use from downstream changes in population health (Glasgow et al., 1999). NIH could similarly assess whether research produces trustworthy findings that reach intended populations, enter practice or policy, and are sustained over time.

Population outcomes are harder to attribute because the population is ultimately uncontrolled. Before asking whether research changed health outcomes, NIH often needs to know whether the resulting intervention, diagnostic, screening strategy, or clinical practice actually reached the population. NIH should incentivize studies linking changes in disease profiles and outcomes to changes in management and uptake, supported by data on where and among whom products and practices are used. This may require linking surveillance and health-system data with commercial, claims, registry, or other sources that capture diffusion beyond traditional research datasets.

Trustworthy evidence is a prerequisite for meaningful public impact. The replication rate of influential NIH-supported findings, tracked over time, would therefore provide an expensive but unusually direct indicator of whether the knowledge being propagated into research, practice, and policy remains supported when tested again.

Two additional indicators worth developing are the estimated resources committed downstream to findings that later failed to replicate, which NIH is uniquely positioned to assess; and the inspectability of clinical decision tools derived from NIH-funded research, including whether clinicians and patients can understand and contest the basis of the model's outputs.

Finally, not all public impact should be reduced to a count. Periodic expert evaluation of influential NIH-supported bodies of work could assess whether and how they changed scientific understanding, practice, policy, or population health over time.

---

8. Other Area: Generative AI, Agents, & Accountability

Generative AI and modern agentic analytic tools interact with every other category, and current trajectories run against the goals of this initiative.

AI-assisted analysis can quickly become an unrecorded selection procedure. An agent that iterates through dozens or hundreds of specifications and surfaces the one that "works best" is performing model selection at a scale exceeding what any analysis plan can reasonably specify in advance. The path taken by an LLM agent is rarely, if ever, reproducible from the prompt alone. If human-led data analysis represents Gelman's "garden of forking paths," an AI agent turns that garden into a random forest.

Mandating a complete audit trail of this search process is technically possible. However, with modern "chain of thought" LLMs, the massive volume of intermediate and erratic "thinking" steps renders any such record highly convoluted. An AI-assisted scientist must then decide among several alternatives. One option is to diligently store a large, complex record of these AI-dictated choices, and to publish this record alongside their research. Another option is to narratively summarize the AI-enabled search, which asks readers and funders alike to trust that whatever decisions were made, justified or not, were appropriate. Yet another option is to proceed as though the selected path was the "right" one all along.

Every incentive pushes the scientist toward this final path of least resistance, which also carries the worst consequences for replicability. Consequently, as AI-assisted analysis becomes normalized, rejected specifications and the rationale for moving among them will disappear — not because scientists intend to conceal them, but because we lack both the norms of analytic accountability and the practical, rigorous tools to support them. This is multiplicity at machine speed, and it is poised to undermine replicability if left unchecked.

Recommendation: prevent multiplicity at the design stage through methodological collaboration. Methodological input on study design should be a fundable, expected component of research, not an unbudgeted courtesy. It has always been true that a specification chosen *a priori* on a firm statistical foundation generates no search to document. Until recently, the effort required to explore many specifications served as an informal check on doing so. AI agents remove that friction, making design-stage input more consequential, yet simultaneously easier to skip. The recommendations that follow address the multiplicity that remains, as it always will, since no analysis plan can anticipate everything a real dataset will demand.

Recommended indicator: an analytic provenance record. Where generative AI materially contributes to analytic code or decisions, require a short log — model and version, date, nature of the contribution, and, where AI was used to generate or compare analytic alternatives, the alternatives considered and the selection process. Systems used for analytic work should make this information exportable rather than requiring investigators to reconstruct it after the fact. NIH could support development of provenance-capture tooling as research infrastructure.

Recommendation: retain locatable human accountability. A named individual should be able to explain and defend every analytic choice. NIH could require an attestation, parallel to existing authorship and conflict attestations, that a responsible analyst reviewed and understood AI-generated analytic code and decisions. AI may propose analytic choices, but a named human must adopt them and be able to state why each was selected over considered alternatives. This shifts analytic authority from AI dictation, which cannot be held to account for its mistakes, back to human judgment.

Recommendation: promote team science through shared analytic artifacts. Collaboration requires a shared, inspectable artifact to collaborate on. If each member works in a private AI-assisted workflow producing unshared code and unlogged decisions, team science becomes several parallel solo efforts with a joint author list (or worse, reverts to solo work with an AI). Shared repositories, executable master scripts, versioned environments, and declared selection procedures are the practical antidote, and NIH should fund and expect them at the team level.

9. Other Area: Unintended Consequences and Feasibility

- **Checklist theater.** Any requirement can become a compliance artifact if checked for existence rather than content; data-availability statements are a cautionary example of the limitations of

existence-based metrics. Mitigation: sample-based audits by qualified reviewers, and metrics based on execution rather than deposit.

- **Chilling exploration.** Pre-specification requirements can be misread as prohibiting exploration. The requirement is *labeling*, not prohibition. Selection disclosure should let exploratory work proceed freely and be reported honestly as exploratory rather than presented as confirmatory.
- **Over-rewarding confirmation.** If replication success becomes a scored metric for individual investigators, the rational response is to study safe questions. This ultimately risks stifled innovation and hinders research exploration. Replication incentives should attach to the enterprise and to replication teams, not to the original investigator's record.
- **Regressive burden.** Container registries, reproducibility reviewers, and synthetic data generation are easier to absorb at well-resourced institutions than at small ones. Mitigation: fund shared infrastructure, allow these costs explicitly in budgets, and scale expectations to award size and complexity rather than applying a uniform standard.
- **Career-stage asymmetry.** Compliance and service expectations impose different burdens across career stages, with early-career investigators often least able to absorb them. Reproducibility service and exemplary artifacts should count concretely in review.
- **Metric capture generally.** The indicators emphasized here, including executed reproductions, audited selection disclosure, candidate-space reporting, and downstream enablement, are relatively resistant to gaming because satisfying them requires producing much of the underlying scientific behavior they are intended to measure.

Feasibility across contexts: selection disclosure, candidate-space reporting, and artifact declaration are relatively low burden when incorporated prospectively into the workflow, because they record decisions as they occur rather than reconstructing them later. Executable-artifact standards carry moderate costs and scale with award size. Independent reproduction services are more resource-intensive and should be centrally or consortium-funded rather than mandated institution by institution.

Closing

Statisticians and biostatisticians have an unusual vantage point on this RFI: we are often among the first to recognize why a finding may not replicate, while not always having the standing or involvement needed to address the problem. The interventions most likely to succeed make analytic choices visible and attributable, because visibility is what allows a collaborator, reviewer, or reader to catch the problem, and because the alternative in an era of AI-accelerated analysis is a scientific record that is easy to re-run and impossible to interrogate.

The American Statistical Association appreciates the opportunity to comment on the Request for Information on Measuring and Rewarding Scientific Impact. Questions or comments may be directed to ASA Director of Science Policy Steve Pierson: spierson@amstat.org.

References cited

- Arshi, B., Wynants, L., Rijnhart, E., Reeve, K., Cowley, L. E., & Smits, L. J. (2025). Number of Publications on New Clinical Prediction Models: A Bibliometric Review. *JMIR Medical Informatics*, *13*, e62710. https://doi.org/10.2196/62710
- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004
- Dahabreh, I. J., Haneuse, S. J. A., Robins, J. M., Robertson, S. E., Buchanan, A. L., Stuart, E. A., & Hernán, M. A. (2021). Study Designs for Extending Causal Inferences From a Randomized Trial to a Target Population. *American Journal of Epidemiology*, *190*(8), 1632–1642. https://doi.org/10.1093/aje/kwaa270
- Gelman, A., & Loken, E.R. (2014). The Statistical Crisis in Science. *American Scientist*, 102, 460. https://doi.org/10.1511/2014.111.460
- Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the Public Health Impact of Health Promotion Interventions: The RE-AIM Framework. *American Journal of Public Health*, *89*(9), 1322–1327. https://doi.org/10.2105/ajph.89.9.1322
- Harrell, F.E. (2015). *Regression Modeling Strategies*, 2nd ed. Springer.
- JASA Reproducibility Guide and Pre-Submission Author Checklist: https://jasa-acs.github.io/repro-guide/
- Lawlor, D. A., Tilling, K., & Davey Smith, G. (2016). Triangulation in Aetiological Epidemiology. *International Journal of Epidemiology*, *45*(6), 1866–1886. https://doi.org/10.1093/ije/dyw314
- Le, J. P., Morrison, J., Malhotra, A., Nemati, S., Wardi, G., & Ford, J. S. (2026). National Institutes of Health-Funded Artificial Intelligence and Machine Learning Research, 2019-2023: Cross-Sectional Study. *Journal of Medical Internet Research*, *28*, e84861. https://doi.org/10.2196/84861
- Li, F., Li, F., Liu, B., Papadogeorgou, G., Bobb, J., Thomas, L., & Wruck, L. (2024). *New Causal Inference Methods for Cluster-Randomized Trials with Postrandomization Selection Bias*. Patient-Centered Outcomes Research Institute (PCORI). https://doi.org/10.25302/10.2024.ME.2019C116146
- Lu, H., Li, F., Lesko, C. R., Fink, D. S., Rudolph, K. E., Harhay, M. O., Rentsch, C. T., Fiellin, D. A., & Gonsalves, G. S. (2024). Four Targets: An Enhanced Framework for Guiding Causal Inference from Observational Data. *International Journal of Epidemiology*, *54*(1), dyaf003. https://doi.org/10.1093/ije/dyaf003
- National Center for Complementary and Integrative Health. (2024a). *PAR-25-274: Feasibility Clinical Trials of Mind and Body Interventions for NCCIH High Priority Research Topics (R34 Clinical Trial Required).* National Institutes of Health. Posted November 21, 2024. Accessed August 17, 2026.
- National Center for Complementary and Integrative Health. (2024b). *PAR-25-268: Investigator Initiated Clinical Trials of Complementary and Integrative Interventions Delivered Remotely or via*

mHealth (R01 Clinical Trial Required). National Institutes of Health. Posted November 21, 2024. Accessed August 17, 2026.

- National Institutes of Health. (2024). *PA-25-305: NIH Research Project Grant (Parent R01 Clinical Trial Required).* Posted December 18, 2024. Accessed August 17, 2026.
- Peterson, R.A., & Cavanaugh, J.E. (2022). Ranked Sparsity: A Cogent Regularization Framework for Selecting and Estimating Feature Interactions and Polynomials. *AStA Advances in Statistical Analysis*, 106, 427–454. https://doi.org/10.1007/s10182-021-00431-7
- Peterson, R.A., McGrath, M., & Cavanaugh, J.E. (2024). Can a Transparent Machine Learning Algorithm Predict Better than Its Black Box Counterparts? A Benchmarking Study Using 110 Data Sets. *Entropy*, 26, 746. https://doi.org/10.3390/e26090746
- Peterson, R.A. (2025a). What Do We Mean by *Glass-box*, Exactly? *Data Diction (blog)*. https://doi.org/10.59350/9dhes-thd51
- Peterson, R.A. (2025b) [written under Hackman, P. pseudonym]. How Can I Guarantee a Significant Result? *Data Diction (blog)*. https://doi.org/10.59350/6mzf8-xzd69
- Peterson, R.A., Bird, S.M., Harris, L.M., Breheny, P.J., & Cavanaugh, J.E. (2026). A Ranked Sparsity Extension to the Bayesian Information Criterion: A Tool for Selecting Variables from Multiple Data Modalities. Preprint (currently in resubmission to *Entropy*). https://doi.org/10.20944/preprints202607.0193.v1
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632
- Tay, J.K., & Tibshirani, R. (2020). Reluctant Generalised Additive Modelling. *International Statistical Review*, 88(Suppl 1), S205–S224. https://doi.org/10.1111/insr.12429
- Wrobel, J., Hector, E.C., Crawford, L., D'Agostino McGowan, L., da Silva, N., Goldsmith, J., Hicks, S., Kane, M., Lee, Y., Mayrink, V., Paciorek, C.J., Usher, T., & Wolfson, J. (2024). Partnering with Authors to Enhance Reproducibility at JASA. *Journal of the American Statistical Association*, 119(546), 795–797. https://doi.org/10.1080/01621459.2024.2340557
- Yu, G., Bien, J., & Tibshirani, R. (2019). Reluctant Interaction Modeling. Preprint. https://doi.org/10.48550/arXiv.1907.08414

213. Americans for Medical Progress

I am responding: On behalf of an organization

Name: Naomi Charalambakis

Name of Organization: Americans for Medical Progress

Type of Organization: Non-Profit Research Organization

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Americans for Medical Progress (AMP) appreciates the opportunity to provide input on approaches for measuring and rewarding scientific impact. As a national nonprofit dedicated to promoting and protecting responsible research, AMP supports NIH's efforts to strengthen rigor, reproducibility, transparency, and reliable scientific findings.

AMP affirms the essential role of animal research in advancing biomedical discovery and furthering NIH's mission to improve human health. We also support NIH's efforts to advance New Approach Methodologies (NAMs) and other human-based research technologies, recognizing that these approaches and animal research are most effective when viewed as complementary tools rather than competing alternatives.

Given AMP's expertise in animal research, we offer recommendations within two of NIH's identified areas:
• Rigor and Reproducibility: recognizing the role of intrinsic and extrinsic variables in research outcomes, reproducibility, and animal welfare.
• Data, Software, and Model Sharing: recognizing model selection, characterization, validation, and scientific appropriateness as necessary foundations for meaningful model sharing.

Rigor and Reproducibility
Reliable, reproducible findings are essential to ensuring that federal research investments generate knowledge that can be trusted, built upon, and translated into meaningful health advances.

NIH has examined these issues extensively in the past, including through its 2021 Advisory Committee to the Director Working Group on Enhancing Rigor, Transparency, and Translatability in Animal Research. AMP encourages NIH to revisit and build upon the recommendations from this committee's final report, particularly those related to identifying, controlling, and reporting extrinsic variables that can influence experimental outcomes. Factors such as genetics, sex, age, diet, microbiome, housing conditions, social environment, handling, acclimation, environmental enrichment, lighting, noise, temperature, transportation, medications, veterinary interventions, and other aspects of animal care can affect physiology, behavior, and how researchers interpret findings.

To account for the conditions that may influence the biological system under investigation, AMP

recommends NIH consider indicators that encourage investigators to identify, manage, justify, and report relevant sources of variability, including those associated with animal care and the research environment. Greater attention to these factors would strengthen reproducibility while also reinforcing the close connection between high-quality animal care and high-quality science.

Potential indicators could include:
• Identification and transparent reporting of intrinsic and extrinsic variables relevant to the research question.
• Scientific justification for which variables are controlled, randomized, stratified, or intentionally varied.
• Incorporation of veterinary, animal-care, behavioral, and comparative-medicine expertise into study design where appropriate.
• Reporting of animal health, housing, husbandry, handling, and environmental conditions that may affect interpretation or reproducibility.
• Use of refinement practices that improve animal welfare while reducing unintended physiological or behavioral variability.

In applying these indicators, AMP advises NIH not to confuse scientific rigor with over-standardization, as excessive standardization may limit the generalizability of findings in many research cases. Depending on the scientific question, investigators may appropriately control certain variables while intentionally incorporating variation in others to determine whether findings remain robust across conditions. The relevant measure of rigor should therefore be whether investigators understand, account for, document, and appropriately interpret the variables that influence their experimental system.

Adopting these indicators also provides an opportunity to recognize more explicitly the scientific contribution of animal care and welfare practices. Appropriate acclimation, handling, enrichment, pain management, housing, training, and veterinary care can influence physiology and behavior and, consequently, the quality and interpretation of experimental data. Recognizing these practices within the context of rigor and reproducibility would reinforce that refinement can contribute directly to both animal welfare and scientific quality.

Taken together, these measures would build on NIH's prior work and support a more complete approach to rigor and reproducibility, one that reflects the biological and environmental context in which animal research is conducted. They would also reinforce the principle that careful attention to animal health, welfare, and experimental conditions is central to responsible biomedical research and supports more reliable and reproducible science.

2. Data, Software, and Model Sharing: AMP supports NIH's goal of increasing the sharing of scientific data, tools, software, and research models. Greater sharing can facilitate independent validation, reduce unnecessary duplication, enable new analyses, and maximize the value of federally funded research.

However, AMP encourages NIH to expand this category to recognize and reward an important component that is not currently captured: model selection, characterization, validation, and scientific appropriateness. Before investigators are incentivized to disseminate a model for broader use, there should be sufficient evidence to define what the model represents, the scientific questions it is suited to address, its known limitations, and the context in which its use is appropriate.

This consideration is especially important in the current policy and communications environment, where animal models and NAMs or other human-based approaches are too often framed as competing alternatives. Modern biomedical research increasingly relies on a diverse range, and often a combination, of experimental systems, each with distinct strengths, limitations, and applications. As a foundational principle, AMP strongly encourages NIH to make clear that no model or methodology should be presumed more rigorous, innovative, predictive, or scientifically valuable solely because it falls within a particular category. These determinations should be based on the scientific question, the evidence supporting the model's use, and its demonstrated fitness for purpose.

Accordingly, NIH should consider indicators that recognize the scientific work required to establish a model's utility and appropriate application before rewarding its dissemination. Potential indicators could include:

- A clear scientific rationale for selection of the model or experimental system.
- Evidence that the model is appropriate for the biological question being studied.
- Characterization of the model's strengths, limitations, and relevant performance expectations.
- Validation or comparative assessment against other relevant approaches, where scientifically appropriate.
- Definition of the model's intended context of use, including circumstances in which its findings may have limited applicability.
- Consideration of whether complementary models or approaches are needed to address the scientific question.
- Availability of sufficient methodological information to support appropriate interpretation and reuse by other investigators.

These considerations are particularly timely as NIH and other federal agencies increase support for NAMs and other human-based research approaches. As these methods are developed and incorporated into biomedical research, it will be important to evaluate them using the same scientific standards applied to any other model, including animal studies. Their value should be demonstrated through evidence of reliability, relevance, and fitness for purpose, rather than inferred from novelty or from reduced animal use alone.

Regarding incentives, AMP recommends NIH be deliberate in defining what it chooses to reward in this area. Incentives should emphasize the evidence needed to determine whether a model is reliable, fit for purpose, and appropriately applied, rather than simply whether it is new, widely adopted, or associated with reduced animal use. Model sharing is most valuable when it rests on a sound scientific foundation and gives other investigators sufficient information to understand not only how a model can be used, but also when, why, and for what purposes its use is appropriate.

By incorporating model selection, characterization, validation, and scientific appropriateness into this category, NIH can ensure that incentives reward the utility, reliability, and appropriate application of shared resources rather than dissemination alone. This approach would also preserve the scientific flexibility necessary for investigators to use animal models, human-based technologies, computational approaches, or combinations of methods according to the needs of the research question.

Conclusion
As NIH considers new ways to recognize and reward scientific impact, AMP urges the agency to preserve the scientific judgment and methodological flexibility that are essential to biomedical progress. Incentives should reward the quality of the science, regardless of the model or methodology used, and should not create a preference for animal models, NAMs, or other human-based approaches simply by virtue of the category into which they fall. Rather, NIH should encourage better model selection and validation, careful attention to intrinsic and extrinsic variables, high standards of animal care and welfare, and transparent reporting.

NIH may also benefit from continued consultation with relevant stakeholder groups, including scientific societies, research institutions, veterinarians, animal care professionals, and organizations such as AMP, as these recommendations move toward implementation. Ongoing engagement can help ensure that any new indicators are scientifically meaningful, practical to apply, and responsive to the realities of research across different settings.

Given the breadth of NIH's research portfolio, flexibility in implementation will also be important. Appropriate indicators may vary by scientific discipline, species, model system, career stage, institution type, and research objective. A framework that establishes clear expectations while allowing for these differences will better support rigor, validation, transparency, and responsible research practices across the diverse programs and institutions NIH supports.

By grounding incentives in scientific appropriateness, rigor, and reliable outcomes, NIH can help ensure that efforts to modernize biomedical research strengthen the scientific foundation needed to advance its mission and improve human and animal health.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/AMP-RFI-Response_MeasuringScientificImpact_FINAL.pdf

Description: PDF copy of AMP's Response to NOT-OD-26-047

NIH Office of the Director
Office of Science Policy
6705 Rockledge Drive, Suite 630
Bethesda, MD 20892

Re: Request for Information (RFI) on Measuring and Rewarding Scientific Impact (NOT-OD-26-047)

Submitted electronically via response portal and e-mail: SciencePolicy@od.nih.gov

Dear Dr. Jorgenson,

Americans for Medical Progress (AMP) appreciates the opportunity to provide input on approaches for measuring and rewarding scientific impact. As a national nonprofit dedicated to promoting and protecting responsible research, AMP supports NIH's efforts to strengthen rigor, reproducibility, transparency, and reliable scientific findings.

AMP affirms the essential role of animal research in advancing biomedical discovery and furthering NIH's mission to improve human health. We also support NIH's efforts to advance New Approach Methodologies (NAMs) and other human-based research technologies, recognizing that these approaches and animal research are most effective when viewed as complementary tools rather than competing alternatives.

Given AMP's expertise in animal research, we offer recommendations within two of NIH's identified areas:

- **Rigor and Reproducibility:** recognizing the role of intrinsic and extrinsic variables in research outcomes, reproducibility, and animal welfare.
- **Data, Software, and Model Sharing:** recognizing model selection, characterization, validation, and scientific appropriateness as necessary foundations for meaningful model sharing.

Rigor and Reproducibility

Reliable, reproducible findings are essential to ensuring that federal research investments generate knowledge that can be trusted, built upon, and translated into meaningful health advances.

NIH has examined these issues extensively in the past, including through its 2021 Advisory Committee to the Director Working Group on Enhancing Rigor, Transparency, and Translatability in Animal Research.[1] AMP encourages NIH to revisit and build upon the recommendations from this committee's final report,[2] particularly those related to identifying, controlling, and reporting extrinsic variables that can influence experimental outcomes. Factors such as genetics, sex, age, diet, microbiome, housing conditions, social environment, handling, acclimation, environmental enrichment, lighting, noise, temperature, transportation, medications, veterinary

[1] National Institutes of Health. (2021). *Advisory Committee to the Director Working Group on Enhancing Rigor, Transparency, and Translatability in Animal Research*. https://acd.od.nih.gov/working-groups/eprar.html
[2] National Institutes of Health, Advisory Committee to the Director, Working Group on Enhancing Rigor, Transparency, and Translatability in Animal Research. (2021, June 11). *Final report.* *https://acd.od.nih.gov/documents/presentations/06112021_ACD_WorkingGroup_FinalReport.pdf*

interventions, and other aspects of animal care can affect physiology, behavior, and how researchers interpret findings.

To account for the conditions that may influence the biological system under investigation, AMP recommends NIH consider indicators that encourage investigators to identify, manage, justify, and report relevant sources of variability, including those associated with animal care and the research environment. Greater attention to these factors would strengthen reproducibility while also reinforcing the close connection between high-quality animal care and high-quality science.

Potential indicators could include:

- Identification and transparent reporting of intrinsic and extrinsic variables relevant to the research question.
- Scientific justification for which variables are controlled, randomized, stratified, or intentionally varied.
- Incorporation of veterinary, animal-care, behavioral, and comparative-medicine expertise into study design where appropriate.
- Reporting of animal health, housing, husbandry, handling, and environmental conditions that may affect interpretation or reproducibility.
- Use of refinement practices that improve animal welfare while reducing unintended physiological or behavioral variability.

In applying these indicators, AMP advises NIH not to confuse scientific rigor with over-standardization, as excessive standardization may limit the generalizability of findings in many research cases. Depending on the scientific question, investigators may appropriately control certain variables while intentionally incorporating variation in others to determine whether findings remain robust across conditions. The relevant measure of rigor should therefore be whether investigators understand, account for, document, and appropriately interpret the variables that influence their experimental system.

Adopting these indicators also provides an opportunity to recognize more explicitly the scientific contribution of animal care and welfare practices. Appropriate acclimation, handling, enrichment, pain management, housing, training, and veterinary care can influence physiology and behavior and, consequently, the quality and interpretation of experimental data. Recognizing these practices within the context of rigor and reproducibility would reinforce that refinement can contribute directly to both animal welfare and scientific quality.

Taken together, these measures would build on NIH's prior work and support a more complete approach to rigor and reproducibility, one that reflects the biological and environmental context in which animal research is conducted. They would also reinforce the principle that careful attention to animal health, welfare, and experimental conditions is central to responsible biomedical research and supports more reliable and reproducible science.

Data, Software, and Model Sharing

AMP supports NIH's goal of increasing the sharing of scientific data, tools, software, and research models. Greater sharing can facilitate independent validation, reduce unnecessary duplication, enable new analyses, and maximize the value of federally funded research.

However, AMP encourages NIH to expand this category to recognize and reward an important component that is not currently captured: **model selection, characterization, validation, and scientific appropriateness**. Before investigators are incentivized to disseminate a model for broader use, there should be sufficient evidence to define what the model represents, the scientific questions it is suited to address, its known limitations, and the context in which its use is appropriate.

This consideration is especially important in the current policy and communications environment, where animal models and NAMs or other human-based approaches are too often framed as competing alternatives. Modern biomedical research increasingly relies on a diverse range, and often a combination, of experimental systems, each with distinct strengths, limitations, and applications. As a foundational principle, AMP strongly encourages NIH to make clear that no model or methodology should be presumed more rigorous, innovative, predictive, or scientifically valuable solely because it falls within a particular category. These determinations should be based on the scientific question, the evidence supporting the model's use, and its demonstrated fitness for purpose.

Accordingly, NIH should consider indicators that recognize the scientific work required to establish a model's utility and appropriate application before rewarding its dissemination. Potential indicators could include:

- A clear scientific rationale for selection of the model or experimental system.
- Evidence that the model is appropriate for the biological question being studied.
- Characterization of the model's strengths, limitations, and relevant performance expectations.
- Validation or comparative assessment against other relevant approaches, where scientifically appropriate.
- Definition of the model's intended context of use, including circumstances in which its findings may have limited applicability.
- Consideration of whether complementary models or approaches are needed to address the scientific question.
- Availability of sufficient methodological information to support appropriate interpretation and reuse by other investigators.

These considerations are particularly timely as NIH and other federal agencies increase support for NAMs and other human-based research approaches. As these methods are developed and incorporated into biomedical research, it will be important to evaluate them using the same scientific standards applied to any other model, including animal studies. Their value should be demonstrated through evidence of reliability, relevance, and fitness for purpose, rather than inferred from novelty or from reduced animal use alone.

Regarding incentives, AMP recommends NIH be deliberate in defining what it chooses to reward in this area. Incentives should emphasize the evidence needed to determine whether a model is reliable, fit for purpose, and appropriately applied, rather than simply whether it is new, widely adopted, or associated with reduced animal use. Model sharing is most valuable when it rests on a sound scientific foundation and gives other investigators sufficient information to understand not only how a model can be used, but also when, why, and for what purposes its use is appropriate.

By incorporating model selection, characterization, validation, and scientific appropriateness into this category, NIH can ensure that incentives reward the utility, reliability, and appropriate application of shared resources rather than dissemination alone. This approach would also preserve the scientific flexibility necessary for

investigators to use animal models, human-based technologies, computational approaches, or combinations of methods according to the needs of the research question.

Conclusion

As NIH considers new ways to recognize and reward scientific impact, AMP urges the agency to preserve the scientific judgment and methodological flexibility that are essential to biomedical progress. Incentives should reward the quality of the science, regardless of the model or methodology used, and should not create a preference for animal models, NAMs, or other human-based approaches simply by virtue of the category into which they fall. Rather, NIH should encourage better model selection and validation, careful attention to intrinsic and extrinsic variables, high standards of animal care and welfare, and transparent reporting.

NIH may also benefit from continued consultation with relevant stakeholder groups, including scientific societies, research institutions, veterinarians, animal care professionals, and organizations such as AMP, as these recommendations move toward implementation. Ongoing engagement can help ensure that any new indicators are scientifically meaningful, practical to apply, and responsive to the realities of research across different settings.

Given the breadth of NIH's research portfolio, flexibility in implementation will also be important. Appropriate indicators may vary by scientific discipline, species, model system, career stage, institution type, and research objective. A framework that establishes clear expectations while allowing for these differences will better support rigor, validation, transparency, and responsible research practices across the diverse programs and institutions NIH supports.

By grounding incentives in scientific appropriateness, rigor, and reliable outcomes, NIH can help ensure that efforts to modernize biomedical research strengthen the scientific foundation needed to advance its mission and improve human and animal health.

Sincerely,



Paula Clifford, MLA, RLATG, CVT
Executive Director
Americans for Medical Progress
paula@amprogress.org
(202) 624-8812

214. Pengyin Shan

I am responding: On behalf of myself

Name: Pengyin Shan

Name of Organization: University of Illinois Urbana-Champaign, National Center for Supercomputing Applications (NCSA)

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Reproducing a computational finding requires identifying the exact software artifact used, and artifact identification fails when a project's own surfaces disagree about version and identifier: in a recent audit of 117 open-source research software projects (arXiv:2608.17159), version agreement across comparable metadata surface pairs was 0.447. A measurable indicator: publications and data deposits carry a resolvable software identifier whose version matches the project's distribution record. As implementation infrastructure, registered analysis protocols with released verification logs make audits themselves replicable; the study above deposited its sampling protocol with a DOI before measurement and released a complete hand-verification log (98.5% verified instrument precision), a pattern NIH could recognize and reward in methodological work generally.

2. Data, Software, and Model Sharing: Incentivizing software sharing presupposes that shared software can be credited, and crediting presupposes that the software's machine-readable citation metadata is consistent. The audit cited above suggests it is not. Across up to seven self-description surfaces per project (CITATION.cff, codemeta.json, .zenodo.json, resolved DOI records, package registry records, README citation content), 83.9% of projects exposing at least two comparable surfaces conflicted with themselves on at least one citation-critical field (title, authors, version, year, or DOI). Half of hand-verified conflicts traced to surfaces describing the software's paper rather than the software, and among projects whose citation file designates a preferred citation, 28 of 32 route citations to a record that disagrees with the software's own metadata (open data: doi.org/10.5281/zenodo.21969769).

Three measurable indicators follow. First, presence of machine-readable citation metadata (a citation file, exchange-schema record, or archival DOI) on funded software: cheap, automatable, and currently a minority practice (41% of audited projects carry CITATION.cff; 5% codemeta.json). Second, cross-surface metadata consistency as a quality indicator, automatically checkable at release time; an open Apache-2.0 audit pipeline demonstrating feasibility has been released (doi.org/10.5281/zenodo.21969695). Third, an archival DOI whose registered metadata agrees with the repository's own declarations, as a condition for counting software as "shared."

One unintended consequence deserves emphasis: attaching rewards to software citation counts before the metadata substrate is consistent would mismeasure, since inconsistent surfaces split citation counts

across records and route credit to papers rather than software. Fixing the measurement substrate is a precondition for the incentive, not an optional refinement. Feasibility is favorable: the audit above was performed by a single investigator with public APIs; the same checks embed in release workflows at negligible cost, and the method is domain-general, with the paper-proxy pattern independently documented in biomedical corpora (Du et al., JASIST 2021; PeerJ Computer Science 2022).

3. Training and Mentorship: N/A

4. Collaboration: Team-science credit requires persistent person identifiers on the surfaces that tooling actually reads. In the audit cited above, measured ORCID coverage varies from 62% of authors in archive metadata and 54% in citation files to 39% in resolved DOI records and 0% in package registry records and README citation blocks, so the surfaces most consumed by automated tooling carry the least contributor-level attribution. A simple indicator: ORCID coverage on machine-readable software surfaces of funded projects. The benefit is contributor-level credit in team-built software, where repository metadata frequently collapses many contributors into an entity string.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/2608.17159v1.pdf

Description: Cited preprint: A Multi-Surface Consistency Audit of Software Citation Metadata (arXiv:2608.17159)

arXiv:2608.17159v1 [cs.SE] 17 Aug 2026

DOES RESEARCH SOFTWARE AGREE WITH ITSELF? A MULTI-SURFACE CONSISTENCY AUDIT OF SOFTWARE CITATION METADATA

A PREPRINT

Pengyin Shan
National Center for Supercomputing Applications
University of Illinois Urbana-Champaign
Urbana, IL 61820
pengyins@illinois.edu

August 19, 2026

ABSTRACT

Research software projects describe themselves in many places at once: citation files in the repository, archive deposits, DOI registry records, package registries, and README text. We treat the software as the underlying object and these machine-readable self-descriptions as its surfaces: the points where people and automated systems read what the project declares about the software. Citation guidance, indexing services, and automated agents may read a different subset of these surfaces, so disagreement between them can silently fragment credit and provenance. This paper asks a simple question that has not been measured directly: when a project's own metadata surfaces are compared with each other, how often do they agree? We audited 117 open-source research software projects, comprising an 87-project high-performance computing and quantum computing corpus and a 30-project registered baseline drawn from the JOSS and pyOpenSci accepted-package lists, across up to seven machine-readable surfaces per project. Using a four-level verdict rubric across six metadata fields, with 98.5% hand-verified verdict precision on a 338-row stratified sample, we found that 52 of the 62 projects exposing at least two comparable surfaces (83.9%) contain at least one core-field conflict, a result that is insensitive to the fuzzy-matching threshold. Half of hand-adjudicated cross-surface conflicts trace to a single mechanism: surfaces describing the software's paper rather than the software itself. Among projects whose CITATION.cff includes a preferred citation, 28 of 32 route citations to a record that disagrees with the software's own metadata. The author lists and titles disagree the most, and the registry surfaces are the least aligned. We release the audit pipeline as an importable library, the corpus, the registered sampling protocol, all raw snapshots, and the complete verification log.

Keywords software citation · metadata consistency · research software · CITATION.cff · provenance · scholarly infrastructure

1 Introduction

The software citation infrastructure has largely solved the problem of *where* a project can declare its citation metadata. CITATION.cff [1], codemeta.json [2], archived deposits with DOIs [3], package registry metadata, and README citation sections are all established, machine-readable, and widely promoted. What this infrastructure does not guarantee is that these surfaces say the same thing. Each could be written at a different time, by a different process, for a different consumer: a CITATION.cff is edited by software developers or maintainers, a Zenodo record may be minted automatically from a GitHub release, a PyPI record is generated at the packaging time, and a README BibTeX block is often pasted directly from a paper's landing page and never updated.

Consumers of citation metadata, whether human authors following a "how to cite" section, automated indexers resolving a DOI, or Large Language Models (LLMs) assembling citations, each read one surface and trust it. If the surfaces

disagree, which record a citation carries depends on which surface the consumer happened to read. The consequences are familiar from software citation research: split citation counts, misattributed credit, version ambiguity, and paper-software conflation [4–6]. What has been missing is a direct measurement of how often a project's own self-descriptions disagree, across surfaces, at scale, with verified instrument precision.

This paper contributes that measurement. We framed the problem as a *supply-side* audit: we frame what projects declare across multiple machine-readable surfaces they expose, and score pairwise consistency. Our contributions:

- **An audit instrument.** An open pipeline (harvest, normalize, compare, report) covering seven surfaces: CITATION.cff (root and preferred-citation), codemeta.json, .zenodo.json, the resolved DOI record, PyPI and npm registry records, and README citation content. The comparison rubric has four verdict levels over six fields, with all thresholds documented and released.
- **A measured corpus with a registered baseline.** 87 HPC and quantum computing projects from a prior, DOI-anchored corpus [7], plus a 30-project baseline drawn from the accepted-package lists of JOSS and pyOpenSci under a sampling protocol fixed and timestamped before candidate acceptance.
- **Verified findings.** 83.9% of projects with at least two comparable surfaces disagree with themselves on at least one core field, with verdict precision hand-verified at 98.5% over a stratified 338-row sample and results insensitive to the matching threshold.
- **A mechanism taxonomy.** Half of adjudicated cross-surface conflicts trace to paper-versus-software conflation, and the remainder cluster into archive staleness, registry ghost metadata, and identity ambiguities, each illustrated with named, snapshot-dated examples.

2 Related Work

The FORCE11 software citation principles established software as a citable research product and specified what a citation must identify: the artifact actually used, through its authors, title, version, date, and a unique persistent identifier [4]. The FAIR principles for research software carry the same position into metadata practice, tying findability and reuse to rich, machine-readable description of the software itself [8]. Reviewing the implementation two years after the principles, Katz and Chue Hong reported that the infrastructure had arrived, with over 50,000 DOIs issued for software at the time, while challenges remained in developer practice and in publisher systems retaining the required metadata [6]. Five of the six fields our rubric compares (title, authors, version, year, and DOI) are the fields this line of work designates as citation-critical; the sixth, license, is carried by every machine-readable surface we audit. Our question is not whether projects declare these fields but whether the surfaces declaring them agree.

Declaration infrastructure followed the principles. The Citation File Format gives maintainers a plain-text citation file in the repository root [1], and GitHub renders that file as the "cite this repository" widget on the repository landing page, so a single YAML file now reaches every visitor. CodeMeta provides a JSON-LD exchange schema with crosswalks between registry and archive vocabularies [2], and archive DOI records are registered against the DataCite metadata schema [3]. Adoption of this infrastructure is measured at registry scale: El Hounsri and Garijo analyzed research software metadata adoption across the software registries of five European Open Science clusters and a multi-disciplinary registry and found recommended practice unevenly followed [9]. Adoption studies of this kind have established whether a surface is present and well-formed, but none has asked whether the surfaces a project has adopted describe the same object in the same way, which is the measurement this paper adds.

Where metadata quality is studied, it is studied one surface at a time. On the paper side, Howison and Bullard examined how software is mentioned in the biology literature and found mentions that were informal and incomplete, lacking the version and identifier information needed to find the artifact used [5]. On the registry side, Bommarito and Bommarito characterized package metadata across 178,592 PyPI packages [10], and Gao et al. measured that existing tools can recover source-repository information for at most 70.5% of PyPI releases because release metadata omits the repository or names the wrong one, and built PyRadar to retrieve and validate those links [11]. The newest large-scale datasets aggregate in the same direction: OpenDORS compiles license, version, and language metadata for over 120,000 openly referenced research software repositories, one record per project [12]. Each of these efforts measures the presence or quality of a single surface across many projects, or merges a project's surfaces into one record. To our knowledge, no prior study compares a project's own self-description surfaces against each other, so the rate at which research software disagrees with itself has not been measured. That distinction also separates this audit from adoption analyses such as [9] as counting surfaces and cross-checking them are different measurements, and only the first exists in the literature.

A separate line of work documents the paper standing in for the software. The Journal of Open Source Software aims to create a peer-reviewed, citable paper, so that software work can be credited through existing publication infrastructure [13]; which citation carries is left to authors and citing practice. Corpus studies show how the proxy pattern emerges

in that practice. The Softcite dataset, built from expert annotation of 4,971 biomedical and economics publications, found that software is mentioned informally far more often than it is formally cited [14], and a follow-up study of the CORD-19 corpus found that even where projects request a specific citation, identifying the canonical record among a project's several candidate sources is difficult [15]. These studies measure conflation where it lands in the citing literature. Our mechanism analysis in Section 4.4 measures it one step earlier: in the machine-readable metadata projects themselves publish, where half of verified cross-surface conflicts trace to a surface describing the paper rather than the software.

Supply-chain security frameworks audit adjacent layers of the same projects. OpenSSF Scorecard scores repository and maintenance practices automatically [16]; SLSA defines graded levels of build provenance, and since the specification is community-maintained without a peer-reviewed primary description, we reference the qualitative study of its deployment challenges by Tamanna et al. [17]; Sigstore provides signing infrastructure for released artifacts [18]. These frameworks address whether code and artifacts are what they claim to be. None checks whether a project's self-descriptions agree with each other, so the audit reported here is complementary to them and overlaps none of their controls.

3 Methods

3.1 Corpus

The corpus contains 117 projects in three strata. The **supercomputing stratum** (labeled `sc26` in the released data files; 87 projects; 44 HPC, 43 quantum computing) is exported verbatim from the tagged v1.0.0 release of our prior provenance-audit corpus [7], where the derivation from the tagged release keeps the corpus identity claim checkable. The **baseline strata** (15 JOSS, 15 pyOpenSci) exist to test whether findings from infrastructure-heavy HPC/Quantum software generalize to community-reviewed research software, and were drawn as follows.

Sampling frames were retrieved on 2026-08-12 from the JOSS published-papers API (3,662 records after retrieval) and the pyOpenSci packages registry (70 records). Corresponding software repositories were required to be GitHub-hosted, since all six eligibility surfaces are probed through the GitHub API. Non-GitHub candidates were excluded at sampling under rule R5 (JOSS: 185, pyOpenSci: 4, hosts reported in the exclusion log). Four further pre-registered rules applied in order: R1 corpus overlap with the 87 (7 exclusions), R2 dual-listing resolved to the pyOpenSci stratum (32), R3 JOSS version papers deduplicated to the most recent (22), R4 missing repository URL (0). Domain bins were assigned from a committed keyword map (first match wins; disciplinary keywords precede generic ones). Candidates with no keyword matched were retained in an explicit "unmapped" bin that participates in the stratified draw rather than being excluded, so the draw does not depend on the coverage of a hand-written map. The registered unmapped share was 33% (JOSS) and 2% (pyOpenSci). One design correction was made before the protocol was registered: an earlier rule that refused to finalize while unmapped candidates appeared in the draw window was unsatisfiable under round-robin interleaving, since the unmapped bin is guaranteed window slots by construction, and it was replaced by unmapped-bin participation plus a 50% frame-level ceiling.

A seeded, domain-stratified draw (seed 202608121338, round-robin over domain bins, oversample 30 per stratum) produced ranked candidate windows. The sampling protocol, ranked candidate lists, and exclusion log were committed (baseline-sampler v0.1.0) and deposited under restricted access with a DOI before any eligibility probing (10.5281/zenodo.21909720; the deposit timestamp precedes the acceptance log timestamp). Eligibility required at least 2 of six machine-readable surfaces, probed per candidate: CITATION.cff, codemeta.json, .zenodo.json in the repository root, a README citation surface (citation section or BibTeX block; the registered rule), a discoverable DOI (from CFF or the README citation section, not resolved at probe time) and a verified registry listing (a PyPI or npm package matching the repository name and linking back to the repository). Acceptance proceeded strictly in rank order to 15 per stratum: JOSS accepted ranks 2-18 (3 skips for insufficient surfaces), pyOpenSci ranks 1-19 (4 skips). One drawn candidate's declared repository did not resolve at probe time and was never reached by acceptance. 15 of the 30 pyOpenSci window candidates are dual-listed with JOSS and therefore carry a JOSS paper DOI, so the DOI-by-construction confound noted below extends partially into that stratum.

Hand-curated concept DOIs were filled for the supercomputing stratum only (20 projects), drawn from the prior corpus's distribution-metadata records; both baseline strata rely entirely on automatic discovery from project surfaces, keeping the baseline availability numbers free of curation effects. Registry package names were detected uniformly for all 117 projects by name-part lookup verified by repository back-link and were hand-reviewed. Two detections were removed as packages not controlled by the project.

3.2 Surfaces

For each project we harvest up to seven surfaces: `cff` (CITATION.cff root), `cff_preferred` (its preferred-citation block, extracted as a pseudo-surface so intra-file disagreement is measurable), `codemeta` (codemeta.json), `zenodo_json` (.zenodo.json), `doi_record` (the DOI's registered metadata, resolved via DataCite [3] with Crossref fallback; DOI priority order: corpus-curated concept DOI, then CFF-declared, then README citation section, with provenance recorded), `pypi` and `npm` (registry records), and `readme` (a heuristic parse of the README's citation section, BibTeX block, and DOIs). GitHub's "cite this repository" widget is generated from CITATION.cff and is covered by the `cff` surface. The README surface is defined as a README file in the repository root; GitHub additionally renders docs/ and .github/ READMEs on the landing page, which this audit does not credit (one corpus project is affected). Every API response is snapshotted with a retrieval timestamp; all snapshots were retrieved on 2026-08-14 and are released in the data deposit.

3.3 Normalization

Each surface is normalized into one record shape: title, authors (family, given, ORCID), version, year, DOI, license. Extraction rules are released with the pipeline, where three deserve note: BibTeX in READMEs is extracted with brace-matching rather than regular expressions, after live testing showed regex extraction silently dropping fields across common forms; DOI strings are sanitized (resolver-URL prefixes, badge-image suffixes, trailing punctuation) and validated against the DOI pattern before resolution; DataCite creator records frequently carry the full name in familyName with no givenName so the normalizer uses record-supplied name parts as-is and falls back to name-splitting only when the record supplies neither part.

3.4 Comparison rubric

For every pair of surfaces present, each of the six fields receives one of four verdicts. **exact**: normalized values identical. **minor**: clearly the same referent with superficial variation (title token-sort fuzzy >= 90; the same author set reordered or fuzzy-matched; year differing by exactly 1; version differing only by a v prefix). **conflict**: both present and materially different. **missing**: at least one side absent. Zenodo concept-versus-version DOI pairs are flagged for hand adjudication rather than auto-resolved. All thresholds live in one released module.

Headline results are insensitive to the fuzzy-match threshold: at 85 and 95 the share of projects whose surfaces contain at least one core-field conflict is unchanged (83.9%, with identical per-stratum rates), while field-level agreement shifts modestly (authors 0.459-0.511 around the reported 0.481; title 0.436-0.442; median per-project pairwise agreement 0.50-0.55). Threshold-sensitivity outputs are released alongside the results.

3.5 Verification protocol and instrument disclosure

We hand-verified the instrument's verdicts against the harvested snapshots rather than the live web, since precision measures whether the pipeline judged correctly given the evidence it retrieved. The verification set comprised every conflict verdict, every minor verdict, every intra-file conflict between CFF root and preferred-citation metadata, and a seeded random sample of 30 exact verdicts. For each judgment, we re-derived the verdict by hand from the recorded value pair; identical value pairs recurring across surface pairs within a project were adjudicated once and the judgment propagated to each covered comparison row. Raw snapshots were consulted where a recorded value appeared anomalous. Rows flagged as possible Zenodo concept-versus-version identifier pairs were additionally adjudicated by resolving both identifiers.

Verified precision was 98.5% over 338 comparison rows (275 deduplicated judgments): conflict 276/281 (98.2%), minor 26/26, exact 31/31. The five disagreeing rows classify into two mechanisms, both documented heuristic limits: threshold boundary (2 rows; an author-entity name variant scoring just below the fuzzy threshold) and parse limitations (3 rows; an organizational author split at an embedded "and", and a LaTeX-escaped accented character blocking one name match). The complete verification log is released.

Instrument defects found and corrected before the reported measurements, in the honest-reporting style this protocol requires: two normalization defect classes were identified in pre-run live testing (YAML date serialization in CITATION.cff snapshots; .zenodo.json license objects producing spurious license conflicts); three DOI-string hygiene classes were identified in first-run review (resolver-URL prefixes, badge-image suffixes, non-DOI identifier values); and hand verification identified one systematic defect (DataCite creator records carrying full names in familyName caused token duplication and depressed author matching), which was corrected and the affected comparisons re-scored and re-verified. The instrument's stage history (v0.2.0 pre-run; v0.2.1 first-run corrections; v0.2.2 verification-corrected

Table 1: Availability of self-description surfaces across the corpus (n = 117; all projects resolvable at snapshot time).

Surface	Availability
README (root-level)[a]	116/117 (99%)
Stable releases	111/117 (95%)
CITATION.cff[b]	48/117 (41%)
Registry listing (PyPI or npm)[c]	46/117 (39%)
Archival DOI (doi_record)	20/117 (17%)
SECURITY.md	16/117 (14%)
.zenodo.json[d]	10/117 (9%)
codemeta.json	6/117 (5%)

[a] One project (preCICE) keeps its README under docs/; under the root-only rule it is counted absent (see Section 3.2). [b] By stratum: 26/87 supercomputing, 10/15 JOSS, 12/15 pyOpenSci. [c] One project is npm-listed; it is also PyPI-listed. [d] By stratum: 3/87 supercomputing, 1/15 JOSS, 6/15 pyOpenSci.

Table 2: Pairwise agreement by field across comparable surface pairs, with and without registry-involved pairs. Registry exclusion is shown only for the two fields where registry pairs change the rate materially.

	All comparable pairs		Excluding registry pairs	
Field	n	Agreement	n	Agreement
Year	50	.780		
DOI	61	.770		
License	47	.766		
Authors	135	.481	98	.612
Version	47	.447		
Title	165	.436	90	.644

Of 93 title conflicts, 61 involve a registry surface.

final) is documented above and in the repository README; the public repository history was flattened for release, and v0.2.2 produced every number reported here.

4 Results

4.1 Surface availability

All 117 repositories resolved at snapshot time. Table 1 reports surface availability. The human-facing substrate is near-universal: README in 116 of 117 projects and stable releases in 95%. The machine-readable citation surfaces are minority practice, led by CITATION.cff at 41% and thinning to codemeta.json at 5%. 62 of 117 projects (53%) expose at least two comparable surfaces; the consistency results below concern those 62. That only half the corpus can be cross-checked at all is itself a finding: for the other half, whatever the single surface says is uncheckable by construction.

4.2 Pairwise consistency

Across 1,104 field comparisons (505 comparable, 599 with at least one side missing), 52 of 62 projects (83.9%) contain at least one core-field conflict (title, authors, version, year, or DOI), and median per-project pairwise agreement is 0.50. By field, agreement among comparable pairs is: year 0.78, DOI 0.77, license 0.77, authors 0.48, version 0.45, title 0.44. Figure 1 shows the composition behind these rates: of 1,104 field comparisons, 599 could not be made at all because at least one surface omits the field, so the agreement rates in Table 2 describe only the comparable remainder.

The registry surfaces drive much of the title and authors disagreement: 61 of 93 title conflicts involve PyPI or npm, largely the structural gap between a registry's package name and a descriptive title ("mpi4py" versus "MPI for Python"). Excluding registry pairs, title agreement rises from 0.44 to 0.64 and authors from 0.48 to 0.61. Figure 2 locates the disagreement: pairs involving a registry surface concentrate the low agreement values, while the repository-native pairs (cff with codemeta, codemeta with .zenodo.json) agree most. We report both because both are real: a human reader distinguishes a package name from a title, but a metadata consumer joining records across surfaces sees a mismatch.

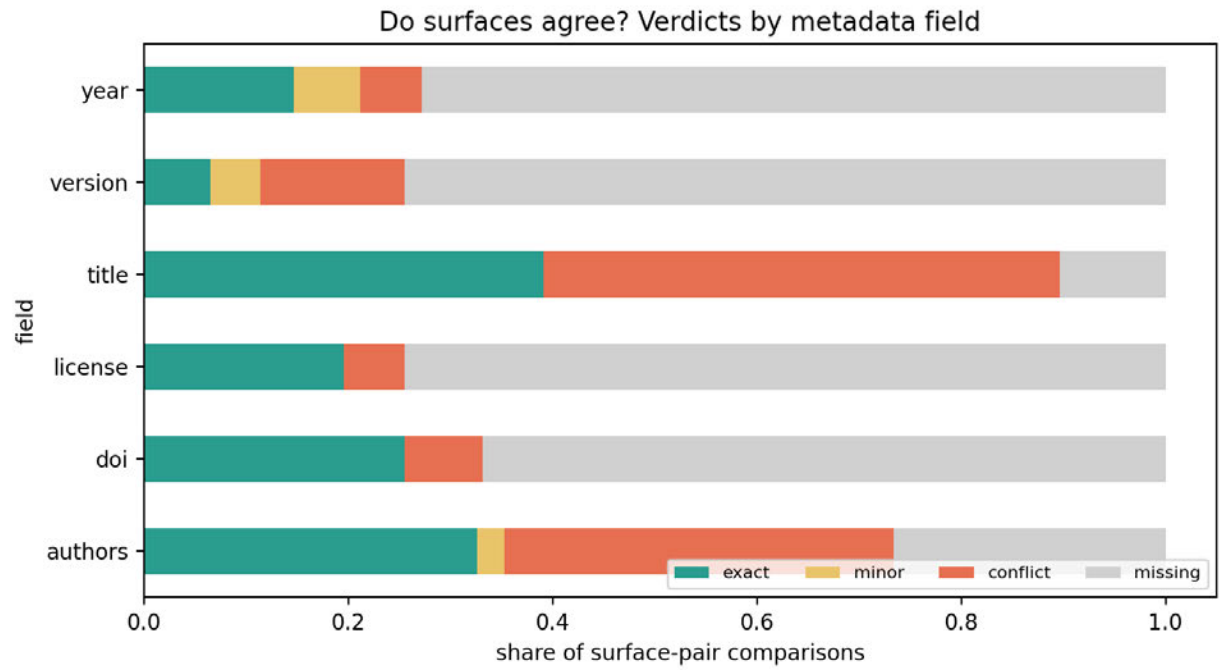


Figure 1: Verdict composition per field across all 1,104 surface-pair field comparisons. Gray indicates the field was not comparable because at least one surface omits it; Table 2 reports agreement among the comparable remainder.

Table 3: Consistency by stratum among projects with at least two comparable surfaces. Agreement is pooled pairwise agreement: exact plus minor verdicts over comparable pairs within the stratum.

Stratum	Projects	Comparable pairs	Agreement	≥1 core-field conflict
supercomputing	36	236	.559	86.1%
JOSS	11	97	.649	63.6%
pyOpenSci	15	172	.494	93.3%
Overall	62	505	.554	83.9%

The median per-project pairwise agreement across the 62 projects is .500.

By stratum, the phenomenon is not an HPC idiosyncrasy: core-field conflict affects 86.1% of supercomputing projects (31/36 with >= 2 surfaces), 63.6% of JOSS projects (7/11), and 93.3% of pyOpenSci projects (14/15). Two cautions attach to the baseline comparison: the JOSS stratum's DOI field is partially DOI-by-construction (every JOSS paper carries a DOI, and 15 of 30 pyOpenSci window candidates are dual-listed), and per-stratum denominators are small (11-36 projects), so we report the direction (community-reviewed packages are not cleaner; pyOpenSci projects expose more surfaces and correspondingly more disagreement) without significance claims. One tautology is excluded from interpretation: when the resolved DOI was discovered from a project's own CFF or README, the DOI-field agreement between the discovering surface and the DOI record is expected; informative DOI comparisons are those between non-source surfaces, including the four adjudicated concept-versus-version pairs.

4.3 The project's own citation file disagrees with itself

32 projects' CITATION.cff files include a preferred-citation block. In 28 of 32 (87.5%), following the preferred-citation yields a record that conflicts with the software's own metadata on at least one field (56 conflicting field comparisons; 39 exact, 1 minor). This is partly the CFF specification working as designed, since preferred-citation exists precisely to route citation to a different object, usually a paper. The measured fact is about the consequence: for seven out of eight such projects, the record a compliant citation tool emits and the record the software's other surfaces describe are different objects, and which one a consumer lands on depends on tooling, not on the project's intent.

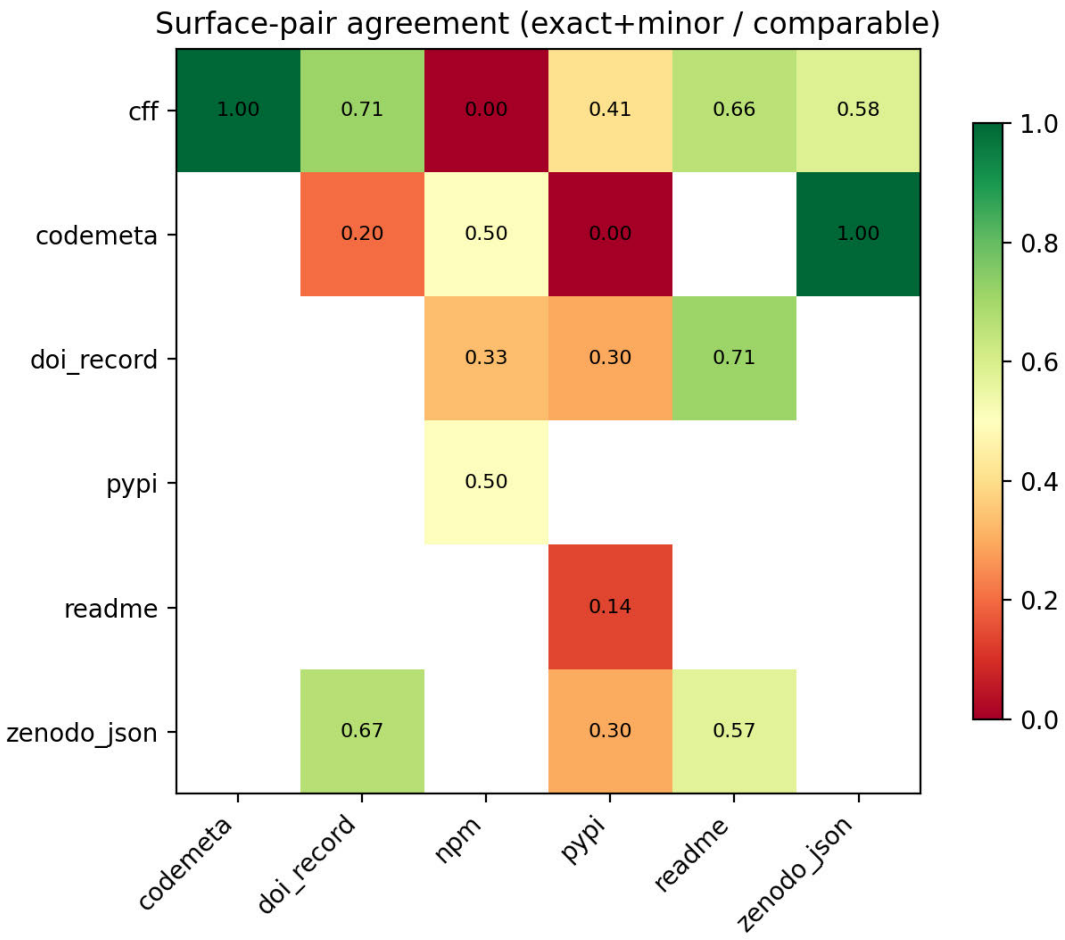


Figure 2: Pairwise agreement (exact plus minor over comparable) by surface pair. Cell denominators vary widely; extreme values, including 0.00 and 1.00, can rest on few comparable pairs, and per-cell counts are in the released results data; the corpus contains a single npm-listed project.

4.4 A mechanism taxonomy of verified conflicts

Hand adjudication of every conflict produced a mechanism classification; the four recurring mechanisms, with named, snapshot-dated examples (retrieved 2026-08-14):

Paper-versus-software conflation (86 of 171 cross-surface conflict judgments, 50%). Surfaces describe the software's paper rather than the software. Two sub-patterns. Divergent: dace's CITATION.cff root titles the software ("DaCe - Data Centric Parallel Programming") while its README BibTeX carries the conference paper's title and its five authors, against 28 in the CFF. *Consistent-but-paper-anchored:* thread-pool's CFF root itself declares the SoftwareX paper's DOI and title, so its surfaces agree with each other while all describing the paper; consistency is not correctness of referent.

Archive staleness. DOI'd archive records lag the distributed release: felupe's Zenodo record says v9.4.0 while PyPI serves 10.1.0; cuda-quantum shows the same mechanism at minimum distance (0.15.0 versus 0.15.1). The rubric counts each non-equal pair once regardless of distance.

Registry ghost metadata. Registry records exist that the project does not control or no longer describes: our review removed two back-link-verified packages (a community npm build and a third-party PyPI wrapper) as not project-controlled; registry author fields are frequently entity strings or empty, and no registry surface in the corpus carries an ORCID (0/25 sampled authors on PyPI).

Identity ambiguity. Entity and person names collide across surfaces: cirq's surfaces disagree only on "Cirq Developers" versus "The Cirq Developers"; mfem's CFF root credits "MFEM Team" while its archive record credits two named individuals and its preferred citation seventeen.

By design, the by-design routing of Section 4.3 (mfem-style preferred-citation) is tabulated separately from cross-surface conflation, since the former is specification-compliant behavior with a fragmenting consequence and the latter is unintended divergence.

4.5 ORCID coverage

Where surfaces carry authors, ORCID coverage varies from 62% of authors in .zenodo.json and 54% in CFF to 39% in resolved DOI records and 0% in registry records and README BibTeX, so the surfaces most likely to be read by packaging tooling carry the least persistent-identifier information.

5 Discussion

Three implications follow from the supply side alone. First, single-surface consumers inherit whichever record their surface carries; at the measured disagreement rates, two compliant consumers reading different surfaces of the same project will disagree about its citation roughly half the time at the field level. Second, the paper-versus-software mechanism dominating verified conflicts (50%) means the fragmentation is not primarily sloppiness: it is the unresolved question of what the citable object is, propagating into machine-readable metadata, consistent with the credit incentives documented in software citation research [4, 6], and it will not be fixed by adoption campaigns for any single format. Third, agreement infrastructure is absent: no existing supply-chain or metadata framework checks cross-surface consistency ([16–18] audit build and release integrity, not self-description coherence), and the near-universal surfaces (README, releases) are precisely the ones with the least structure.

The audit engine is released as an importable library whose per-project primitive (harvest, normalize, compare, report for one repository) is designed for downstream reuse in release-time integrity checks and citation and provenance checker tools; the corpus-scale study reported here is one instantiation. The demand side, how automated consumers weight these surfaces when they disagree, is measurable with the same instrument as ground truth and is deliberately out of scope here.

6 Limitations

The corpus is GitHub-hosted by criterion and skewed toward HPC/QC by design, with small baseline strata. README parsing is heuristic; verified precision (98.5%) bounds but does not eliminate instrument error, and the released log classifies every known disagreement. The registry surface covers PyPI and npm only, so R and Julia packages face a five-surface ceiling, and their registry practice is unmeasured. Registry name-versus-title conflicts are structural and reported both included and excluded. The eligibility filter conditions the baseline on having >= 2 surfaces, so baseline agreement rates describe metadata-active projects, not the registries at large. Snapshots date to a single retrieval window; projects change. Concept-versus-version DOI adjudication covered flagged Zenodo pairs only. Verdict counts weight all conflicts equally regardless of severity.

7 Conclusion

Measured across their own machine-readable surfaces, five of six research software projects that can be cross-checked disagree with themselves on at least one citation-critical field, and the single largest verified mechanism is the paper standing in for the software. The instrument, corpus, registered protocol, snapshots, and verification log are released so the measurement can be repeated, extended to other ecosystems, and used as ground truth for the demand-side question of what metadata consumers do when the surfaces disagree. Extending the corpus beyond GitHub-hosted projects and additional registries, and re-measuring the same projects over time and assessing mitigation in which a single authoritative record generates the other surfaces, is planned future work on the same instrument.

8 Data and Code Availability

Pipeline: https://github.com/pengyin-shan/rda-audit-pipeline (Apache-2.0). Measurements were produced with v0.2.2; the archived release v0.2.3 adds only license, citation, and README files, with measurement code identical

(10.5281/zenodo.21969695). Corpus, raw snapshots (retrieval-timestamped), results, verification log, and sensitivity outputs: *10.5281/zenodo.21969769*, CC-BY-4.0. Sampling protocol, ranked candidate lists, and exclusion log: *10.5281/zenodo.21909720* (restricted until publication; opened on posting); the same files are public in the sampler repository. Sampler: baseline-sampler v0.1.0 (registered draw commit 64b979e), https://github.com/pengyin-shan/baseline-sampler. Prior corpus: *10.5281/zenodo.21443211*.

9 Acknowledgments

The author thanks Daniel S. Katz for comments on a draft. Portions of the manuscript text were drafted with writing assistance from Claude (Anthropic); the author reviewed and verified all content, and all data collection, measurement, and verification were performed by the author. The corpus derives from work conducted during the author's 2025-2026 Trusted CI Fellowship with Trusted CI, the NSF Cybersecurity Center of Excellence; the fellowship provided no financial support for this work. The author thanks the National Center for Supercomputing Applications (NCSA), University of Illinois Urbana-Champaign, for a supportive professional environment; this work was conducted in the author's individual capacity and received no dedicated funding. The author declares no competing interests.

References

[1] Stephan Druskat, Jurriaan H. Spaaks, Neil Chue Hong, Robert Haines, James Baker, Spencer Bliven, Egon Willighagen, David Pérez-Suárez, and Alexander Konovalov. Citation file format, August 2021. URL `https://doi.org/10.5281/zenodo.5171937`.

[2] Matthew B. Jones, Carl Boettiger, Abby Cabunoc Mayes, Arfon Smith, Peter Slaughter, Kyle Niemeyer, Yolanda Gil, Martin Fenner, Krzysztof Nowak, Mark Hahnel, Luke Coy, Alice Allen, Mercè Crosas, Ashley Sands, Neil Chue Hong, Patricia Cruse, Daniel S. Katz, and Carole Goble. CodeMeta: an exchange schema for software metadata, 2017. URL `https://doi.org/10.5063/schema/codemeta-2.0`. Version 2.0.

[3] DataCite Metadata Working Group. DataCite metadata schema for the publication and citation of research data and other research outputs, v4.5, 2024. URL `https://doi.org/10.14454/g8e5-6293`.

[4] Arfon M. Smith, Daniel S. Katz, Kyle E. Niemeyer, and FORCE11 Software Citation Working Group. Software citation principles. *PeerJ Computer Science*, 2:e86, 2016. doi:10.7717/peerj-cs.86.

[5] James Howison and Julia Bullard. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, 67(9):2137–2155, 2016. doi:10.1002/asi.23538.

[6] Daniel S. Katz and Neil P. Chue Hong. Software citation in theory and practice. In James H. Davenport, Manuel Kauers, George Labahn, and Josef Urban, editors, *Mathematical Software – ICMS 2018*, volume 10931 of *Lecture Notes in Computer Science*, pages 289–296, Cham, 2018. Springer. doi:10.1007/978-3-319-96418-8_34.

[7] Pengyin Shan. Channel-based provenance signals for 87 HPC and quantum-computing open-source projects, 2026. URL `https://doi.org/10.5281/zenodo.21443211`. Version 1.0.0, dataset.

[8] Michelle Barker, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, Leyla Jael Castro, Morane Gruenpeter, Paula Andrea Martinez, and Tom Honeyman. Introducing the FAIR principles for research software. *Scientific Data*, 9(1):622, 2022. doi:10.1038/s41597-022-01710-x.

[9] Anas El Hounsri and Daniel Garijo. Good practice versus reality: A landscape analysis of research software metadata adoption in european open science clusters. In *2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR)*, pages 116–128, 2025. doi:10.1109/MSR66628.2025.00028.

[10] Ethan Bommarito and Michael J. Bommarito. An empirical analysis of the Python Package Index (PyPI), 2019.

[11] Kai Gao, Weiwei Xu, Wenhao Yang, and Minghui Zhou. PyRadar: Towards automatically retrieving and validating source code repository information for PyPI packages. *Proceedings of the ACM on Software Engineering*, 1(FSE): 2608–2631, 2024. doi:10.1145/3660822. Article 115.

[12] Stephan Druskat and Lars Grunske. OpenDORS: A dataset of openly referenced open research software, 2025. URL `https://doi.org/10.48550/arXiv.2512.01570`. arXiv:2512.01570.

[13] Daniel S. Katz, Kyle E. Niemeyer, and Arfon M. Smith. Publish your software: Introducing the journal of open source software (JOSS). *Computing in Science & Engineering*, 20(3):84–88, 2018. doi:10.1109/MCSE.2018.03221930.

[14] Caifan Du, Johanna Cohoon, Patrice Lopez, and James Howison. Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, 72(7):870–884, 2021. doi:10.1002/asi.24454.

[15] Caifan Du, Johanna Cohoon, Patrice Lopez, and James Howison. Understanding progress in software citation: a study of software citation in the CORD-19 corpus. *PeerJ Computer Science*, 8:e1022, 2022. doi:10.7717/peerj-cs.1022.

[16] Nusrat Zahan, Parth Kanakiya, Brian Hambleton, Shohanuzzaman Shohan, and Laurie Williams. OpenSSF scorecard: On the path toward ecosystem-wide automated security metrics. *IEEE Security & Privacy*, 21(6): 76–88, 2023. doi:10.1109/MSEC.2023.3279773.

[17] Mahzabin Tamanna, Sivana Hamer, Mindy Tran, Sascha Fahl, Yasemin Acar, and Laurie Williams. Analyzing challenges in deployment of the SLSA framework for software supply chain security, 2024. URL `https://doi.org/10.48550/arXiv.2409.05014`. arXiv:2409.05014.

[18] Zachary Newman, John Speed Meyers, and Santiago Torres-Arias. Sigstore: Software signing for everybody. In *Proc. 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2022. doi:10.1145/3548606.3560596.

215. UsAgainstAlzheimer's

I am responding: On behalf of an organization

Name: Erika Sward

Name of Organization: UsAgainstAlzheimer's

Type of Organization: Research Participant/Patient Advocacy Organization/Patient Advocate/Caregiver

Role: Research Participant/Patient/Advocate/Caregiver

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/UsAgainstAlzheimers_Comments_to_NIH-outcomes-RFI-8-19-26.pdf

Description: Please see attached comments from UsAgainstAlzheimer's

UsAgainstAlzheimer's

August 19, 2026

Lyric Jorgenson, PhD
Associate Director for Science Policy
Office of Science Policy
National Institutes of Health
Bethesda, MD

Re: Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Dear Dr. Jorgenson:

UsAgainstAlzheimer's is pleased to submit comments in response to the National Institutes of Health (NIH) request for information. UsAgainstAlzheimer's mission is to end Alzheimer's disease (AD). Our strategic priorities are to advance prevention and risk reduction; work to promote early detection and diagnosis; and ensure access to treatments and care. We are guided by the experiences of individuals living with AD and Alzheimer's disease related dementias (ADRDs, which include frontotemporal dementia, Lewy body dementia, vascular contributions to cognitive impairment and dementia, and rare, related dementias, e.g., MAPT-FTD and FTD-GRN), their families, and the communities most at risk.

Preventing and treating AD/ADRDs is an urgent national priority for American families and for the social, economic, and fiscal health of our nation — and it is within reach. Our nation is not prepared for the current toll of Alzheimer's — let alone its projected growth. The scale of this disease already exceeds what our care infrastructure was built for. An estimated 7.4 million Americans age 65 and older are living with AD in 2026, with about 1 in 9 people age 65 and older affected.[1] By 2050, the number of Americans living with AD is projected to rise to nearly 13 million.[2] Many Americans are at uniquely high risk: women are twice as likely to develop AD than men, African Americans are twice as likely and Latinos are 1.5 as likely to develop AD than white older Americans.[3] Rural adults have a higher burden of modifiable dementia risk factors.[4] And individuals with Down syndrome are also at extremely high risk, and often develop the disease at much younger ages.[5] The cost curve is just as alarming: health and long-term care costs for people living with AD and other dementias are projected to reach $409 billion in 2026 — not including the value of unpaid caregiving — and total payments are projected to climb to nearly $1 trillion by 2050.[6]

We commend the NIH for outlining goals for scientific excellence in its report, "Leading in Gold Standard Science." Many aspects of this publication, including the promotion of collaboration, data sharing, public access, and dissemination (among others) are consistent with tenets of reproducibility, transparency, communication, and avoidance of

conflict of interest – which we also endorse. We believe the research reward and impact assessment system must include these goals, but UsAgainstAlzheimer's is extremely concerned that the Gold Standard Science policy is inconsistent with other administration policies, including those advanced in a recent proposed rule from the Office of Management and Budget, "Regulation for Federal Financial Assistance." Ultimately, UsAgainstAlzheimer's urges that science and promise for breakthroughs be what drives investment, not politics.

The RFI states that "NIH seeks to better align its measures of success with its core mission; generating knowledge that enhances health, lengthens life, and reduces illness and disability," and it requests input on potential indicators and incentives in the following areas:

Rigor and Reproducibility

We agree with the NIH assertion that advancing foundational knowledge should explicitly prioritize rigor and generalizability as core scientific principles. Such priorities are particularly urgent for research on AD/ADRD, where gaps in representativeness have limited progress and the generalizability of findings. Several populations that are at increased risk are noted above; ensuring that research considers the variability across different groups (e.g., genetics, exposome, geography, age) from the outset will strengthen the external validity and translational impact of research on AD/ADRD and other conditions.

Similarly, strengthening rigor and reproducibility also requires heightened emphasis on implementation and real-world translation. It is essential to ensure that discoveries from the laboratory are moved into clinical practice and community settings, particularly in very high therapeutic need fields such as AD/ADRD. NIH must continue to expand support for pragmatic trials, community-based research, and public-private partnerships, as well as collaboration with other agencies, e.g., the Centers for Disease Control and Prevention (CDC), to ensure scientific advances translate into meaningful, generalizable improvements in health outcomes for women and men, and in diverse populations.

Along these lines, while AD/ADRD research has a long history of testing and retesting hypotheses (e.g., the amyloid hypothesis) before translation into rigorous clinical trials, some areas of AD/ADRD research – e.g., research on caregiving interventions – do not easily lend themselves to assessment in typical double-blind randomized clinical trials. NIH's sister agency – the Agency for Healthcare Research and Quality (AHRQ) – has assessed how other research approaches, including quasi-experimental designs, might be weighed when considering a body of research evidence for translation, and we would urge NIH leadership to consider working with AHRQ to investigate options for reinforcing rigor across these fields, while also fostering collaborative discussion about where and when specific design and statistical approaches are appropriate to incentivize.

Data, Software, and Model Sharing

NIH must continue to build and modernize research infrastructure, emphasizing data use and protections. Expanding existing platforms to capture longitudinal, harmonized, and interoperable datasets – and ensure that sex differences (noting the higher burden of AD/ADRD on women) are built in at both the design and analysis end – is essential to advancing science; data standards and sharing requirements must be consistently enforced to maximize the value of NIH-funded research and enable collaboration.

Ensuring that clinical trial data are relevant to all Americans should also be an NIH priority. Clinical trial recruitment and monitoring efforts must be prioritized, made more efficient, and adequately funded. NIH has launched hundreds of AD/ADRD clinical trials in the past decade, but clinical trial recruitment remains a challenge across populations, and it's imperative that everyone has access to participation. Several sub-groups have a differential risk to AD, as noted above: women, African Americans, Latinos, individuals living rurally, and individuals living with Down syndrome are all at higher risk. Recruitment and the development and maintenance of related infrastructure that ensures broad inclusion of these groups lead to non-trivial costs, and investments to support these structures and actions must be prioritized.

As just one example, most trials are implemented with traditional cognitive assessment measures that rely on trained human raters to evaluate patients. The availability, retention, training costs and variability of raters all pose challenges – and must be considered when budgets are developed. Monitoring and enforcing recruitment from the federal government is also key to success, and the National Institute on Aging's (NIA's) tracking has only relatively recently been modernized (via its Clinical Research Operations & Management System). The success of this tracking and the enforcement opportunities it provides are key to the inclusion of appropriate gender and population representation in NIH-funded AD/ADRD trials. It is unclear what the status of this effort is, but future success will require an investment both in programs such as CROMS that support data collection on recruitment, and in programmatic staff capable of utilizing the data generated to ensure compliance by investigators.

Lastly, investments in clinical trials are impacted by the overall destabilization of federal research funding that is occurring, in ways that seriously hamper progress in AD/ADRD research: researchers will not be able to secure the funding they need for the trials, including the costs of adequate recruitment efforts, and participants will be apprehensive about joining a trial, at the risk that the trial might be de-funded mid-way through the research.

Disruptions to ongoing studies also have consequences that extend well beyond financial costs. Prematurely interrupting or terminating research may result in the loss of valuable scientific data and waste significant public investments in grant funding and institutional infrastructure. Of utmost importance, such disruptions may also adversely affect study

participants and their care partners, who often rely on research participation not only for access to investigational therapies or non-pharmacological interventions, but also for specialized clinical monitoring, care coordination, cognitive assessments, and other health and support services provided through the study. It is particularly dangerous to the health and safety of patients who may face sudden and unplanned disruption or discontinuation of medical interventions where their medical team has no time and resources, and therefore no ability, to manage a medically prudent transition (e.g., tapering off of a medicine, transitioning to a new provider).

In the case of AD/ADRD research, a number of current drug treatments in clinical trials have potentially serious side effects such as amyloid-related imaging abnormalities (ARIA) that require careful, consistent, and complex safety monitoring. Abruptly suspending or terminating grant funding for such a trial literally risks serious avoidable harm and potential death for study participants – in addition to wasting data and financial resources already invested.

NIH should also invest in emerging technologies that expand research capacity and accelerate discovery. Innovations including but not limited to CRISPR and AI-driven analytics hold significant promise for identifying disease mechanisms and therapeutic targets. Research tools such as digital cognitive assessments offer the potential for earlier, longitudinal, more sensitive, and more efficient measurement of cognitive change in diseases like AD/ADRD, while also reducing trial burden and improving participant experience. As novel therapies move earlier in the disease course, increased sensitivity of clinical measures will be indispensable to demonstrate efficacy. NIH's leadership in funding and coordinating validation of these resources would help resolve a key research bottleneck.

One challenge in the realm of optimizing large datasets that is somewhat unique to the field of AD/ADRD research is the need to study age-related cognitive changes as an add-on or ancillary activity to a clinical trial for which the initial outcomes were not cognitive (e.g., cardiovascular, endocrine). Sustained and reliable funding streams are necessary to ensure that investments made in the “base” trials are not then wasted in the realm of AD/ADRD, if a study is de-funded, or funding is reduced, such that the data collection needed at distal study endpoints to assess cognition (often imaging) is not feasible.

Lastly, but critically, there is a significant need for optimizing data and datasets in the realm of women’s brain health research. Currently, NIH invests only 3% of its AD/ADRD funding on work that is relevant to women’s health (based on FY2024 RePORTER data). Women make up two-thirds of the people living with AD/ADRD, and research over the last decade has highlighted that women experience a range of unique and differential life exposures, as well as sex-based differences in cardiovascular, immune, endocrine systems, etc. – all of these influence AD/ADRD risk and response to interventions.

NIH should prioritize collaborations across women's health and AD/ADRD research teams that enable women's health-related data to be collected in AD/ADRD studies, and should consider incentivizing prospective high value datasets (that would capture life course data following risks and exposures of women, as well as outcomes across a wide range of systems, including endocrine and cognitive). It is only through significant incentivization and investment at the level of the NIH Office of the Director – as well as a sustained commitment to enforce NIH's own 2016 Sex as a Biological Variable (SABV) policy – that we will yield substantive progress in this area.

Training and Mentorship

NIH must prioritize workforce development as a core strategic investment. NIH's shift toward parent awards and multi-year funding models have devastated the training community; greater attention is needed to ensure investigator-focused training mechanisms are not overlooked in future strategic planning efforts.

NIH's commitment to mentorship is vital and should be a formal, valued component of the research enterprise. NIH must ensure adequate salary support and protected time to enable established researchers – including both mid-career and established investigators – to nurture the next generation of researchers. In parallel, small but meaningful investments in training programs (e.g., increasing stipends) improve trainee recruitment and retention, particularly those from underserved backgrounds and in fields facing workforce shortages (e.g., geriatrics). These investments demonstrate the value placed on training and help reduce financial barriers disproportionately affecting individuals entering biomedical research careers.

Emerging scientific challenges require interdisciplinary expertise; accordingly, the strategic plan should also prioritize interdisciplinary training and support career transitions across fields. A strong research workforce requires investing in scientists from underserved and underrepresented backgrounds, whose perspectives and experiences strengthen scientific rigor, innovation, and research relevance to broader populations. To move science toward a precision medicine model, NIH must restore funding mechanisms (e.g., diversity supplements) to support these researchers.

Collaboration

Recognizing contributions to team-based science has been a longstanding problem across biomedical research, and fields like AD/ADRD research demand an increasingly complex array of contributing disciplines – including geneticists, molecular biologists, clinicians, statisticians, data scientists, imaging experts, and much, much more. The contributions of these individuals and their expertise is critical to the progress of science, yet each individual contributor must weigh their interest in and commitment to scientific discovery with their goals for academic stability and advancement within their organization. Progress is already underway at the NIH to provide appropriate incentives to team science, including

but not limited to the multiple-PI option for grant applicants, but it should continue to make novel funding opportunities available to collaborative teams and supporting roles that facilitate team science; essentially, those that would be recognized by traditional academic review committees. It is notable, though, that the many forces at play within the current NIH landscape that are destabilizing research funding are not likely to incentivize research collaborations.

Entrepreneurship and Translation

We agree that innovation and entrepreneurship should be fostered by the NIH, towards preventing and treating conditions like AD/ADRD. The expansion of NIA's small business and technology programs as AD/ADRD funding has increased – as well as the development of challenge prize programs at the NIA – have demonstrated the contributions that are possible with these investments. We urge the NIH to continue to incentivize these programs, and to think more broadly about how more entrepreneurial approaches might be applied to unique and innovative directions of disease research in the future, at the basic, translational, and clinical realms. A specific goal could be the launch of another EUREKA (Ensuring Useful Research Expenditures is Key for Alzheimer's) prize, targeted to AD/ADRD research.

Foundational Scientific Exploration

We agree that both basic and high-risk/high-reward research should be given fair consideration within the overall research realm. For AD/ADRD research, we also encourage continued diversification of the target pipeline, both for treatment and disease prevention, and we hope NIH will closely monitor emerging lines of research that may yield innovative strategies, such as the potential of shingles and other safe and effective vaccines to protect against development of AD/cognitive impairment, as studies designed to validate this possibility would be a meaningful contribution to the field.[7,8] Similarly, one of the challenges in AD/ADRD research is predicting the disease trajectory in different individuals – discovery in this area would be very high reward for both healthcare providers as well as those designing clinical trials. Validating models to predict which patients are likely to progress will be a major support for the field.[9,10]

Many other aspects of AD/ADRD research play a critical role in ensuring that the investments in NIH yield optimal impact – including caregiving research and studies that explore interventions for AD-related neuropsychiatric symptoms (e.g., agitation and psychosis). While outcomes in these areas are incredibly meaningful for disease management, such studies might not be considered "high risk." Nevertheless, the payoff when interventions prove effective, as was the case recently with the NIH-funded LiBBY study, is immense for families, caregivers, and healthcare providers. In addition to addressing agitation and other late-stage symptoms of AD/ADRD, we also encourage the NIH to consider outcomes related to hospice, bereavement, and end of life decision-making as important and relevant in the overall spectrum of disease research.

At the same time, we need better data on how treatment, care planning, and family burden differ when a diagnosis involves mixed etiology or a related dementia – rather than "pure" AD, since these conditions often carry different symptom trajectories, caregiver burden profiles, and treatment eligibility (including for emerging AD-specific therapies), yet families frequently receive the same generic guidance regardless of the underlying pathology.

Public Impact

The RFI lists many contributions to public impact that may be considered, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.

With respect to ***health outcomes***, NIH AD/ADRD research investments have already borne fruit in the pharmacological sector, with two relatively recent drug approvals having been made by the Food and Drug Administration (FDA) for early-stage disease. Prevention and risk reduction for AD has also advanced (including both pharmacological and non-pharmacological approaches) and the means to achieve it at population scale are either known or in development: proven life-course risk-reduction strategies, active and passive immunotherapies, and other emerging modalities, including vaccines aimed at reducing AD/ADRD risk. Even more effective solutions can emerge from the world class research laboratories across our country. NIH must prioritize prevention in addition to treatment research, and it should expand these studies, including treatments and risk reduction strategies applicable to the preclinical phase of disease and combinations of interventions that include both pharmacological and non-pharmacological approaches.

Patient preferences should also be a high priority for the NIH when considering what constitutes “success” with respect to health outcomes. UsAgainstAlzheimer’s Alzheimer’s Disease Patient and Caregiver Engagement (AD PACE) Initiative is designed to address this need, i.e., to identify the outcomes and care goals that are most meaningful to those living with AD/ADRD and their families, caregivers, and care partners. The focus of the What Matters Most (WMM) Research Program of AD PACE has collected data on over 600 people to improve our understanding and measurement of these outcomes. The next stage of this research is ongoing, and adds in the much-needed focus on aiding FDA and companies to set clinical outcome measures related to WMM to families and on enhancing communication between clinicians and patients about the disease and potential outcomes, to ensure ongoing care is appropriately informed by the outcomes and care goals found to be most meaningful – literally “What Matters Most” – to those affected.

Through the last 15 years, drug development has particularly focused on the earliest stages of disease, when those with the diagnosis can report for themselves about the outcomes and care goals that are most important to them. The connection to the FDA Patient-Focused Drug Development optimizes the investment in both initiatives (FDA Patient-

Focused Drug Development Guidance Series for Enhancing the Incorporation of the Patient's Voice in Medical Product Development and Regulatory Decision Making | FDA).

This Initiative is contributing detailed information about specific areas useful for measuring the impact of treatments as well. As patients and payment-related organizations address the reimbursement challenges created by the growing arsenal of disease treatments, more detailed data about outcomes important to those affected are increasingly needed. These are the data that the AD PACE Initiative generates, and NIH's future planning efforts should address the ongoing need for longitudinal data that focuses on patient and carer perspectives. Connecting patient-centered data to ongoing federally funded initiatives, like the NIH-funded Alzheimer's Disease Neuroimaging Initiative (ADNI), is another fruitful area for emphasis in the coming years. Maintaining funding for these initiatives is therefore even more important as the views of people with lived experience are increasingly incorporated into the work. AD PACE is a model for working directly with people with lived experience to better understand their priorities across the disease spectrum AD PACE is also a model appropriate for other dementias. We would encourage the NIH to utilize our data to ensure the recognition of patient priorities as it incentivizes and prioritizes specific research areas.

Within the ***economic growth*** realm, there continues to be a significant lack of caregiver data — particularly economic data needed to support reimbursement decisions. Caregivers are the backbone of the current system, yet the data needed to value, support, and reimburse their work remain fragmented and inconsistent. More than 12 million family members and other unpaid caregivers provided an estimated 19.6 billion hours of care to people living with AD/ADRD.[11] The most recent AARP analysis found that the average value of one caregiving hour rose from $16.59 to $20.41 in 2024[12]— yet estimates like this vary widely depending on methodology, and there is no standardized, government-collected dataset that payers can rely on consistently. This is not just an academic problem: a peer-reviewed analysis of the indirect costs AD found explicitly that there are currently no publicly available estimates of certain affected populations, and that data gaps and a lack of consensus persist concerning how to value caregiver forgone earnings and disutility.[13] Without consistent, high-quality economic data on caregiving — hours provided, income and career impact, health costs borne by caregivers themselves — payers cannot make sound reimbursement decisions, whether through Medicaid Home- and Community-based Services (HCBS) waivers, paid family leave, tax credits, or emerging CMS models like GUIDE. We are essentially asking decision-makers to build reimbursement decisions on incomplete evidence.

The recent FDA approvals of lecanemab and donanemab have advanced the offerings in ***clinical practice*** for AD/ADRD, but the diagnostic challenges surrounding AD/ADRD have long been highly complex. The NIA contributed a number of years ago to the development of research criteria for AD/ADRD[14], but has since shifted away from those activities as they have morphed into guidance that is increasingly comparable to clinical practice guidelines. Nevertheless, NIA (and the NIH) do have a role in contributing the research that informs practice, and continued investments in understanding the basic science of the disease, as

well as its biomarkers and manifestations that can lead to more accurate diagnoses, will be important, as well contributions to clinical practice in the area of care for those living with advanced AD/ADRD.

Fostering public trust in AD/ADRD research, as well as ***accessibility*** of these findings, is essential – but currently an immense obstacle, as the NIH website (and its institute and center sites) have been trusted public health information resources for decades. Within the AD/ADRD realm, the science is evolving very quickly, and messaging needs to not only reflect advances in drug development but also advances in behavior change and modifiable risk factors (such as blood pressure control and several additional behavior change strategies). NIH must ensure timely, transparent, and scientifically grounded decision-making that supports both the research community and the public. Effective stewardship depends on clear communication channels between NIH leadership and staff with the extramural research community and the public, as well as consistent and efficient funding processes that ensure that federal commitments to research awards are honored.

The reduction and centralization of public-facing functions across ICs, including but not limited to communications, legislative affairs, policy, and evaluation offices is deeply troubling. While central coordination may appear to improve consistency, the elimination and consolidation of these functions compounded by understaffing and loss of expertise has threatened information accuracy, reduced visibility of NIH-supported science, and limited direct engagement with the public and scientific stakeholders. This is particularly consequential given the scale of NIH's mission; no longer are there dedicated channels to the community on specific health conditions such as AD/ADRD, common comorbid conditions, and actionable risk reduction evidence. For example, NIA previously reached more than 125,000 subscribers through weekly e-alerts, and its social media platforms engaged more than 187,000 followers (*HHS National Plan to Address Alzheimer's Disease: 2024 Update*). The absence of NIA communications staff is calamitous for the AD community, as their outreach channels were essential tools for disseminating research advances, particularly at a time when any new intervention must be communicated quickly and accurately with the public. These channels should be strategically reestablished in the near future.

In addition, a lack of transparency has emerged in the last year between program officers and grantees, one of the most essential relationships for NIH. Historically, program officials played a critical role in identifying emerging scientific priorities and rapidly translating them into funding opportunities and were able to share research priorities early with the scientific community. Unfortunately, since 2025, changes to NIH structures and systems have dramatically slowed and undermined scientific opportunities. For example, while some reductions and consolidations in the number of Notices of Funding Opportunities (NOFOs) in redundancy are important, the dramatic decrease has only served to limit the ability of program staff to fill pivotal gaps in highly specialized research fields and disincentivize active communications with prospective applicants and current grantees.

Lastly, ***societal benefit*** should continue to be first and foremost in the minds of NIH leaders as research areas are considered and prioritized. And as such, discovery must benefit women and men equally, as well as diverse, at-risk populations, and NIH should seek to ensure that research portfolios are continually assessed with this goal in mind.

NIH should also prioritize life course considerations with respect to disease research. Within the AD/ADRD research realm, we must better understand the impact of Alzheimer's disease across the lifespan – not just for the person diagnosed, but for the generation(s) caring for them. Emerging research on caregivers for people living with AD/ADRD causes chronic stress, disrupted sleep, and immune dysregulation, and studies increasingly point to a bidirectional relationship in which the strain of caregiving may itself accelerate biological aging and elevate the caregiver's own future dementia risk.[15]

This matters enormously for genetically predisposed adult children and grandchildren – those carrying APOE-e4 or other risk variants –who are simultaneously caring for a parent or grandparent with AD and, some research suggests, may be experiencing years of chronic cortisol exposure and inflammatory stress that could interact with that genetic vulnerability to hasten their own cognitive decline. Much of this evidence is still early and associative rather than causal, but it raises urgent questions about whether we are inadvertently watching the disease propagate across generations through the caregiving experience itself, not just through inherited genes.

Further, societal benefit cannot be fully considered by NIH in a vacuum; collaborations must be fostered with sister agencies under the Department of Health and Human Services (HHS), even when they are not defined in statute. A particular between-agency challenge that NIH leadership could begin to rectify (and which is meaningful to outcomes and "benefit") is related to how agencies under HHS define "evidence," especially in the non-pharmacological realm, that supports broad implementation. For example, within HHS, NIH and the ACL contribute complementary perspectives on what constitutes evidence for improving dementia care. NIH's mission emphasizes generating rigorous, generalizable scientific evidence through clinical trials (particularly randomized clinical trials), observational studies, and systematic evidence reviews to determine whether interventions are effective, for whom they work, and under what conditions. ACL builds on this evidence while also valuing practice-based evidence generated through community programs, service delivery, program evaluation, quality improvement, and the lived experiences of people living with dementia, care partners, and service providers to assess whether interventions are feasible in real-world settings. While these perspectives are complementary, the absence of a shared framework for determining when an intervention has accumulated sufficient evidence to move from research to broad implementation can slow dissemination and implementation, creating uncertainty for researchers, health systems, and community organizations.

In closing, the NIH is well-positioned to continue to make unprecedented progress in AD/ADRD research, but the considerations above are critical to the prioritization and

incentivization of research, and to true “gold standard” science. Thank you for the opportunity to submit these comments. We look forward to working with the NIH to advance our shared priorities and end this disease.

Sincerely,



Catherine Patterson
Acting Chief Operating Officer

---

[1] Alzheimer's Association, *2026 Alzheimer's Disease Facts and Figures*, https://www.alz.org/alzheimers-dementia/facts-figures
[2] USC Schaeffer Center, "Dementia Will Cost the U.S. $818 Billion in 2026, USC-Led Study Finds," June 24, 2026, https://schaeffer.usc.edu/research/dementia-cost-us-818-billion-2026-study/
[3] Alzheimer's Association. (2026). *2026 Alzheimer's disease facts and figures*. https://www.alz.org/getmedia/ef8f48f9-ad36-48ea-87f9-b74034635c1e/alzheimers-facts-and-figures.pdf
[4] Xie Z, Hu J, Stallings-Smith S, et al. Rural-urban differences in modifiable dementia risk factors among U.S. populations aged 45 years or older. *J Alzheimers Dis Rep*. 2025 Nov 6;9:25424823251395318.
[5] Rubenstein E, Tewolde S, Michals A et al. Alzheimer Dementia Among Individuals With Down Syndrome. *JAMA Netw Open*. 2024;7;(9):e2435018.
[6] Alzheimer's Association, *2026 Alzheimer's Disease Facts and Figures*, https://www.alz.org/alzheimers-dementia/facts-figures
[7]Geldsetzer, P. Why large-scale randomized trials of live-attenuated shingles vaccination for dementia prevention are urgently needed. *Nat Med* (2026).
[8]Maggi S, Fulop T, De Vita E, et al. Association between vaccinations and risk of dementia: a systematic review and meta-analysis. *Age Ageing*. 2025;54(11).
[9] Petersen KK, Mila-Aloma M, Li Y, et al. Predicting onset of symptomatic Alzheimer's disease with plasma p-tau217 clocks. *Nat Med*. 2026;32:1085-1094.
[10] Buckley RF, Townsend DL, Birkenbihl CJ, et al. Prognostic value of blood-based p-tau217 levels for progression to cognitive impairment. *JAMA*. Published online July 14, 2026.
[11] Alzheimer's Association, *2026 Alzheimer's Disease Facts and Figures*, Wiley Online Library, https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.71345
[12] AARP Public Policy Institute, *Valuing the Invaluable 2026: Family Caregivers' Contribution Reaches $1 Trillion*, March 26, 2026, https://www.aarp.org/caregiving/financial-legal/valuing-the-invaluable-report-2026/
[13] Indirect Costs of Alzheimer's Disease: Unpaid Caregiver Burden and Patient Productivity Loss," *Value in Health, ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S1098301524067597

[14] Jack, CR, Bennett, DA, Blennow, K et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement*. 2018 Apr;14(4):535–562.
[15] "Stress, Stress Management, and Dementia: A Narrative Review," National Center for Biotechnology Information (NCBI), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12604644/

216. National Society for Histotechnology

I am responding: On behalf of an organization

Name: Aubrey MJ Wanner

Name of Organization: National Society for Histotechnology

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Association Professional

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/NOT-OD-26-087_-NSH-Response.pdf

NATIONAL SOCIETY FOR HISTOTECHNOLOGY

3545 Ellicott Mills Drive, PMB 204
Ellicott City, MD 20143
P: 443-535-4060
E: histo@nsh.org

Response to NIH RFI NOT-OD-26-087
NIH Office of Science Policy
SciencePolicy@od.nih.gov

Submitted by: National Society for Histotechnology
Date: August 18, 2026

The National Society for Histotechnology (NSH) appreciates the opportunity to provide feedback on NIH's Request for Information regarding the measurement and rewarding of scientific impact. As the leading professional society representing histologists and other tissue-based laboratory professionals, NSH strongly supports NIH's efforts to modernize scientific impact assessment and align incentives with the practices that most effectively advance biomedical discovery and improve public health.

We encourage NIH to expand scientific impact metrics beyond traditional publication and citation measures to more fully recognize contributions that enhance reproducibility, accelerate translation, strengthen collaboration, and improve research quality.

Reproducibility Should Be a Core Measure of Scientific Impact

NSH believes that reproducibility and replicability should be elevated as primary measures of scientific impact. Novel findings have limited long-term value if other investigators cannot evaluate, reproduce, validate, or build upon them. Reproducible research creates a dependable foundation for subsequent discovery, while independent replication tests whether findings remain consistent when examined with new data or in different settings. Both activities strengthen confidence in scientific knowledge and should be recognized as substantive scientific contributions, not treated as secondary quality-control activities.

The importance of methodological transparency is particularly evident in tissue-based biomedical research [1]. In *Guide for collecting and reporting metadata on protocol variables and parameters from slide-based histotechnology assays to enhance reproducibility*, Chiriboga and colleagues identify major categories in which deficiencies can undermine reproducibility. They further note that insufficient reporting of reagents, protocols, and other experimental variables can prevent researchers from replicating or building upon published findings [2]. The recognition of this issue is not limited to the scientific histology community, the lack of rigor and reproducibility has been noted by others and a number of tools to identify materials (https://rrid.site/), laboratories (https://coremarketplace.org/), and researchers (https://orcid.org/) have been successfully implemented in the last several years [3-6]. These tools should play a larger role as metrics of reproducible research practices, with stricter requirements for their use in publications, grant applications, and evaluations of research laboratory performance [7-10].

Real-world histology research illustrates why reproducibility must be evaluated across laboratories rather than assumed from performance at a single site. In *The Impact of Stain Variation and Colour Normalisation on Tissue Stain Intensity and Automated Nuclear Counting: A Multi-Site Study*, investigators found statistically significant differences among laboratories, with greater variation between sites even though intra-laboratory consistency was relatively high [11] . In fact, instrument reproducibility is also subject to significant variation[12]. These examples demonstrate that reproducibility requires attention not only at the analytical steps but real-world performance across

NATIONAL SOCIETY FOR HISTOTECHNOLOGY

3545 Ellicott Mills Drive, PMB 204
Ellicott City, MD 20143
P: 443-535-4060
E: histo@nsh.org

institutions. This variation is a significant barrier to the implementation of computational models [11, 13, 14].

Reproducibility is essential in immunohistochemistry (IHC). In *Immunohistochemistry as an assay*, authors emphasize that IHC is a complex assay requiring analytical standardization and rigorous quality-control standards [15, 16]. This includes implementation of newly developed quantitative standards for statistical process control in immunohistochemistry [17, 18]. Immunohistochemical assay standardization is all the more important in the era of companion diagnostics as lack of rigor can have profound effects on outcomes [19-21]. Establishing best practices with reproducible standards is critical for rigorous scientific investigation [22-24]. Recognizing the scientific contributions of histology scientists would help correct an impact-assessment system that credits the final intellectual interpretation while overlooking the expertise that made a reliable and reproducible test result possible.

In emerging fields such as spatial biology, reproducibility also depends on the development of community standards and best practices [25]. The 2025 article, *Histotechnology at the crossroad of spatial biology bridging legacy and innovation,* observes that spatial-omics currently lacks fully standardized nomenclature, frameworks, and protocols [26]. New technology is vulnerable to artifacts because their effects on outcomes are often unknown. Moreover, when such technology is built on techniques or methodologies with a history of irreproducibility, it poses a significant risk. They and others identify actionable best practices to “standardize workflows, enhance reproducibility, and address logistical challenges [27-29]. Scientific impact evaluation should therefore recognize investigators, technical professionals, societies, and consortia that develop and maintain the shared technological standards underpinning reproducibility, even when those contributions are not fully captured by first-author publications or citation counts.

Investigators and scientists should be recognized and rewarded for following standardized workflows that enhance reproducibility. Likewise, those who develop these workflows should receive clear recognition, including citation, so that existing and emerging tracking tools can capture relevant metrics and help measure scientific reproducibility. NIH should therefore consider rewarding investigators who:

- Develop and publish validated protocols or workflows and participate in replication studies.
- Sharing of well-documented data, code, computational workflows, metadata, and analytical decision rules when ethically and legally permissible.
- Creation of shared reference datasets, control materials, repositories, proficiency-testing resources, and other infrastructure that allows reproducibility to be evaluated across institutions
- Contribute to standards development and demonstrating adherence to best-practice reporting frameworks.

Collaboration as a Measure of Scientific Impact

NSH strongly encourages NIH to recognize and reward scientific collaboration as a meaningful indicator of scientific impact. Many of today's most significant advances in biomedical research are the product of multidisciplinary teams that bring together complementary expertise [30, 31]. This includes the core facilities which exemplify collaborative science [32, 33]

NATIONAL SOCIETY FOR HISTOTECHNOLOGY

3545 Ellicott Mills Drive, PMB 204
Ellicott City, MD 20143
P: 443-535-4060
E: histo@nsh.org

Scientific progress increasingly depends on collaborative models of research. In a multi-industry consortium focused on spatial biology for drug discovery and biomarker development, the authors concluded that: *"Key to the success of these studies is the collaboration among experts from various fields, including molecular biology, pathology, histotechnology, imaging, and data science."* They further noted that *"This multidisciplinary team is essential not only in the design and execution of the experiment but also in the interpretation of the results."* and that *"Early discussions among these experts are vital"* for assessing technical feasibility and ensuring reliable data generation and analysis [27]. Collaboration is a key driver of reproducibility. Scientific teams that include subject matter experts across multiple disciplines are better positioned to identify methodological weaknesses, establish consensus standards, and generate reproducible results.

Professional societies play an important role as infrastructure for collaboration. The spatial biology editorial specifically identifies NSH and its partnerships with organizations such as the College of American Pathologists (CAP), American Society for Clinical Pathology, The Human BioMolecular Atlas Program (HuBMAP), the Human Tumor Atlas Network (HTAN), The Celluar Senescence Network (SenNet), and industry stakeholders as mechanisms for advancing education, workforce development, quality systems, and scientific innovation.

For these reasons, NSH recommends that NIH expand scientific impact assessments beyond traditional publication and citation metrics to recognize meaningful contributions to team science. Activities such as developing consensus standards, participating in scientific consortia, creating shared resources, establishing reproducibility frameworks, mentoring across disciplines, and fostering cross-sector collaborations generate substantial scientific value and accelerate translation fundamentally supporting the biomedical enterprise. As above, recognizing and rewarding these contributions would better align incentives with the collaborative nature of modern biomedical research and ultimately advance the NIH mission

Data, Software, and Model Sharing

Histology scientists manage large volumes of samples that exist in multiple forms, including human, animal, and model organism specimens; fresh, frozen, and archival tissues; images; plus, all the materials required to process these different samples on demand [34]. These specimens contain a wealth of scientific (-omic) information and hold significant research value [35, 36]. Because histologists have deep expertise in how tissue samples are collected, processed, managed, and stored, they are uniquely positioned to guide their use as data resources [36, 37]. Biorepository histology laboratories play an instrumental role in stewarding this powerful source of data [32, 38]. However, many histology biorepositories and core laboratories are understaffed and struggle to meet growing demand because of lack of recognition that stems directly from the lack of quantifiable metrics of the contributions of these facilities to the biomedical research enterprise [39-41].

Training and Mentorship

Core histology laboratories not only support the research enterprise but also provide training to graduate students, residents, fellows, and scientists for all critical steps of tissue processing from preanalytical to post analytical analysis for humans, animals, and model systems, including patient-derived xenografts, xenotransplants, and organoids, all collected for research purposes [42]. This

NATIONAL SOCIETY FOR HISTOTECHNOLOGY

3545 Ellicott Mills Drive, PMB 204
Ellicott City, MD 20143
P: 443-535-4060
E: histo@nsh.org

training is essential to ensuring rigor and reproducibility [42]. However, it often goes unnoticed or unrecognized in publications and grant applications because of its informal nature [39, 41].

Public Impact

The lack of recognition for the contributions of histology scientists has significantly contributed to workforce shortages, affecting not only patient-directed clinical histology laboratories but also core laboratories, where this expertise is in high demand [43]. Histologists play a vital role in both basic academic research and clinical practice. Together, these areas represent a substantial economic resource and make important contributions to the biomedical research enterprise. Greater recognition and visibility would help strengthen public trust and support improved health outcomes [44, 45].

Conclusion

Biomedical research is increasingly collaborative, data-intensive, and translational. As NIH modernizes how scientific impact is assessed, it has an important opportunity to reward the behaviors that produce trustworthy science and meaningful improvements in human health.

NSH encourages NIH to place reproducibility and collaboration at the center of scientific impact evaluation. Researchers who establish standards, share resources, validate methods, participate in replication efforts, mentor colleagues, and build collaborative networks contribute substantially to scientific progress.

These activities strengthen public trust, accelerate translation, improve return on federal investment, and ultimately advance the NIH mission of improving health and reducing disease.

References

1. Ioannidis, J.P., *Why most published research findings are false.* PLoS Med, 2005. **2**(8): p. e124.
2. Chiriboga, L., et al., *Guide for collecting and reporting metadata on protocol variables and parameters from slide-based histotechnology assays to enhance reproducibility.* J Histotechnol, 2022. **45**(4): p. 132-147.
3. Babic, Z., et al., *Incidences of problematic cell lines are lower in papers that use RRIDs to identify cell lines.* Elife, 2019. **8**.
4. Bandrowski, A., *A decade of GigaScience: What can be learned from half a million RRIDs in the scientific literature?* Gigascience, 2022. **11**.
5. Bandrowski, A., et al., *The Resource Identification Initiative: A cultural shift in publishing.* F1000Res, 2015. **4**: p. 134.
6. Bandrowski, A., et al., *The Antibody Registry: ten years of registering antibodies.* Nucleic Acids Res, 2023. **51**(D1): p. D358-d367.
7. Hofseth, L.J., *Getting rigorous with scientific rigor.* Carcinogenesis, 2018. **39**(1): p. 21-25.
8. Menke, J., et al., *Establishing Institutional Scores With the Rigor and Transparency Index: Large-scale Analysis of Scientific Reporting Quality.* J Med Internet Res, 2022. **24**(6): p. e37324.
9. Menke, J., et al., *The Rigor and Transparency Index Quality Metric for Assessing Biological and Medical Science Methods.* iScience, 2020. **23**(11): p. 101698.

NATIONAL SOCIETY FOR HISTOTECHNOLOGY

3545 Ellicott Mills Drive, PMB 204
Ellicott City, MD 20143
P: 443-535-4060
E: histo@nsh.org

10. Serghiou, S., et al., *Assessment of transparency indicators across the biomedical literature: How open is open?* PLoS Biol, 2021. **19**(3): p. e3001107.
11. Dunn, C., et al., *The impact of stain variation and colour normalisation on tissue stain intensity and automated nuclear counting: a multi-site study.* J Histotechnol, 2026: p. 1-16.
12. Pye, H., et al., *A look at scanner introduced variation in contrast, resolution, and colour across nine different models of whole slide imaging (WSI) scanner.* J Histotechnol, 2026: p. 1-12.
13. Chen, R.J., et al., *Towards a general-purpose foundation model for computational pathology.* Nat Med, 2024.
14. Schoenpflug, L.A., et al., *A protocol for evaluating robustness to H&E staining variation in computational pathology models.* Journal of Pathology Informatics, 2026: p. 100702.
15. Bell, M., et al., *Immunohistochemistry as an assay.* J Histotechnol, 2023. **46**(4): p. 156-157.
16. Dabbs, D.J., et al., *In Support of Magnani and Taylor.* Arch Pathol Lab Med, 2024. **148**(1): p. 11.
17. Bogen, S.A., *A Root Cause Analysis Into the High Error Rate in Clinical Immunohistochemistry.* Appl Immunohistochem Mol Morphol, 2019.
18. Bogen, S.A., et al., *A Consortium for Analytic Standardization in Immunohistochemistry.* Arch Pathol Lab Med, 2022. **147**(5): p. 584-590.
19. Eren, O.C., et al., *Interobserver and Interassay Concordance in HER2 Immunohistochemistry: Implications on HER2-Low Breast Cancer.* Appl Immunohistochem Mol Morphol, 2026. **34**(1): p. 15-21.
20. Troxell, M.L., *Repurposing HER2 IHC: Pan-HER2 and Low-HER2.* Appl Immunohistochem Mol Morphol, 2026. **34**(3): p. 135-42.
21. Yaziji, H., J.L. Hornick, and M.L. Troxell, *The Suitability of Repurposed, Legacy HER2 Immunohistochemistry Assays for the Detection of HER2 Low and Ultralow Expression: Current Limitations and Potential Considerations.* Appl Immunohistochem Mol Morphol, 2026. **34**(1): p. 1-4.
22. Torlakovic, E.E., et al., *Development and Validation of Measurement Traceability for In Situ Immunoassays.* Clin Chem, 2021. **67**(5): p. 763-771.
23. Nielsen, S., et al., *Lessons Learned, Challenges Taken, and Actions Made for "Precision" Immunohistochemistry. Analysis and Perspectives From the NordiQC Proficiency Testing Program.* Appl Immunohistochem Mol Morphol, 2023. **31**(7): p. 452-458.
24. Torlakovic, E.E. and C.C. Cheung, *Immunohistochemistry: The Importance of Precision Ontology to Precision Oncology.* Appl Immunohistochem Mol Morphol, 2024. **32**(2): p. 61-63.
25. Marx, V., *Method of the Year: spatially resolved transcriptomics.* Nat Methods, 2021. **18**(1): p. 9-14.
26. Wang, S., Y. Wang, and D. Krull, *Histotechnology at the crossroad of spatial biology: bridging legacy and innovation.* J Histotechnol, 2025. **48**(1): p. 1-6.
27. Krull, D., et al., *A best practices framework for spatial biology studies in drug discovery and development: enabling successful cohort studies using digital spatial profiling.* J Histotechnol, 2025. **48**(1): p. 7-26.
28. Cervilla, S., et al., *A technical comparison of spatial transcriptomics platforms across six cancer types.* Genome Biol, 2026. **27**(1): p. 22.
29. Georges, K., et al., *Mitigating Fixation Artifacts in Spatial Transcriptomics: Methodological Insights from Human Brain Tissue.* Mol Neurobiol, 2026. **63**(1).
30. Wagner, C.S., et al., *Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature.* Journal of Informetrics, 2011. **5**(1): p. 14-26.
31. Hall, K.L., et al., *The Science of Team Science: A Review of the Empirical Evidence and Research Gaps on Collaboration in Science.* American Psychologist, 2018. **73**(4): p. 532-548.

NATIONAL SOCIETY FOR HISTOTECHNOLOGY

3545 Ellicott Mills Drive, PMB 204
Ellicott City, MD 20143
P: 443-535-4060
E: histo@nsh.org

32. Zwick, M.E., *Organizing core facilities as force multipliers: strategies for research universities.* J Biomol Tech, 2021. **32**(1): p. 36-41.
33. Ambulos, N.P., Jr., S.M. Mische, and J.W. Fox, *Greater Reliance on Shared Resources Can Enrich the Future Success of Academic Research.* J Biomol Tech, 2025. **36**(2).
34. Gaffney, E.F., et al., *Factors that drive the increasing use of FFPE tissue in basic and translational cancer research.* Biotechnic & Histochemistry, 2018. **93**(5): p. 373-386.
35. Massett, H.A., et al., *Assessing the need for a standardized cancer HUman Biobank (caHUB): findings from a national survey with cancer researchers.* J Natl Cancer Inst Monogr, 2011. **2011**(42): p. 8-15.
36. Sanderson-November, M., et al., *Biorepository best practices for research and clinical investigations.* Contemporary Clinical Trials, 2022. **116**.
37. Vaught, J.B., M.K. Henderson, and C.C. Compton, *Biospecimens and Biorepositories: From Afterthought to Science.* Cancer Epidemiology Biomarkers & Prevention, 2012. **21**(2): p. 253-255.
38. Smith, D.B., A. Ferdous, and J.T. Oxford, *Shared Core Facilities Serve as Hubs for Biomedical Research Network at Institutions of Emerging Excellence.* J Biomol Tech, 2023. **34**(1).
39. Turpen, P.B., et al., *Metrics for Success: Strategies for Enabling Core Facility Performance and Assessing Outcomes.* J Biomol Tech, 2016. **27**(1): p. 25-39.
40. Smith, D.B., et al., *Data Management Tools to Measure the Impact of Core Facilities.* J Biomol Tech, 2021. **32**(2): p. 63-73.
41. Clissold, K.A. and M.S. Itano, *How do you measure the success and impact of a core facility?* Biotechniques, 2025. **77**(7-8): p. 263-270.
42. Ricco, E., et al., *Elevating the Educational Mission of "Full-Service" Core Facilities through Formal Biotechnology Workshops.* J Biomol Tech, 2023. **34**(4).
43. Chlipala, E., et al., *What will it take for histologists to be recognized under CLIA?* J Histotechnol, 2024. **47**(3): p. 93-96.
44. Marcelus, H. and D. Packert, *Addressing the histotechnologist shortage through improved classification and recognition.* J Histotechnol, 2024. **47**(4): p. 143-145.
45. Nieto Sierra, J.A. and D. Gefen, *35 years after CLIA 1988: Key insights and policy implications among laboratory professionals.* PLoS One, 2024. **19**(9): p. e0311251.

217. Bethany Laursen

I am responding: On behalf of myself

Name: Bethany Laursen

Name of Organization: University of Michigan

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: The CRediT taxonomy has advanced the recognition and reward of many kinds of contributions to a research product. However, CRediT does not include contributions to transdisciplinary collaborations, including relationship building and reflective analysis. Transdisciplinary collaborations include patient- and community-engaged research, university/government/industry partnerships, and art-science collaborations, and these are central to translational breakthroughs because they span knowledge groups traditionally kept in siloes. I encourage NIH to sponsor the further development of CRediT-fAIR, an expanded taxonomy of contributorship roles, to the level of a NISO standard so it can be widely adopted like CRediT. CRediT-fAIR has been developed through the art-science journal Ground Works in collaboration with the Alliance for Arts in Research Universities: https://groundworks.io/credit-fair.

I also encourage NIH to review and leverage existing indicators of collaboration and interdisciplinarity in the following open access dataset: Laursen, B. K., Anderson, K. J., & Motzer, N. (2022). Systematic review of the literature assessing interdisciplinarity from 2000 to 2019: A qualitative dataset. https://doi.org/10.7910/dvn/6ucnkw.

The dataset, review papers, and interactive visualizations are listed here: https://aea365.org/blog/reflections-on-practicum-assignments-part-iii-supervisor-perspectives-by-marshall-bailly-assetou-barry-beverly-peters-danuta-dobosz-talatha-kiazolu-reeves-jillian-klarman-michael-petillo-and-2/

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: The Translational Science Benefits Model (TSBM) is an excellent resource. However, it only showcases the ultimate societal benefits of health research and has no place for intermediate indicators of ultimate success. For example, there are no benefits about a "more cohesive research

team". These upstream indicators are essential for tracking progress before societal benefits are realized. Without these intermediate indicators, current practice with the TSBM is to guess at which societal benefits are likely. This guesswork depends on individual judgments of likelihood rather than intersubjective evidence. NIH NCATS explicitly sponsors research and development that targets these upstream impact processes in its quest to understand the science of translation. NIH NCATS could sponsor the development of a corollary TSBM that focuses on features of the translational process itself, and fund research that shows how these are early indicators of the original TSBM outcomes.

8. Other Areas: Regardless of the specific process or outcome being evaluated, I strongly encourage NIH to adopt transparent evaluative reasoning. Such reasoning makes clear what are the criteria and standards being used as thresholds to determine the "worth, merit, and value" of the research being evaluated. If no such thresholds or value determinations are desired, then NIH should be clear that any measurement of the evaluand is simply assessment for ongoing monitoring or exploration, and not for determination of its value. When indicators are tracked but the value we attach to them is not made transparent, then researchers guess and game the system, trying to inflate indicators without understanding the underlying merit they should be pursuing. Laursen et al (2022, attached) explain how this works when it comes to assessing interdisciplinarity.

Description: Laursen, B. K., Motzer, N., & Anderson, K. J. (2022). Pathways for assessing interdisciplinarity: A systematic review. Research Evaluation, 31(3), 326–343. https://doi.org/10.1093/reseval/rvac013

218. Tufts CTSI

I am responding: On behalf of an organization

Name: Hannah M Santos

Name of Organization: Tufts CTSI

Type of Organization: Non-Profit Research Organization

Role: Other

Role - Other: Program Director

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: NIH could support the application of real-world data by funding repositories and funding investigator access and use of the data (similar to the NC3 or All of Us databases).

2. Data, Software, and Model Sharing: Rather than simply denying funding to an investigator unwilling to share data, it would be helpful if grant reviewers rewarded proposals with well-structured data-sharing plans with higher scores.

3. Training and Mentorship: i. Some current NIH policies make it difficult for early-stage investigators to work collaboratively with more senior investigators and thus be mentored by them. For example, NIH disallows early-career status for individual investigators on multi-PI grants. This discourages the inclusion of junior faculty members, effectively blocking a key opportunity for mentoring them.
ii. Also, submitting more than one K award proposal at the same time is not allowed. Given the long turnaround time for the review process, this often leads to gaps in funding for early career researchers. While more than one K can't be funded at the same time, we believe that an investigator should be allowed to apply for an institutional K award while an NIH award is being considered.

4. Collaboration: i. The role of MPI should be exempt from any limit on number of grants on which one can be PI. The MPI role promotes the very goals that the limitation of PI roles intends to achieve. By allowing multiple PIs younger investigators can be included in the MPI role even if they would not yet be eligible for sole PI status. Also, having MPIs promotes team and cross-disciplinary research, which are objectives of NIH.
ii. Include a requirement for broader range of relevant stakeholder groups to provide input on the project. For example, the Army's SBIR Catalyst Program requires partnerships between the small business with a promising technology, Army transition partners, and technology integrators, who work together to speed translation and adoption of the technology research.
iii. Break Through Cancer is a good example of supporting collaboration and data sharing within a disease area. This nonprofit provides a framework for cross-disciplinary collaborations with other researchers working on the same form of cancer. The framework reduces administrative barriers for cross-institutional collaborations through harmonized contracts, a cloud-based data science hub, and 'Team Labs': deeply integrated, multi-institutional research teams working in real-time to solve focused

cancer challenges. Junior faculty are given leadership positions within the 'Team Labs.'
iv. If NIH prioritizes collaborative grants and MPI research, this will encourage universities to prioritize the same in their promotion and tenure decisions.
v. Most research projects are multi-disciplinary and involve multiple investigators. Guidelines on reviewing grants should include recognition of /education on authorship requirements within specific disciplines and the importance of authors who are neither first nor last author.

5. Entrepreneurship and Translation: NCATS or another NIH agency could provide financial support for CTSAs to provide dedicated seed grants for projects seeking commercialization. This would be separate from SBIR/STTR funding mechanisms.

6. Foundational Scientific Exploration: Reinforcing the importance of negative results, providing a searchable online space to record the aims and hypotheses that had negative results, and rewarding investigators for disseminating those results would provide incentives for academic researchers to innovate and take risks.

7. Public Impact: i. Translational Science Benefits Model (TSBM)contains several metrics or benefits under four domains: Clinical, Community, Economic, and Policy
• NIH could require grant applicants to include expected benefits based on this model. For example, does this research improve public health access or improve life expectancy and quality of life?
ii. Grant applicants could be encouraged to use a tool like the TSBM to describe the impact of the proposed research in plain language.
iii. One question that should be part of every grant RFA/RFP is: "What will you do to make sure the general public understands your work?"
iv. To support investigators in communicating results outside of academic venues, grant budgets could include funding for public communication or communication specialists, or providing a service for communicating in plain language could be emphasized in CTSA RFAs.
v. NIH should allocate funds to specifically support engagement with stakeholders, insofar as the applicant demonstrates that engagement can improve public impact.

8. Other Areas: N/A

219. Chronic Pain Research Alliance- an Initiative of The TMJ Association

I am responding: On behalf of an organization

Name: Christin Veasley

Name of Organization: Chronic Pain Research Alliance - an Initiative of The TMJ Association

Type of Organization: Research Participant/Patient Advocacy Organization/Patient Advocate/Caregiver

Role: Research Participant/Patient/Advocate/Caregiver

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: From October 2023 through November 2025, the NIH convened the ENGAGE Working Group to develop a unified vision, framework, roadmap, and recommendations for making meaningful researcher-community partnerships standard practice across NIH-funded clinical research. In December 2025 the NIH Director endorsed the report and committed to implementing the recommendations agency-wide. As NIH moves in this direction, its measures of scientific impact should recognize the education, applied training, and mentorship that enable NIH-funded researchers to partner effectively with people with lived experience (PWLE) and community organizations.

EDUCATION
NIH should recognize and reward researchers who pursue foundational education in PWLE/community-engaged research principles, best practices, and implementation methods. This education should address how meaningful engagement differs from recruitment, research participation, outreach, or one-time consultation; how to assess readiness; establish roles and responsibilities of all partners; identify and connect with appropriate partners; and ensure broad representation. It should teach researchers specific implementation methods to partner with PWLE and communities across the research lifecycle – from developing research questions and designing studies through grant development, study launch and conduct, data interpretation, development and dissemination of knowledge products, and study close-out. Cross-cutting topics should include shared values and decision-making; equitable compensation; bidirectional communication and transparency; accessibility, accommodations, and flexible participation; partner preparation and skills development; recognition and credit; engagement measurement and iterative improvement; and documentation of how partner input changes the research. Whenever possible, educational activities should be co-developed and co-delivered by PWLE and community partners. An exemplar is the research engagement curriculum – Foundations of PWLE Engagement in Clinical Research – which was co-developed and co-taught by PWLE members of the NIH HEAL National K12 Clinical Pain Career Development Program.

Potential metrics:
• Completion of substantive education in PWLE/community-engaged research.
• Demonstrated understanding through knowledge assessments or applied exercises.

• Inclusion of engagement-related learning goals in career-development or research-team plans.
• Documentation of how this education informed the researcher's engagement approach or research decisions.
• Continued education as the researcher assumes responsibility for additional and/or expanded partnerships.

Potential NIH incentives include the following:
• Include engagement education as a review consideration in appropriate career-development, fellowship, institutional training, center, and clinical research awards.
• Allow grant funds and protected time for engagement education.
• Include progress toward educational goals in progress reports and renewals.
• Allow/encourage these activities to be documented in NIH biosketches, NIH RePORTER, and grant products.
• Consider evidence that researchers applied this education in subsequent funding decisions.

NIH should assess what researchers learned and applied - not simply count courses, hours, or certificates.

APPLIED TRAINING AND PWLE MENTORSHIP
NIH should separately recognize and reward researchers who undertake applied, competency-based training in developing, implementing, evaluating, and sustaining meaningful partnerships with PWLE and/or community partners. This training should be co-developed and co-delivered by PWLE and community partners and, where appropriate, include a compensated PWLE or community mentor who contributes to the researcher's career development and research. The NIH HEAL National K12 Clinical Pain Career Development Program provides a promising model: scholars are matched with mentors with lived pain experience, receive training in PWLE and community engagement, develop a PWLE/community engagement plan, and apply what they learn in their own research.

Potential metrics include:
• Completion of applied, competency-based engagement training.
• Development and implementation of study-specific engagement plans, timelines, budgets, partner roles, and evaluation approaches.
• Inclusion of a compensated PWLE and/or community mentors with clearly defined goals and responsibilities.
• Assessments from both the researcher and PWLE mentor regarding the quality and value of the mentorship.
• Documentation of changes made to the research because of the training or mentorship.
• Feedback from PWLE and/or community partners regarding the researcher's engagement practices.
• Continued application of engagement competencies in subsequent research.

Potential NIH incentives include the following:
• Include applied engagement training and PWLE/community mentorship as scored criteria in career-development and training awards.
• Allow dedicated funds and protected time for training and mentorship, including compensation and

administrative support for PWLE mentors.
• Offer administrative supplements for researchers seeking applied training or adding a PWLE/community mentor.
• Include engagement competencies and mentorship outcomes in progress reports and renewal decisions.
• Allow/encourage researchers to document these contributions in NIH biosketches, NIH RePORTER, and grant products.

RESOURCES/REFERENCES
- NIH HEAL K12 Clinical Pain Research Career Development Award (https://www.jpain.org/article/S1526-5900(25)01059-4/abstract, https://www.nih.gov/heal/research/training-workforce)

4. Collaboration: NIH should recognize meaningful partnership with people with lived experience (PWLE) and community members/organizations as a valued form of scientific collaboration and team science. Investigators should be incentivized to involve PWLE and community partners throughout the research lifecycle – from identifying health priorities and developing research questions through study design, grant development, study preparation and conduct, data analysis and interpretation, return of results, and dissemination and implementation. PWLE and community partners should be recognized as contributors of expertise rather than treated only as advisors or recruitment resources.

NIH should not use the number of advisory-panel members or meetings as a measure of success. Recognition should depend on the quality, continuity, and demonstrated influence of the partnership.

Possible metrics include:
1. The point at which PWLE and community partners became involved and the research lifecycle stages in which they participated.
2. The degree of clearly defined roles, responsibilities, bi-directional communication, and decision-making authority.
3. A dedicated engagement budget, including timely and appropriate compensation for individual partners and community organizations.
4. Representation of the populations affected by or expected to use the research, including accommodations needed to support participation.
5. Documentation of specific decisions, materials, methods, outcomes, analyses, or dissemination strategies changed because of partner input.
6. Partner-reported assessments of trust, respect, communication, support, reciprocity, and ability to influence decisions.
7. Evidence that partners received appropriate authorship, contributor credit, presentation opportunities, or other recognition.
8. Continuity of the partnership and efforts to sustain relationships beyond a single grant.

Ways NIH could incentivize and reward these activities include:
1. Ask institutions applying for training, career development, center, and cooperative agreement awards to describe how community-engaged scholarship and team-based contributions are recognized in faculty evaluation, promotion, and tenure.
2. Require a study-specific engagement plan and budget in funding opportunities and make engagement

quality a scored review criterion.
3. Include compensated PWLE and community reviewers on peer-review panels evaluating engagement plans.
4. Provide planning awards or pre-application funding so researchers and communities can develop partnerships before the research question and protocol are developed and finalized.
5. Offer administrative supplements to strengthen or expand engagement during an existing award.
6. Require investigators to report progress against engagement milestones in annual progress reports and consider performance in continuation and renewal decisions.
7. Allow PWLE and community partners to be formally identified and credited as members of the scientific team in applications, NIH RePORTER, biosketches, and grant products.
8. Establish awards recognizing research teams that demonstrate exemplary and sustained research-community partnership.

Resources and Examples include:
- NIH ENGAGE Working Group Report (https://osp.od.nih.gov/wp-content/uploads/2025/12/NExTRACs-ENGAGE-Report_FINAL.pdf): Provides an engagement framework spanning the research lifecycle and recommends NIH requirements, support, measurement, and accountability.
- PCORI Foundational Expectations for Partnerships in Research (https://www.pcori.org/engagement-research/engagement-resources/foundational-expectations): establishes expectations for early and ongoing engagement, partner compensation, shared decision-making, team capacity, and continuous assessment.
- Medical College of Wisconsin (https://www.mcw.edu/-/media/MCW/Departments/Faculty-Affairs/MCW-Faculty-Handbook.pdf): explicitly recognizes Community Engagement as an accomplishment area in faculty promotion, with criteria spanning involvement, achievement, and leadership. Examples include sustained community partnerships, community-engaged scholarship, dissemination to community and academic audiences, and demonstrated impact within both the community and the faculty member's field. Community engagement can also inform tenure through MCW's assessment of contributions to its institutional missions and scholarly and programmatic achievements.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: We applaud NIH's goal of expanding the definition of scientific impact beyond grants, publications, citations, and presentations at scientific meetings. Publicly funded research has greater impact when it addresses questions that are important to affected communities, produces findings that intended users can understand and apply, and contributes to improvements in healthcare, coverage, education, community programs, policy, public trust, and/or health.

Studies should be designed from the outset with dissemination and, where appropriate, implementation in mind. Investigators should identify the intended end users of their findings – such as patients, caregivers, clinicians, insurers and other payers, health systems, community organizations, employers, educators, and policymakers – and partner with representatives of these groups when selecting research questions and outcomes, designing the study, and planning dissemination. This early involvement can help ensure that the study generates findings and outputs that are relevant, accessible, and useful.

NIH should require dissemination and implementation plans in appropriate grant applications and provide sufficient funding to carry them out. Plans should identify relevant end-users of study data, intended audiences, knowledge products, communication channels, partner roles, timelines, accessibility considerations, evaluation methods, and the intended use of the findings. Researchers should work with PWLE, community partners, and other end users to co-develop, test, refine, and disseminate audience-specific products such as plain-language summaries, infographics, webinars, videos, decision tools, clinical resources, payer or policy briefs, toolkits, and community presentations. These activities should be recognized as substantive scientific contributions, not optional outreach.

Metrics NIH could use include:
1. Near-term impact: Early identification and involvement of intended study data end-users; a funded dissemination and implementation plan; relevance of the research questions and outcomes; changes made because of partner input; improvements in study feasibility, acceptability, recruitment, retention, or participant burden; and appropriate compensation and credit for partners.
2. Study outputs: Co-development of audience-specific knowledge products; timely return of aggregate and, when appropriate, individual results; and assessment by intended users of the products' accessibility, clarity, relevance, cultural appropriateness, and usefulness.
3. Intermediate impact: Reach through trusted channels, interpreted in context; use of findings or products by patients, clinicians, payers, health systems, community organizations, educators, or policymakers; adoption in training, programs, guidelines, coverage decisions, or organizational practices; and continuation of partnerships.
4. Longer-term impact: Changes in clinical practice, healthcare access or coverage, policy, public health programs, health outcomes, costs, or public trust when reasonably attributable to the research and measurable within an appropriate timeframe.
5. Public-impact narrative: A concise account of who benefited, what changed, the investigator's contribution, and the evidence supporting the claimed impact.
6. End-user feedback: Input from PWLE, community partners, and other intended users regarding the relevance, usefulness, uptake, and effects of the research and its outputs.

Ways NIH could incentivize and reward these activities include:
- Require a dissemination and implementation plan in appropriate grant applications and make its quality a scored review criterion.
- Require applicants to budget for end-user engagement, partner compensation, development and testing of knowledge products, accessible formats, return of results, and co-dissemination.
- Allow dissemination and implementation planning costs from the beginning of the award rather than limiting support to the study's final stages.
- Provide final-year dissemination supplements and follow-on implementation awards after findings become available.
- Include progress toward dissemination and public-impact goals in annual and final reports, competing renewals, and subsequent funding decisions.
- Allow nontraditional outputs, knowledge products, engagement contributions, and evidence of public impact to be reported in NIH biosketches, NIH RePORTER, and future applications.
- Establish NIH awards or showcases recognizing investigators, community partners, and teams that

demonstrate exemplary co-dissemination or public impact.
- Ask institutions applying for training, center, and cooperative agreement awards how engagement, dissemination, and public impact are recognized in annual review, promotion, and tenure.

NIH should reward evidence that dissemination was planned early, appropriately funded, co-developed with intended users, and resulted in useful and used products, not simply the number of products created, downloads, or social-media impressions. Investigators should not be penalized for null or negative findings or for long-term outcomes beyond their control. They should be recognized for responsible communication, fulfillment of dissemination commitments, and credible evidence of impact within their sphere of influence.

Resources and examples include:
- PCORI Dissemination and Implementation Framework and Toolkit (https://www.pcori.org/implementation-evidence/putting-evidence-work/dissemination-and-implementation-framework-and-toolkit): Integrates early audience identification, partner engagement, dissemination, implementation, and evaluation.
- PCORI Engagement Award: Dissemination Initiative (https://www.pcori.org/funding-opportunities/announcement/engagement-award-dissemination-initiative-spring-2026-cycle): Demonstrates how a funder can provide dedicated support for dissemination through trusted organizations and networks.
- Washington University Translational Science Benefits Model and Translating for Impact Toolkit (https://www.translationalsciencebenefits.wustl.edu/about-the-model/): Provide a framework for planning, tracking, and demonstrating clinical, community, economic, and policy impacts.
- CDC Clear Communication Index (https://www.cdc.gov/ccindex/widget.html): Provides research-based criteria for developing and evaluating public-facing communication products.

8. Other Areas: N/A

220. Frontiers Media Inc.

I am responding: On behalf of an organization

Name: Eleonora Colangelo

Name of Organization: Frontiers Media Inc.

Type of Organization: Publisher

Role: Other

Role - Other: Policy Lead

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Frontiers welcomes NIH's Replication and Reproducibility Initiative and its push toward gold standard science, which we see as the premise of good publishing policy rather than its endpoint. As set out in our response to the NIH Request for Information NOT-OD-26-047 (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-047.html), the largest gains available to NIH sit in a part of the research system that is currently being significantly changed by the advent of AI: the editorial and integrity workflows that catch error before it enters the record. Frontiers has spent the past eight years engineering that layer, and we would encourage NIH to similarly use integrity in the published scientific records as a clear indicator of rigor. Every indicator NIH is considering depends on accuracy and integrity in practice. A replication study is only informative if the original record is accurate, and preregistration only constrains analysis if deviations are caught at review. Whether NIH is measuring reproducibility, data and software sharing, collaboration, or public impact, accuracy and integrity are essential, and this response draws on Frontiers' operating experience to inform each of those areas below.

The stakes are large and quantified. OSTP's July 2026 report to the President, Science: A New Golden Age (https://www.whitehouse.gov/wp-content/uploads/2026/07/Science-A-New-Golden-Age.pdf), cites published estimates that irreproducible findings in preclinical biomedical research alone misdirect roughly $28 billion a year, and it warns that AI will make plausible results easier to generate than to verify. The same report sets out the Genesis Mission's ambition to unlock federal scientific data and build AI-ready datasets capable of driving discovery at scale. Both goals rest on the same requirement, which is a supply of high quality, structured, reviewed information for AI systems to work from. Peer-reviewed literature and curated, well-determined data are among the most dependable inputs available for that purpose. Frontiers would welcome a conversation with NIH about contributing reviewed content and AI-ready data to this effort.
Every manuscript submitted to Frontiers passes through AIRA, our AI Review Assistant, which runs more than 40 automated integrity and quality checks before a single reviewer sees the paper, covering language quality, image manipulation, plagiarism, conflicts of interest, scope, and potentially fabricated content. That screening blocks or rejects roughly 32% of submissions before peer review begins. Our 66-person research integrity team includes a dedicated auditing division, created in 2022, that proactively investigates paper-mill networks rather than waiting for a report to arrive. Furthermore, Frontiers holds

a COPE Trustee seat, chairs the COPE Papermill Working Group, is a signatory of United2Act, and contributes to the STM Integrity Hub, the shared infrastructure publishers use to flag paper-mill activity, duplicate submissions, and citations to retracted work across the industry before they contaminate someone else's reference list. These are precisely the workflows that must continue to improve as AI generated and hallucinated content reaches the literature in volume, and Frontiers sees real value in iterating and improving these processes with NIH.

None of these pre-review analyses shows up in a citation count, but all of it is what makes a citation trustworthy.

NIH's framework should reward publication of replication studies, null results and confirmatory research; use of registered reports and preregistration; ensuring availability of underlying data, code and protocols; adherence to field-specific reporting standards; and the integrity functions of peer review, editorial assessment, corrections and retractions. Frontiers would add one further indicator: direct recognition of pre-publication screening and post-publication correction activity as evidence of how rigorously the record is being defended, and the quality of each publication validated. Merit review and peer-reviewed publication are already integral to how federal science is evaluated, and Frontiers agrees with the focus of this RFI on broadening the measures of scientific impact. Data, code, protocols, negative results and methodological work deserve to be discoverable and credited too, rather than sitting where neither other researchers nor the public can benefit from them. Team-based science raises a related measurement problem: a paper with dozens of authors does not tell NIH who did which aspects or what percentage of the work. Frontiers adopted the CRediT contributor taxonomy in 2023 (https://www.frontiersin.org/news/2023/07/20/frontiers-adopts-credit-to-enhance-clarity-in-research-authorship/), which disaggregates authorship into 14 defined roles, from conceptualization and methodology to data curation and funding acquisition. That structured record is directly useful for assessing individual contributions to team-based science, and NIH's framework should recognize CRediT style role disclosure as an indicator in its own right and explore how it can be expanded for data, software, and additional research outputs.

Two cautions. First, NIH should not trade one narrow metric, the publication count, for another equally narrow one such as a single reproducibility or deposit indicator. NIH's own Data Sharing Index (S-index) Challenge models the better approach (https://www.nih.gov/challenges/nih-data-sharing-index-s-index-challenge). It scores how effectively a researcher shares data, weighing FAIR adherence, timeliness, completeness of annotation, actual reuse by others, and the publications and patents that follow. That set of data is a measure of overall quality and impact, the logic we would encourage NIH to carry across the rest of the framework. Second, what replication looks like varies by research field. Clinical, laboratory, computational and population-health research each need different reproducibility infrastructure, and a single checklist will fit none of them well.

Frontiers and its industry peers already run, at scale, systems NIH can build this framework on. We would welcome the chance to share our integrity data directly with NIH in the context of any post-submission follow-up, and to explore how these workflows could support the Replication and Reproducibility Initiative in practice.

2. Data, Software, and Model Sharing: Frontiers supports NIH's effort to reward genuine data and code sharing, a standard that we apply to our own authors. Our policies and publication ethics require that the minimum dataset supporting a paper's conclusions be made available, that standardized data be deposited in the relevant community repositories, and that every article carries an explicit data availability statement (https://www.frontiersin.org/guidelines/policies-and-publication-ethics).

Since October 2025, we have gone further with our FAIR² Data Management service. It publishes data itself as a citable, peer-reviewed article type, with an interactive portal offering AI chat, visualization tools and notebooks for reuse, and an AI data steward that helps researchers get metadata right the first time (https://www.frontiersin.org/about/fair-data-management). FAIR² extends the FAIR principles — Findable, Accessible, Interoperable, Reusable — with AI-readiness and governance, and it is aligned with NIH's and NSF's own data-sharing mandates.

That experience tells us where NIH's incentive design will succeed: on quality rather than volume. OSTP's July 2026 report makes the same point about federally funded data, noting that valuable datasets are "routinely abandoned, sometimes due to a lack of stable funding for storage and curation, and other times because there is no signal on the value of the information." A deposit count rewards quantity and does nothing to improve a dataset that is incomplete, undocumented, or disconnected from the paper that produced it. NIH's framework should instead reward deposition in trusted repositories with links back to the literature, persistent identifiers and community metadata standards, documented evidence of reuse or citation, clear versioning and licensing, and appropriate access controls where privacy, security or participant protection genuinely require them. That is the same balance NIH Director Bhattacharya has set for the agency's own Real-World Data Platform, being built "consistent with a deep respect for individual privacy rights" (https://www.nih.gov/about-nih/nih-director/statements/advancing-nihs-mission-through-unified-strategy).

Making any of this measurable depends on metadata. Crossref and DataCite both operate services that connect datasets and software to the articles describing them, through data and software citation deposits (https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/data-citation/) and the relationships registered between DOIs (https://www.crossref.org/documentation/retrieve-metadata/data-citations/). The value and longer-term impact of those links depend on publishers supplying complete, accurate metadata at scale. Frontiers integrated with the OA Switchboard in March 2026 to automate metadata exchange with institutions, libraries, and funders (https://www.frontiersin.org/news/2026/03/30/frontiers-integrates-with-oa-switchboard-to-streamline-open-access-workflows). Curated metadata is what makes open science, rigor, and reproducibility monitorable at all.

Publishers are building this infrastructure (i.e., data-article linking, software citation, repository integrations, metadata enhancement) because it is where the field is going. It is also where OSTP's own vision points: Science: A New Golden Age anticipates researchers releasing "shorter outputs more frequently, including datasets, code, preliminary findings, and methodological notes," with "dynamic papers" that "update automatically as underlying data changes." A natural role for a publisher in that future is helping researchers document and link their work so that others can find it, build on it and apply it. Done properly, this is what produces AI-ready data and the richer linked research graphs that

connect papers, datasets, code, methods and people. That is exactly the substrate programs that the Replication and Reproducibility Initiative and the Genesis Mission need, and it is where we think publishers can be most useful to NIH. The incentive framework should accelerate that connectivity of research, data and outputs, rather than creating a parallel compliance regime that counts deposits and cannot see the connections between them.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Real translational benefit from biomedical research often takes years to surface, which is the same long horizon the Science: A New Golden Age report gives as the rationale for federal science funding itself. Short-term signals such as news coverage, downloads, views, and social media posts measure attention rather than health outcomes, and the report makes the same point about the incentives around publication: “Journal editors want clear narratives. University press offices demand headlines. Nuance dies, as a result, in the race for attention.”

Nevertheless, the pathway to public benefit runs through publication. What NIH can measure with confidence is whether that pathway is intact: whether findings are peer reviewed, indexed, correctable and openly available. NIH's own Public Access Policy says as much. Increasing access to NIH-funded publications “can accelerate future research, lead to collaboration, and allow interested readers and patients to follow the latest advances more closely,” and reflects NIH's commitment to the “responsible stewardship of the Nation's investment in biomedical research by improving transparency and accessibility of taxpayer-funded research, an essential component of fostering trust in research.” Frontiers has delivered exactly that pathway since 2007. Every article is immediately and permanently free to read under CC-BY 4.0, with no embargo and no paywalled version of record, across more than 759,000 articles in 222 journals. Frontiers delivers this open access model at a comparatively lower cost to researchers, institutions, and funding administrations alike — improving efficiency and enabling highly curated distribution of scientific knowledge, while helping administrators save money relative to prevailing market rates.

Frontiers asks NIH to weigh this RFI against a live financial reality it is also facing. GAO reported in May 2026 that federal agencies spent roughly $295 million on article processing charges in 2024, projected to reach $582 to $937 million by 2030 (https://www.gao.gov/products/gao-26-107738). A separate rulemaking now under consideration at OMB would respond to that pressure with a blanket ban on publication and subscription costs, a step Frontiers has urged OMB to withdraw because it would defund the peer review, integrity screening, indexing and correction functions that make the scholarly record trustworthy in the first place. A framework that rewards trusted dissemination as a component of public impact, while a parallel rule removes the funding that sustains it, sends NIH-funded researchers in two directions at once. We ask NIH to treat reasonable publication costs as impact infrastructure, not overhead.

Publication is also where American research competes globally and NIH's direction can strengthen that

position rather than simply defend it. China published approximately 1.1 million research articles in 2024, compared with 880,000 from the United States, and the Science: A New Golden Age report describes the same shift across the wider enterprise, observing that “America no longer stands alone as the unchallenged leader in science and technology” while China's R&D spending “has surged from a negligible fraction of U.S. levels to full parity on a purchasing-power-adjusted basis.” The answer to that is to broaden what American research shares and how usable it is: open results, reviewed data, documented methods and machine-readable outputs that let computational and AI-driven research move faster on top of them. A framework that treats publication and structured dissemination as central to the mission, and resources them accordingly, is one of the more direct levers NIH has for unlocking U.S. advantage in the quality of our scientific output.

The link between publications and patents merits further attention. Patent citation of the underlying literature is one of the more direct ways to trace research into downstream industrial activity, and it remains comparatively under-tracked next to citation counts. Understanding that connection, and how it varies across fields and institutions, is a natural next step before NIH builds an indicator to assess public impacts. Frontiers’ own publication workflows already build the infrastructure necessary to trace impacts: persistent identifiers, structured metadata and DOI-to-DOI relationships registered through Crossref and DataCite, the same mechanisms that let a patent examiner’s citation be traced back to the paper it drew on. Surfacing those connections for authors and institutions, rather than leaving them buried in citation databases, is work we would welcome doing alongside NIH. Public impact also depends on infrastructure that connects research to the people who can act on it. Frontiers Science House convened over 100 researchers, policymakers, industry leaders and Nobel laureates from 16 countries alongside the World Economic Forum's January 2026 Annual Meeting in Davos, translating findings into conversations with the people who set policy and allocate capital (https://www.alpsp.org/news-publications/industry-news/100-scientific-leaders-convene-in-davos-to-shape-global-decision-making/). That kind of engagement is the credible middle step between publication and public benefit, and NIH's framework should treat structured translation activity of this kind as evidence of impact.

8. Other Areas: Frontiers urges NIH to state plainly, in whatever framework results from this RFI, that publication and dissemination are not adjacent to NIH's mission. They are how that mission is completed. NIH's own public access, reproducibility and gold standard science initiatives all state this in different words already.

OSTP's July 2026 report raises pointed questions about how research is published, reviewed and rewarded, including the risk that publication metrics “fall to Goodhart's Law as gameable targets” as AI makes it easier to generate plausible-looking research at scale. Frontiers shares that concern and welcomes debate about possible approaches. The scholarly ecosystem exists to enable collaboration and scrutiny, and we look forward to modernizing it in the direction OSTP describes, particularly around AI. Part of what the report asks for is already how we operate. Every Frontiers article is open to every reader immediately, with no embargo, and the evidence on open access is consistent: once age and discipline are accounted for, openly available articles receive around 18% more citations than average, and open outputs are reused and built on more widely (https://peerj.com/articles/4375/). Extending that same openness, review, and curation to data, code, and methods is the practical path to the system the report envisions, and it is a path we would like to take with NIH.

Reform of narrow bibliometrics is needed but should not be read as meaning that publication itself becomes dispensable. Correcting the over-reliance on publication counts and impact factors allows a

richer picture of what publication does, supporting peer review, transparency, correctability, and reuse. Publication is essential for sharing and building on research outputs, and no aspect of research benefits from a smaller role for publication in the research lifecycle. That richer picture of the roles for publication is still being drawn, and Frontiers wants to help draw it with NIH rather than working in parallel.
Frontiers, as a fully open-access publisher operating since 2007, would welcome a direct conversation with NIH on how Frontiers' scientific-impact framework, data and AI programs, as well as its costs and access policies can pull in the directions that match NIH's goals.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/NOT-OD-26-087_Frontiers_20260819.pdf

frontiers

August 18th, 2026

Subject: Response to NOT-OD-26-087: Measuring and Rewarding Scientific Impact

Frontiers welcomes the opportunity to respond to NIH's Request for Information on Measuring and Rewarding Scientific Impact. As one of the world's largest and most impactful research publishers — and the fifth most-cited scientific publisher worldwide — Frontiers is dedicated to making peer-reviewed, quality-certified science openly accessible to researchers, clinicians, policymakers, and the taxpayer. We are a fully Gold Open Access publisher and Open Science platform, supporting American researchers by publishing their work in ways that maximize the visibility, accessibility, and impact of publicly funded science. Born digital in 2007, Frontiers has combined human editorial expertise with proprietary AI from the outset — most notably AIRA (AI Review Assistant, in use since 2017), which runs 40+ automated integrity and quality checks on every submission, and FAIR² (Fair Squared, launched October 2025), our peer-reviewed data publication service that turns research data into citable, machine-actionable, AI-ready assets aligned with NIH, Horizon Europe, and UKRI mandates. We submit these comments as a long-standing partner to the U.S. research community — committed to helping NIH translate its priorities into trustworthy, openly accessible science that powers American innovation in health, biotechnology, and the bioeconomy — and we offer the comments below in support of a vision that treats high-quality, open scholarly publishing as a necessary condition for delivering on three main pillars.

Rigor and Reproducibility

Frontiers welcomes NIH's Replication and Reproducibility Initiative and its push toward gold standard science, which we see as the *premise* of good publishing policy rather than its endpoint. As set out in our response to the NIH Request for Information NOT-OD-26-047, the largest gains available to NIH sit in a part of the research system that is currently being significantly changed by the advent of AI: the editorial and integrity workflows that catch error before it enters the record. Frontiers has spent the past eight years engineering that layer, and we would encourage NIH to similarly use integrity in the published scientific records as a clear indicator of rigor. Every indicator NIH is considering depends on accuracy and integrity in practice. A replication study is only informative if the original record is accurate, and preregistration only constrains analysis if deviations are caught at review. Whether NIH is measuring reproducibility, data and software sharing, collaboration, or public impact, accuracy and integrity are essential, and this response draws on Frontiers' operating experience to inform each of those areas below.

The stakes are large and quantified. OSTP's July 2026 report to the President, Science: A New Golden Age, cites published estimates that irreproducible findings in preclinical biomedical research alone misdirect roughly $28 billion a year, and it warns that AI will make plausible results easier to generate than to verify. The same report sets out the Genesis Mission's ambition to unlock federal scientific data and build AI-ready datasets capable of driving discovery at scale. Both goals rest on the same requirement, which is a supply of high quality, structured, reviewed information for AI systems to work from. Peer-reviewed literature and curated, well-determined data are among the most dependable inputs available for that

purpose. Frontiers would welcome a conversation with NIH about contributing reviewed content and AI-ready data to this effort.

Every manuscript submitted to Frontiers passes through AIRA, our AI Review Assistant, which runs more than 40 automated integrity and quality checks before a single reviewer sees the paper, covering language quality, image manipulation, plagiarism, conflicts of interest, scope, and potentially fabricated content. That screening blocks or rejects roughly 32% of submissions before peer review begins. Our 66-person research integrity team includes a dedicated auditing division, created in 2022, that proactively investigates paper-mill networks rather than waiting for a report to arrive. Furthermore, Frontiers holds a COPE Trustee seat, chairs the COPE Papermill Working Group, is a signatory of United2Act, and contributes to the STM Integrity Hub, the shared infrastructure publishers use to flag paper-mill activity, duplicate submissions, and citations to retracted work across the industry before they contaminate someone else's reference list. These are precisely the workflows that must continue to improve as AI generated and hallucinated content reaches the literature in volume, and Frontiers sees real value in iterating and improving these processes with NIH.

None of these pre-review analyses shows up in a citation count, but all of it is what makes a citation trustworthy.

NIH's framework should reward publication of replication studies, null results and confirmatory research; use of registered reports and preregistration; ensuring availability of underlying data, code and protocols; adherence to field-specific reporting standards; and the integrity functions of peer review, editorial assessment, corrections and retractions. Frontiers would add one further indicator: direct recognition of pre-publication screening and post-publication correction activity as evidence of how rigorously the record is being defended, and the quality of each publication validated. Merit review and peer-reviewed publication are already integral to how federal science is evaluated, and Frontiers agrees with the focus of this RFI on broadening the measures of scientific impact. Data, code, protocols, negative results and methodological work deserve to be discoverable and credited too, rather than sitting where neither other researchers nor the public can benefit from them. Team-based science raises a related measurement problem: a paper with dozens of authors does not tell NIH who did which aspects or what percentage of the work. Frontiers adopted the CRediT contributor taxonomy in 2023, which disaggregates authorship into 14 defined roles, from conceptualization and methodology to data curation and funding acquisition. That structured record is directly useful for assessing individual contributions to team-based science, and NIH's framework should recognize CRediT style role disclosure as an indicator in its own right and explore how it can be expanded for data, software, and additional research outputs.

Two cautions. First, NIH should not trade one narrow metric, the publication count, for another equally narrow one such as a single reproducibility or deposit indicator. NIH's own Data Sharing Index (S-index) Challenge models the better approach. It scores how effectively a researcher shares data, weighing FAIR adherence, timeliness, completeness of annotation, actual reuse by others, and the publications and patents that follow. That set of data is a measure of overall quality and impact, the logic we would encourage NIH to carry across the rest of the framework. Second, what replication looks like varies by research field. Clinical, laboratory, computational and population-health research each need different reproducibility infrastructure, and a single checklist will fit none of them well.

Frontiers and its industry peers already run, at scale, systems NIH can build this framework on. We would welcome the chance to share our integrity data directly with NIH in the context of any post-submission follow-up, and to explore how these workflows could support the Replication and Reproducibility Initiative in practice.

frontiers

Data, Software, and Model Sharing

Frontiers supports NIH's effort to reward genuine data and code sharing, a standard that we apply to our own authors. Our policies and publication ethics require that the minimum dataset supporting a paper's conclusions be made available, that standardized data be deposited in the relevant community repositories, and that every article carries an explicit data availability statement. Since October 2025, we have gone further with our FAIR² Data Management service. It publishes data itself as a citable, peer-reviewed article type, with an interactive portal offering AI chat, visualization tools and notebooks for reuse, and an AI data steward that helps researchers get metadata right the first time. FAIR² extends the FAIR principles — Findable, Accessible, Interoperable, Reusable — with AI-readiness and governance, and it is aligned with NIH's and NSF's own data-sharing mandates.

That experience tells us where NIH's incentive design will succeed: on quality rather than volume. OSTP's July 2026 report makes the same point about federally funded data, noting that valuable datasets are "routinely abandoned, sometimes due to a lack of stable funding for storage and curation, and other times because there is no signal on the value of the information." A deposit count rewards quantity and does nothing to improve a dataset that is incomplete, undocumented, or disconnected from the paper that produced it. NIH's framework should instead reward deposition in trusted repositories with links back to the literature, persistent identifiers and community metadata standards, documented evidence of reuse or citation, clear versioning and licensing, and appropriate access controls where privacy, security or participant protection genuinely require them. That is the same balance NIH Director Bhattacharya has set for the agency's own Real-World Data Platform, being built "consistent with a deep respect for individual privacy rights."

Making any of this measurable depends on metadata. Crossref and DataCite both operate services that connect datasets and software to the articles describing them, through data and software citation deposits and the relationships registered between DOIs. The value and longer-term impact of those links depend on publishers supplying complete, accurate metadata at scale. Frontiers integrated with the OA Switchboard in March 2026 to automate metadata exchange with institutions, libraries, and funders. Curated metadata is what makes open science, rigor, and reproducibility monitorable at all.

Publishers are building this infrastructure (i.e., data-article linking, software citation, repository integrations, metadata enhancement) because it is where the field is going. It is also where OSTP's own vision points: *Science: A New Golden Age* anticipates researchers releasing "shorter outputs more frequently, including datasets, code, preliminary findings, and methodological notes," with "dynamic papers" that "update automatically as underlying data changes." A natural role for a publisher in that future is helping researchers document and link their work so that others can find it, build on it and apply it. Done properly, this is what produces AI-ready data and the richer linked research graphs that connect papers, datasets, code, methods and people. That is exactly the substrate programs that the Replication and Reproducibility Initiative and the Genesis Mission need, and it is where we think publishers can be most useful to NIH. The incentive framework should accelerate that connectivity of research, data and outputs, rather than creating a parallel compliance regime that counts deposits and cannot see the connections between them.

Public Impact

Real translational benefit from biomedical research often takes years to surface, which is the same long horizon the *Science: A New Golden Age* report gives as the rationale for federal

science funding itself. Short-term signals such as news coverage, downloads, views, and social media posts measure attention rather than health outcomes, and the report makes the same point about the incentives around publication: “Journal editors want clear narratives. University press offices demand headlines. Nuance dies, as a result, in the race for attention.”

Nevertheless, the pathway to public benefit runs through publication. What NIH can measure with confidence is whether that pathway is intact: whether findings are peer reviewed, indexed, correctable and openly available. NIH's own Public Access Policy says as much. Increasing access to NIH-funded publications “can accelerate future research, lead to collaboration, and allow interested readers and patients to follow the latest advances more closely,” and reflects NIH's commitment to the “responsible stewardship of the Nation's investment in biomedical research by improving transparency and accessibility of taxpayer-funded research, an essential component of fostering trust in research.” Frontiers has delivered exactly that pathway since 2007. Every article is immediately and permanently free to read under CC-BY 4.0, with no embargo and no paywalled version of record, across more than 759,000 articles in 222 journals. Frontiers delivers this open access model at a comparatively lower cost to researchers, institutions, and funding administrations alike — improving efficiency and enabling highly curated distribution of scientific knowledge, while helping administrators save money relative to prevailing market rates.

Frontiers asks NIH to weigh this RFI against a live financial reality it is also facing. GAO reported in May 2026 that federal agencies spent roughly $295 million on article processing charges in 2024, projected to reach $582 to $937 million by 2030. A separate rulemaking now under consideration at OMB would respond to that pressure with a blanket ban on publication and subscription costs, a step Frontiers has urged OMB to withdraw because it would defund the peer review, integrity screening, indexing and correction functions that make the scholarly record trustworthy in the first place. A framework that rewards trusted dissemination as a component of public impact, while a parallel rule removes the funding that sustains it, sends NIH-funded researchers in two directions at once. We ask NIH to treat reasonable publication costs as impact infrastructure, not overhead.

Publication is also where American research competes globally and NIH's direction can strengthen that position rather than simply defend it. China published approximately 1.1 million research articles in 2024, compared with 880,000 from the United States, and the *Science: A New Golden Age* report describes the same shift across the wider enterprise, observing that “America no longer stands alone as the unchallenged leader in science and technology” while China's R&D spending “has surged from a negligible fraction of U.S. levels to full parity on a purchasing-power-adjusted basis.” The answer to that is to broaden what American research shares and how usable it is: open results, reviewed data, documented methods and machine-readable outputs that let computational and AI-driven research move faster on top of them. A framework that treats publication and structured dissemination as central to the mission, and resources them accordingly, is one of the more direct levers NIH has for unlocking U.S. advantage in the quality of our scientific output.

The link between publications and patents merits further attention. Patent citation of the underlying literature is one of the more direct ways to trace research into downstream industrial activity, and it remains comparatively under-tracked next to citation counts. Understanding that connection, and how it varies across fields and institutions, is a natural next step before NIH builds an indicator to assess public impacts. Frontiers’ own publication workflows already build the infrastructure necessary to trace impacts: persistent identifiers, structured metadata and DOI-to-DOI relationships registered through Crossref and DataCite, the same mechanisms that let a patent examiner’s citation be traced back to the paper it drew on. Surfacing those connections for authors and institutions, rather than leaving them buried in citation databases, is

work we would welcome doing alongside NIH. Public impact also depends on infrastructure that connects research to the people who can act on it. Frontiers Science House convened over 100 researchers, policymakers, industry leaders and Nobel laureates from 16 countries alongside the World Economic Forum's January 2026 Annual Meeting in Davos, translating findings into conversations with the people who set policy and allocate capital. That kind of engagement is the credible middle step between publication and public benefit, and NIH's framework should treat structured translation activity of this kind as evidence of impact.

Other Areas

Frontiers urges NIH to state plainly, in whatever framework results from this RFI, that publication and dissemination are not adjacent to NIH's mission. They are how that mission is completed. NIH's own public access, reproducibility and gold standard science initiatives all state this in different words already.

OSTP's July 2026 report raises pointed questions about how research is published, reviewed and rewarded, including the risk that publication metrics "fall to Goodhart's Law as gameable targets" as AI makes it easier to generate plausible-looking research at scale. Frontiers shares that concern and welcomes debate about possible approaches. The scholarly ecosystem exists to enable collaboration and scrutiny, and we look forward to modernizing it in the direction OSTP describes, particularly around AI. Part of what the report asks for is already how we operate. Every Frontiers article is open to every reader immediately, with no embargo, and the evidence on open access is consistent: once age and discipline are accounted for, openly available articles receive around 18% more citations than average, and open outputs are reused and built on more widely. Extending that same openness, review, and curation to data, code, and methods is the practical path to the system the report envisions, and it is a path we would like to take with NIH.

In closing

Reform of narrow bibliometrics is needed but should not be read as meaning that publication itself becomes dispensable. Correcting the over-reliance on publication counts and impact factors allows a richer picture of what publication does, supporting peer review, transparency, correctability, and reuse. Publication is essential for sharing and building on research outputs, and no aspect of research benefits from a smaller role for publication in the research lifecycle. That richer picture of the roles for publication is still being drawn, and Frontiers wants to help draw it with NIH rather than working in parallel.

Frontiers, as a fully open-access publisher operating since 2007, would welcome a direct conversation with NIH on how Frontiers' scientific-impact framework, data and AI programs, as well as its costs and access policies can pull in the directions that match NIH's goals.

Respectfully submitted,

Christopher Charles Erdmann, Head of Public Affairs, North America – christopher.erdmann@frontiersin.org
Stephan Kuster, Director of External Affairs – stephan.kuster@frontiersin.org
Eleonora Colangelo, Policy Lead – External Affairs – eleonora.colangelo@frontiersin.org

221. Alexei Morozov

I am responding: On behalf of myself

Name: Alexei Morozov

Name of Organization: Fralin Biomedical Research Institute, Virginia Tech

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: A) To identify findings that are ripe for replication, NIH needs to solicit proposals to replicate published research. The proposals must be evaluated for their merit and any other NIH-funded grants. Thematic Special Emphasis Panels can evaluate such proposals based on a) replication need, b) qualifications of the applicant, and b) potential conflicts of interest between the authors of the original publication and the replication applicant.

B) To enforce rigor and reproducibility, NIH must engage with post-publication reviewing systems, such as PubPeer. Any PubPeer comment on NIH-supported studies must be evaluated by the Program Officers responsible for the award linked to the publication in question.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: The number of trained PhD students is an important criterion for a PI promotion at many universities. However, the quality of PhD studies is evaluated only by the thesis committee, hand-picked by the PhD advisor, interested in graduating their students. The system guarantees a thesis defense regardless of research outcomes. As a result, PhD programs often recruit unprepared students and produce poorly trained PhDs, while using NIH funds. NIH can improve the situation by requiring more rigid graduation criteria for NIH-supported PhD students, at a minimum, a couple of published first-author papers.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: Moving away from percentile scoring by study sections, which punishes high-risk proposals, to evaluations by IC programs will give the high-potential proposals and ideas a better chance.

7. Public Impact: Two things can help NIH earn public trust. 1) NIH needs to purge the research enterprise of administrators whose business models focus on recruiting investigators with high promises of bringing in grants and high indirect costs rather than forming unique research schools. 2) Research misconduct, particularly data manipulation, must result in severe punishment, and such cases must be broadly publicized to convince the public that NIH aims to eradicate such behaviors.

8. Other Areas: N/A

222. The 3Rs Collaborative

I am responding: On behalf of an organization

Name: Megan LaFollette

Name of Organization: The 3Rs Collaborative

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: The 3Rs Collaborative (3RsC) recommends that NIH treat the 3Rs of animal research, replacement, reduction, and refinement, as infrastructure supporting rigorous, reproducible, and replicable biomedical research rather than solely as an animal welfare framework. Together, they help define the best methods to answer a scientific question, including model selection, study design, and experimental conditions. In all cases, models should be evaluated for fit for purpose, studies designed for rigor and robustness, and methods reported transparently.

Refinement can directly improve rigor and reproducibility by reducing hidden sources of biological variability. Housing, handling, and husbandry practices that induce chronic stress may alter neuroendocrine, metabolic, and immune responses, creating unrecognized confounding variables. We recommend NIH fund research evaluating the impact of refinement practices on data variability and require reporting of husbandry, enrichment, and handling variables. Potential indicators include the proportion of funded studies reporting these variables and reductions in baseline variability across laboratories adopting evidence-based refinement practices such as provision of species-specific enrichment such as sufficient nesting material for mice and tunnel or cup handling of mice.

Reduction is similarly aligned with rigorous study design. Reduction means using the minimum number of animals necessary for a valid experimental and statistical design to answer a justifiable question. NIH should continue promoting randomization, blinded outcome assessment, prespecified outcomes, sample size justification, and consideration of sex as a biological variable such as requiring the Sex Inclusive Research Framework (https://openinnovation.astrazeneca.com/preclinical-research/sex-inclusive-research-framework.html) in applications to help NIH reviewers determine when single sex justifications are valid or not (Landis et al., Nature, 2012). NIH should also support research evaluating model validity and approaches that balance increasing external validity (such as adding deliberate heterogenization) with efficient study design (Voelkl et al., PLOS Biology, 2018). Reviewer training in experimental design, validity, and translation, as well as recruiting statisticians to sit in on review panels, is equally important, since reviewer expectations strongly influence research behavior.

The 3Rs Collaborative's 3Rs Certificate Course, which teaches key concepts in replacement, reduction, and refinement, could be adopted nationally as part of infrastructure to support the reproduction and replication of animal research.

A key opportunity for NIH is to strengthen adoption and enforcement of existing standards. PREPARE, ARRIVE, and the recently released DRIVER guidelines provide established frameworks for planning and reporting in vivo and in vitro research (Smith et al., Laboratory Animals, 2018; Percie du Sert et al., PLOS Biology, 2020; NC3Rs 2026). We recommend NIH move beyond endorsement and incorporate verification of compliance through progress reporting, audits, and renewal review.

To support reproducibility across research models, NIH should require standardized reporting of cell source and animal-derived reagents (as required by the DRIVER guidelines), including source, species, and lot information, and support development of chemically defined, animal-free reagents (https://nc3rs.org.uk/3rs-resource-library/moving-animal-free-scaffolds-construction-3d-organotypic-models). Such reporting and development would reduce an important but underappreciated source of experimental variability and validity.

Finally, NIH should support replication studies by prioritizing findings with substantial downstream scientific or translational impact and recognizing replication efforts as valuable scientific contributions.

2. Data, Software, and Model Sharing: 3RsC recommends that NIH incentivize standardized data templates and metadata practices aligned with FAIR principles and existing reporting standards. Standardization enables meaningful data reuse and reduces unnecessary duplicative experimentation by helping investigators identify questions that have already been addressed through robust evidence.

NIH should also support mechanisms for sharing control datasets and developing validated approaches for virtual or historical controls, where scientifically appropriate. Reuse of existing data has significant potential to improve efficiency while reducing unnecessary animal use.

We further recommend NIH establish safe-harbor mechanisms that encourage publication of negative results, failed replications, and unsuccessful model-validation studies. These findings are rarely shared despite their considerable value in guiding future research decisions and preventing duplication of effort.

Importantly, NIH should reward not only data deposition but also demonstrated reuse. Indicators might include standards-compliant dataset submission rates, documented dataset reuse, use of shared controls, publication of negative findings, and evidence that shared resources informed subsequent studies. Existing NIH-supported collaborative programs, including Complement-ARIE, ORIVA, the Technology Development Centers, and the FNIH NAMs Validation and Qualification Network, provide useful models for this approach.

3. Training and Mentorship: The 3RsC considers training and mentorship the areas where NIH has the most leverage. Most failures of rigor originate upstream of execution, in how the question was framed and whether it needed answering. Reporting checklists do not catch this, since a checklist assumes the study was worth doing.

NIH should support training programs that build interdisciplinary competency across comparative medicine, animal science, computational biology, experimental design, statistics, and regulatory science rather than training researchers within methodological silos. The ability to evaluate and select the most appropriate approach for a scientific question is increasingly important in modern biomedical research.

This includes re-training for established faculty on the most up to date methods.

In addition to technical skills, training programs should emphasize question formulation, critical appraisal of the literature, model evaluation, harm-benefit analysis, and study design principles that support reproducibility and translational success. These competencies influence research quality before any experiment begins. Appropriate scientific question formulation helps drive key experimental constraints and hypothesis generation. Critical appraisal of the literature, including systematic searching, ensures investigators do not repeat settled experiments or build programs on fragile findings. Model evaluation is essential to ensure the selected model is able to answer the scientific questions at hand, while considering both model validity and translation. Harm-benefit analysis should be conducted during design rather than as a form completed for a committee, since institutional review performs a version of this analysis only after the design is fixed.

As a measurable indicator, NIH could track adoption of standardized 3Rs curricula such as the 3Rs Collaborative's 3Rs Certificate Course as part of Responsible Conduct of Research. Course completion should be paired with downstream outcomes including reporting quality, study design practices, and research reproducibility rather than relying solely on participation metrics. Additionally, to ensure workforce stability, retention should be tracked of all key personnel, including not only investigators but also key support staff such as comparative medicine employees.

We recommend NIH fund T32, K, and professional development mechanisms building competency across comparative medicine, animal science, computational biology, statistics, and regulatory science. Workforce indicators should also recognize veterinarians, animal care professionals, technicians, and IACUC personnel, whose expertise is essential for implementation of the 3Rs and overall research quality.

4. Collaboration: The 3Rs Collaborative (3RsC) convenes academia, industry, technology providers, contract research organizations, and federal agencies, and comments from that experience.

As biomedical research becomes increasingly interdisciplinary, NIH should recognize collaboration that bridges traditionally separate communities, including animal researchers, developers of new approach methodologies (LaFollette et al, 2025, doi: 10.1177/02611929241310566) , computational scientists, clinicians, and regulatory experts. Scientific progress is strengthened when these groups work toward shared objectives rather than operating in isolation.

Potential indicators include cross-sector and cross-disciplinary co-authorship, participation in collaborative consortia, contribution to shared resources, and diverse representation on study sections and advisory groups. NIH should also encourage use of contribution taxonomies that better recognize activities such as model validation, method development, data curation, and technical expertise.

Because incentive structures drive behavior, collaboration should be recognized in biosketches, progress reports, and review criteria. Special attention should be paid to ensuring that early-career investigators receive appropriate credit for team-science contributions.

NIH should also be attentive to collaboration health: harassment, vilification, or dismissal of investigators

based on the methods they use undermines both scientific progress and public trust, while driving investigators away from key methodological conversation. NIH-funded infrastructure providing neutral, trusted convening space merits recognition and support, and 3RsC would welcome expanded partnership with NIH in that capacity.

5. Entrepreneurship and Translation: Indicators of translational success should focus on a model or method's demonstrated ability to improve decision-making and reduce costly late-stage failures in research and development. This principle applies equally to animal-based models, human-based models, NAMs, and integrated approaches alike.

NIH should support systematic evaluation of where specific models perform well, where limitations exist, and where new approaches may offer complementary or superior predictive value for defined contexts of use. Context of use is the operative concept: a method is valid for a defined question, in a defined biological system, against defined endpoints, rather than valid in general. Characterization and validation efforts, including public-private collaborations, are essential infrastructure for scientific translation. This includes funding characterization and validation studies such as the 3RsC's drug-induced liver injury (DILI) project with the FDA and 8 commercial providers (LaFollette et al. 2026, doi: https://doi.org/10.1016/j.yrtph.2026.106189).

Incentivizing entrepreneurship. Method and platform development is frequently commercialized through small companies, and the academic reward structure gives limited credit for it. We recommend NIH recognize entrepreneurial contribution directly through SBIR and STTR participation, spinouts and licensing arising from method development, and adoption of validated methods by other laboratories or by industry and encourage institutions to count these as scholarly output in promotion review.

6. Foundational Scientific Exploration: The 3Rs Collaborative (3RsC) notes that basic and high-risk research characterizing where models, including animal models, diverge from human physiology is chronically underfunded, because it yields no immediate product even though it establishes the evidentiary base later validation depends on. The same holds for early-stage validation science, which carries high scientific risk, long timelines, and low publication yield.

We recommend NIH protect funding for domain-mapping research and early-stage validation and treat both as infrastructure. Indicators include portfolio share devoted to model characterization and early validation, time from characterization to defined context-of-use qualification, and downstream dependence of applied programs on characterization outputs.

We recommend NIH state explicitly in review guidance that rigorous negative characterization and completed validation studies count as productivity, encourage institutions to weight them in promotion and tenure review, and support protected time for investigators carrying long-timeline characterization work. Institutions follow what NIH scores, so guidance to review panels is the most direct available lever.

7. Public Impact: We recommend that NIH publicly report portfolio-level indicators of 3Rs implementation, including use of valid models and designs, extent of refinement adoption, and replacement efforts, where appropriate. Transparent reporting of these indicators, along with publication of negative findings and validation studies, would help demonstrate NIH's commitment to rigorous, accountable science.

3Rs implementation bears on the dimensions NIH names. On health outcomes and clinical practice, methods validated for a defined context of use reduce the late-stage failures that delay therapies reaching patients. On economic growth, avoided duplicative studies, reusable datasets, and reduced late-stage attrition represent recoverable cost, and reproducibility failures carry a documented and substantial price (Freedman et al., PLOS Biology, 2015). On accessibility, standardized protocols, shared control data, and free reporting instruments lower the barrier for institutions without large research infrastructure. On public trust and societal benefit, the public has a durable interest in how research animals are treated, and 3Rs progress is among the few research-quality measures the public can readily understand and evaluate.

On recognition rather than assessment, we recommend NIH report these indicators publicly at portfolio level and name 3Rs progress explicitly when describing research quality, since what an agency reports on is what institutions come to prioritize.

To avoid unintended consequences, we recommend that indicators for all 3Rs should be reported and interpreted together rather than individually. Public indicators should also be normalized by institution size and award volume, since raw comparisons would penalize smaller institutions without dedicated 3Rs staff. For example, only counting animal numbers could result in underpowered studies or studies moved to a different institution, so reduction should include indicators for design-quality.

8. Other Areas: We encourage NIH to recognize progress on the 3Rs themselves (validated replacement, demonstrated reduction, and measurable refinement) as an important category of scientific impact. Progress in the 3Rs directly supports the rigor, reproducibility, efficiency, and translational value that NIH seeks to promote through this initiative. It is inseparable from the rigor, reproducibility, and translational value this RFI seeks to reward, and currently the dimension of research quality least visible in NIH's indicators.

Tracking it at portfolio scale calls for a body positioned to develop standards, deliver training, curate shared data, and measure adoption nationally. The United Kingdom has had one for two decades in the National Centre for the Replacement, Refinement and Reduction of Animals in Research, (NC3Rs) which supplies training, data resources, funding calls, guideline development, and cross-sector convening, and which developed the DRIVER guidelines released in July 2026. The United States has no equivalent, so 3Rs progress here is distributed across individual institutions, investigators, and voluntary effort, which makes it difficult to capture in any portfolio-level metric.

We recommend NIH support establishing an independent, U.S.-based 3Rs center on that model, and expand federal partnership with and funding for The 3Rs Collaborative in that capacity. 3RsC is a nonpartisan nonprofit working across all three Rs and already collaborates with NIH and FDA through the FNIH NAMs Validation and Qualification Network and through federal representation on its board and initiatives. We have developed a 3Rs Certificate Course suitable for wider adoption, conducted validation work including a DILI project with FDA and eight commercial providers (LaFollette et al. 2026, doi: https://doi.org/10.1016/j.yrtph.2026.106189), and convene cross-sector dialogue spanning academia, industry, technology providers, contract research organizations, and regulators.

We thank NIH for the opportunity to comment and would welcome the opportunity to work with the agency on the indicators described in this response.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/2026_NIH_RFI_Scientific-Impact.pdf

Description: The 3Rs Collaboratives Response (Formatted)

The 3Rs Collaborative's Response to NIH's RFI on Measuring and Rewarding Scientific Input

Background: The National Institutes of Health (NIH) recently launched a new agency-wide Replication and Reproducibility Initiative to incentivize and institutionalize replication and reproducibility efforts across the biomedical research enterprise. NIH aims to incentivize and reward rigorous research that delivers value to the American public, but current metrics of scientific success rely heavily on individual-level indicators of productivity, such as publication counts and citation metrics. NIH seeks to better align its measures of success with its core mission; generating knowledge that enhances health, lengthens life, and reduces illness and disability.

To support these efforts, NIH has issued a Request for Information which requests stakeholders to identify specific, measurable indicators and incentives that reflect the collaborative, rigorous, and mission-drive nature of modern biomedical science.

While NIH welcomes all comments, we are specifically interested in receiving feedback related to topics listed below. NIH is particularly interested in comments that include specific examples of measurable indicators, implementation approaches, potential benefits or unintended consequences, and considerations for feasibility across career stages, disciplines, and institution types.

Box 1: Rigor and Reproducibility. Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings.

The 3Rs Collaborative (3RsC) recommends that NIH treat the 3Rs of animal research, replacement, reduction, and refinement, as infrastructure supporting rigorous, reproducible, and replicable biomedical research rather than solely as an animal welfare framework. Together, they help define the best methods to answer a scientific question, including model selection, study design, and experimental conditions. In all cases, models should be evaluated for **fit** for purpose, studies designed for rigor and robustness, and methods reported transparently.

Refinement can directly improve rigor and reproducibility by reducing hidden sources of biological variability. Housing, handling, and husbandry practices that induce chronic stress may alter neuroendocrine, metabolic, and immune responses, creating unrecognized confounding variables. We recommend NIH fund research evaluating the impact of refinement practices on data variability and require reporting of husbandry, enrichment, and handling variables. Potential indicators include the proportion of funded studies reporting these variables and reductions in

baseline variability across laboratories adopting evidence-based refinement practices such as provision of species-specific enrichment such as sufficient nesting material for mice and tunnel or cup handling of mice.

Reduction is similarly aligned with rigorous study design. Reduction means using the minimum number of animals necessary for a valid experimental and statistical design to answer a justifiable question. NIH should continue promoting randomization, blinded outcome assessment, prespecified outcomes, sample size justification, and consideration of sex as a biological variable such as requiring the Sex Inclusive Research Framework (https://openinnovation.astrazeneca.com/preclinical-research/sex-inclusive-research-framework.html) in applications to help NIH reviewers determine when single sex justifications are valid or not (Landis et al., Nature, 2012). NIH should also support research evaluating model validity and approaches that balance increasing external validity (such as adding deliberate heterogenization) with efficient study design (Voelkl et al., PLOS Biology, 2018). Reviewer training in experimental design, validity, and translation, as well as recruiting statisticians to sit in on review panels, is equally important, since reviewer expectations strongly influence research behavior.

The 3Rs Collaborative's 3Rs Certificate Course, which teaches key concepts in replacement, reduction, and refinement, could be adopted nationally as part of infrastructure to support the reproduction and replication of animal research.

A key opportunity for NIH is to strengthen adoption and enforcement of existing standards. PREPARE, ARRIVE, and the recently released DRIVER guidelines provide established frameworks for planning and reporting in vivo and in vitro research (Smith et al., Laboratory Animals, 2018; Percie du Sert et al., PLOS Biology, 2020; NC3Rs 2026). We recommend NIH move beyond endorsement and incorporate verification of compliance through progress reporting, audits, and renewal review.

To support reproducibility across research models, NIH should require standardized reporting of cell source and animal-derived reagents (as required by the DRIVER guidelines), including source, species, and lot information, and support development of chemically defined, animal-free reagents (https://nc3rs.org.uk/3rs-resource-library/moving-animal-free-scaffolds-construction-3d-organotypic-models). Such reporting and development would reduce an important but underappreciated source of experimental variability and validity.

Finally, NIH should support replication studies by prioritizing findings with substantial downstream scientific or translational impact and recognizing replication efforts as valuable scientific contributions.

BOX 2: Data, Software, and Model Sharing. **Sharing scientific data and tools accelerates biomedical research discovery, in part, by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies. NIH seeks to incentivize and reward these activities to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.**

3RsC recommends that NIH incentivize standardized data templates and metadata practices aligned with FAIR principles and existing reporting standards. Standardization enables meaningful data reuse and reduces unnecessary duplicative experimentation by helping investigators identify questions that have already been addressed through robust evidence.

NIH should also support mechanisms for sharing control datasets and developing validated approaches for virtual or historical controls, where scientifically appropriate. Reuse of existing data has significant potential to improve efficiency while reducing unnecessary animal use.

We further recommend NIH establish safe-harbor mechanisms that encourage publication of negative results, failed replications, and unsuccessful model-validation studies. These findings are rarely shared despite their considerable value in guiding future research decisions and preventing duplication of effort.

Importantly, NIH should reward not only data deposition but also demonstrated reuse. Indicators might include standards-compliant dataset submission rates, documented dataset reuse, use of shared controls, publication of negative findings, and evidence that shared resources informed subsequent studies. Existing NIH-supported collaborative programs, including Complement-ARIE, ORIVA, the Technology Development Centers, and the FNIH NAMs Validation and Qualification Network, provide useful models for this approach.

BOX 3: Training and Mentorship. **NIH recognizes that advancing its mission requires promoting the growth and stability of the biomedical research workforce. While NIH supports training and protects time through existing policies and programs, additional indicators are needed to assess and quantify the impact of these efforts.**

The 3RsC considers training and mentorship the areas where NIH has the most leverage. Most failures of rigor originate upstream of execution, in how the question was framed and whether it

needed answering. Reporting checklists do not catch this, since a checklist assumes the study was worth doing.

NIH should support training programs that build interdisciplinary competency across comparative medicine, animal science, computational biology, experimental design, statistics, and regulatory science rather than training researchers within methodological silos. The ability to evaluate and select the most appropriate approach for a scientific question is increasingly important in modern biomedical research. This includes re-training for established faculty on the most up to date methods.

In addition to technical skills, training programs should emphasize question formulation, critical appraisal of the literature, model evaluation, harm-benefit analysis, and study design principles that support reproducibility and translational success. These competencies influence research quality before any experiment begins. Appropriate scientific question formulation helps drive key experimental constraints and hypothesis generation. Critical appraisal of the literature, including systematic searching, ensures investigators do not repeat settled experiments or build programs on fragile findings. Model evaluation is essential to ensure the selected model is able to answer the scientific questions at hand, while considering both model validity and translation. Harm-benefit analysis should be conducted during design rather than as a form completed for a committee, since institutional review performs a version of this analysis only after the design is fixed.

As a measurable indicator, NIH could track adoption of standardized 3Rs curricula such as the 3Rs Collaborative's 3Rs Certificate Course as part of Responsible Conduct of Research. Course completion should be paired with downstream outcomes including reporting quality, study design practices, and research reproducibility rather than relying solely on participation metrics. Additionally, to ensure workforce stability, retention should be tracked of all key personnel, including not only investigators but also key support staff such as comparative medicine employees.

We recommend NIH fund T32, K, and professional development mechanisms building competency across comparative medicine, animal science, computational biology, statistics, and regulatory science. Workforce indicators should also recognize veterinarians, animal care professionals, technicians, and IACUC personnel, whose expertise is essential for implementation of the 3Rs and overall research quality.

BOX 4: Collaboration. As biomedical research becomes increasingly interdisciplinary and convergent, collaboration is growing in both scale and complexity. NIH seeks new approaches for assessing and recognizing contributions to team-based science.

The 3Rs Collaborative (3RsC) convenes academia, industry, technology providers, contract research organizations, and federal agencies, and comments from that experience.

As biomedical research becomes increasingly interdisciplinary, NIH should recognize collaboration that bridges traditionally separate communities, including animal researchers, developers of new approach methodologies (LaFollette et al, 2025, doi: 10.1177/02611929241310566) , computational scientists, clinicians, and regulatory experts. Scientific progress is strengthened when these groups work toward shared objectives rather than operating in isolation.

Potential indicators include cross-sector and cross-disciplinary co-authorship, participation in collaborative consortia, contribution to shared resources, and diverse representation on study sections and advisory groups. NIH should also encourage use of contribution taxonomies that better recognize activities such as model validation, method development, data curation, and technical expertise.

Because incentive structures drive behavior, collaboration should be recognized in biosketches, progress reports, and review criteria. Special attention should be paid to ensuring that early-career investigators receive appropriate credit for team-science contributions.

NIH should also be attentive to collaboration health: harassment, vilification, or dismissal of investigators based on the methods they use undermines both scientific progress and public trust, while driving investigators away from key methodological conversation. NIH-funded infrastructure providing neutral, trusted convening space merits recognition and support, and 3RsC would welcome expanded partnership with NIH in that capacity.

BOX 5: Entrepreneurship and Translation. To advance the translation of discoveries into products for people, NIH fosters the maturation of discoveries from bench to bedside and promotes innovation across sectors. More can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission.

Indicators of translational success should focus on a model or method's demonstrated ability to improve decision-making and reduce costly late-stage failures in research and development. This principle applies equally to animal-based models, human-based models, NAMs, and integrated approaches alike.

NIH should support systematic evaluation of where specific models perform well, where limitations exist, and where new approaches may offer complementary or superior predictive value for defined contexts of use. Context of use is the operative concept: a method is valid for a defined question, in a defined biological system, against defined endpoints, rather than valid in general. Characterization and validation efforts, including public-private collaborations, are essential infrastructure for scientific translation. This includes funding characterization and validation studies such as the 3RsC's drug-induced liver injury (DILI) project with the FDA and 8 commercial providers (LaFollette et al. 2026, doi: https://doi.org/10.1016/j.yrtph.2026.106189).

Incentivizing entrepreneurship. Method and platform development is frequently commercialized through small companies, and the academic reward structure gives limited credit for it. We recommend NIH recognize entrepreneurial contribution directly through SBIR and STTR participation, spinouts and licensing arising from method development, and adoption of validated methods by other laboratories or by industry and encourage institutions to count these as scholarly output in promotion review.

BOX 6: Foundational Scientific Exploration. Basic research and high-risk, high-reward initiatives represent foundational scientific exploration with an inherent risk of failure but the potential for paradigm-shifting breakthroughs. Further embedding these approaches within the academic enterprise is essential to ensuring continued support for bold and innovative research directions.

The 3Rs Collaborative (3RsC) notes that basic and high-risk research characterizing where models, including animal models, diverge from human physiology is chronically underfunded, because it yields no immediate product even though it establishes the evidentiary base later validation depends on. The same holds for early-stage validation science, which carries high scientific risk, long timelines, and low publication yield.

We recommend NIH protect funding for domain-mapping research and early-stage validation and treat both as infrastructure. Indicators include portfolio share devoted to model characterization and early validation, time from characterization to defined context-of-use qualification, and downstream dependence of applied programs on characterization outputs.

We recommend NIH state explicitly in review guidance that rigorous negative characterization and completed validation studies count as productivity, encourage institutions to weight them in promotion and tenure review, and support protected time for investigators carrying long-timeline characterization work. Institutions follow what NIH scores, so guidance to review panels is the most direct available lever.

BOX 7: Public Impact. NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.

We recommend that NIH publicly report portfolio-level indicators of 3Rs implementation, including use of valid models and designs, extent of refinement adoption, and replacement efforts, where appropriate. Transparent reporting of these indicators, along with publication of negative findings and validation studies, would help demonstrate NIH's commitment to rigorous, accountable science.

3Rs implementation bears on the dimensions NIH names. On health outcomes and clinical practice, methods validated for a defined context of use reduce the late-stage failures that delay therapies reaching patients. On economic growth, avoided duplicative studies, reusable datasets, and reduced late-stage attrition represent recoverable cost, and reproducibility failures carry a documented and substantial price (Freedman et al., PLOS Biology, 2015). On accessibility, standardized protocols, shared control data, and free reporting instruments lower the barrier for institutions without large research infrastructure. On public trust and societal benefit, the public has a durable interest in how research animals are treated, and 3Rs progress is among the few research-quality measures the public can readily understand and evaluate.

On recognition rather than assessment, we recommend NIH report these indicators publicly at portfolio level and name 3Rs progress explicitly when describing research quality, since what an agency reports on is what institutions come to prioritize.

To avoid unintended consequences, we recommend that indicators for all 3Rs should be reported and interpreted together rather than individually. Public indicators should also be normalized by institution size and award volume, since raw comparisons would penalize smaller institutions without dedicated 3Rs staff. For example, only counting animal numbers could result in underpowered studies or studies moved to a different institution, so reduction should include indicators for design-quality.

BOX 8. Other Areas.

We encourage NIH to recognize progress on the 3Rs themselves (validated replacement, demonstrated reduction, and measurable refinement) as an important category of scientific impact. Progress in the 3Rs directly supports the rigor, reproducibility, efficiency, and translational value that NIH seeks to promote through this initiative. It is inseparable from the rigor, reproducibility, and translational value this RFI seeks to reward, and currently the dimension of research quality least visible in NIH's indicators.

Tracking it at portfolio scale calls for a body positioned to develop standards, deliver training, curate shared data, and measure adoption nationally. The United Kingdom has had one for two decades in the National Centre for the Replacement, Refinement and Reduction of Animals in Research, (NC3Rs) which supplies training, data resources, funding calls, guideline development, and cross-sector convening, and which developed the DRIVER guidelines released in July 2026. The United States has no equivalent, so 3Rs progress here is distributed across individual institutions, investigators, and voluntary effort, which makes it difficult to capture in any portfolio-level metric.

We recommend NIH support establishing an independent, U.S.-based 3Rs center on that model, and expand federal partnership with and funding for The 3Rs Collaborative in that capacity. 3RsC is a nonpartisan nonprofit working across all three Rs and already collaborates with NIH and FDA through the FNIH NAMs Validation and Qualification Network and through federal representation on its board and initiatives. We have developed a 3Rs Certificate Course suitable for wider adoption, conducted validation work including a DILI project with FDA and eight commercial providers (LaFollette et al. 2026, doi: https://doi.org/10.1016/j.yrtph.2026.106189), and convene cross-sector dialogue spanning academia, industry, technology providers, contract research organizations, and regulators.

We thank NIH for the opportunity to comment and would welcome the opportunity to work with the agency on the indicators described in this response.

223. SPARC

I am responding: On behalf of an organization

Name: Corinna Turbes

Name of Organization: SPARC

Type of Organization: Other

Type of Organization - Other: nonprofit

Role: Other

Role - Other: government affairs manager

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Response-to_-Request-for-Information-on-Measuring-and-Rewarding-Scientific-Impact-NOT-OD-26-087.pdf

SPARC
1333 New Hampshire Ave NW Suite 200 • Washington, DC 20036
(202) 630-5090 • www.sparcopen.org

Aug 19, 2026

Re: Comments on the Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Submitted electronically

SPARC, a non-profit organization of more than 190 libraries and academic organizations across the United States that supports research and education systems that enable everyone, everywhere to access, contribute to, and benefit from shared knowledge, appreciates the opportunity to comment on NIH's Request for Information on Measuring and Rewarding Scientific Impact.

SPARC commends NIH for using this RFI to reconsider how scientific success is measured. NIH's stated aim, to align its measures of success with its core mission of generating knowledge that improves health, is one SPARC strongly supports, and it is a natural complement to NIH's Replication and Reproducibility Initiative. SPARC also commends NIH for its leadership for early steps to give investigators room to describe and be credited for a broader range of research products, such as the recent Biosketch revisions.

Our comments focus on two recommendations that speak directly to that aim: NIH should (1) move away from proxy metrics that are poorly correlated with research quality, rigor, or impact, and (2) name and reward open research practices explicitly as indicators of scientific success.

To that end, SPARC recommends that NIH:

- Explicitly deprioritize publication counts and journal-level metrics, such as the Journal Impact Factor, as indicators of investigator quality, rigor, or impact in any indicators or guidance that result from this RFI, consistent with the San Francisco Declaration on Research Assessment (DORA).
- Recognize open, reusable data, software, and other research outputs as core indicators of scientific contribution.
- Require or incentivize the use of persistent identifiers for researchers, organizations, and research outputs across NIH-funded research.
- Encourage adoption of contribution-based credit frameworks in place of author-order-driven credit, particularly for team-based and interdisciplinary science.
- Favor non-proprietary, community-governed infrastructure over proprietary vendor metrics and platforms when NIH adopts tools to measure and report on the indicators identified through this RFI.

- Adopt a holistic, multi-indicator approach to research assessment and avoid overreliance on any single metric, however well-intentioned, since a single metric used as a target for reward or funding decisions is inherently vulnerable to gaming.

Below is a more in-depth look at the recommendations to help NIH reach its stated goals.

Move Away from Proxies

Publication counts, journal prestige, and related bibliometric indicators such as Journal Impact Factor are poor proxies for research quality, rigor, or real-word impact. The Journal Impact Factor, in particular, was not designed to evaluate individual researchers or individual articles; it was created as a tool for librarians making journal purchasing decisions, and it carries well-documented deficiencies as a measure of research quality.[1] That is why the San Francisco Declaration on Research Assessment (DORA), now signed by thousands of individual researchers and hundreds of research organizations and funders, calls on institutions not to use journal-based metrics as a surrogate for the quality of an individual scientist's contributions or in funding decisions.[2] Peer national funding agencies have already acted on this principle. Canada's tri-agencies, for example, are DORA signatories and have updated their peer review guidance to direct reviewers away from journal-based metrics and toward a broader, more qualitative assessment of contribution and impact.[3]

Continuing to rely on these proxies is not a neutral choice. It actively incentivizes the behavior NIH is trying to move away from, such as:

- **Rewarding volume over quality.** Counting publications creates pressure to fragment findings into the smallest publishable unit, rather than to produce fewer, more complete, and more rigorous studies.
- **Rewarding novelty over replication.** Journals and reviewers favor novel, positive findings, which discourages the replication and null-result reporting that NIH's own Replication and Reproducibility Initiative identifies as foundational to rigorous science.
- **Undervaluing collaboration.** Author-order and citation-based metrics were built for an era of small, single-investigator science and struggle to fairly credit the large, interdisciplinary, team-based projects that this RFI's Collaboration category identifies as increasingly central to biomedical research.

SPARC recommends that NIH explicitly deprioritize publication counts and journal-level metrics as indicators of investigator or research quality in any indicators or guidance that emerge from this RFI, and instead consider indicators that assess the research itself: its rigor, transparency, and reusability, rather than where or how often it was published.

Move Toward Open Research Practices

---

[1] Seglen, Per O. "Why the Impact Factor of Journals Should Not Be Used for Evaluating Research." *BMJ* 314, no. 7079 (1997): 498–502. https://pmc.ncbi.nlm.nih.gov/articles/PMC2126010/pdf/9056804.pdf
[2] San Francisco Declaration on Research Assessment, "Read the Declaration," https://sfdora.org/read/.
[3] Canadian Institutes of Health Research, “Responsible Research Assessment at CIHR” https://cihr-irsc.gc.ca/e/51731.html.

Open research practices directly support NIH's stated goals around reproducibility and data sharing, and this RFI is an opportunity for NIH to name these practices explicitly as indicators worth measuring and rewarding, rather than leaving them as implicit or secondary considerations.

Requiring and rewarding open accessibility of research outputs. Making data, code, software, protocols, and other research outputs openly available, not merely "available upon request," should be recognized as a core scientific contribution in its own right, not an afterthought to publication. Openly shared outputs are what make replication and reuse possible in the first place, and NIH should consider both requiring and then rewarding indicators that capture whether outputs are measurably and actively reusable.

Adoption of persistent identifiers. Consistent use of persistent identifiers (PIDs), including identifiers such as ORCID iDs for researchers, ROR IDs for organizations, and DOIs or ARKs for datasets and software makes contributions traceable, attributable, and machine-readable across the research lifecycle. Persistent identifiers are also the connective tissue that makes many of the other indicators in this RFI measurable at all. Without them, NIH cannot reliably track data reuse, software citation, or team contribution at scale. NIH is well positioned to require or strongly incentivize their use across their intramural research programs and the grants it funds.

Contribution-based credit instead of author-order-driven credit. Author order is a poor and increasingly outdated signal of who did what on a project, particularly for large, team-based, or interdisciplinary science. For example, the Contributor Roles Taxonomy (CRediT), which defines fourteen distinct contributor roles, offers a structured alternative that captures roles like data curation, software development, and project administration alongside traditional writing and conceptualization roles.[4] NIH's own Intramural Research Program already uses CRediT to adjudicate authorship disputes, resolving roughly a dozen such conflicts a year with a common vocabulary for contribution.[5] This is the type of credit system that NIH is well positioned to extend outward that can encourage structured, contribution statements across the research it funds so that credit reflects actual roles rather than position in an author list. Finding ways to both surface contributions as well as provide meaningful attribution (such as with the Open Exchange Architecture[6])would directly support NIH's interest in better recognizing collaboration and team science.

Use of non-proprietary, community-governed platforms and repositories. Reliance on proprietary vendor metrics and platforms concentrates control over how scientific value is defined in the hands of commercial entities whose incentives may not align with NIH's mission. As NIH adopts tools to track the indicators identified through this RFI, SPARC urges the agency

---

[4] NISO, "Contributor Role Taxonomy (CRediT)," https://credit.niso.org/.
[5] Partin, K. & Hosseini, M., "Using the contributor role taxonomy (CRediT) as a tool in resolving authorship disputes at the NIH," *Accountability in Research* (2025), https://doi.org/10.1080/08989621.2025.2596063
[6] Rowan Cockett, "Introducing the Open Exchange Architecture," OXA, January 28, 2026 https://oxa.dev/articles/introducing-oxa

to favor infrastructure, including repositories, identifier systems, and evaluation tools, that is non-commercial and openly governed by the scientific community, ensuring that the metrics used to evaluate NIH-funded science remain aligned with public, mission-driven goals rather than commercial ones.

Guard Against Overreliance on Any Single Metric

SPARC's recommendation calls on NIH to move away from proxy metrics and toward supporting open research communication practices. Any metric or indicator that NIH ultimately adopts through this process should be part of a holistic, multi-indicator approach to assessment rather than a singular number that stands in for scientific quality. A metric of data-sharing activity, persistent identifier use, or open-source contribution can be just as easily gamed as a citation count or a journal placement once researchers understand that it drives funding or advancement decisions. Compliance can become a matter of checking a box rather than genuinely advancing rigor, reproducibility, or reuse.

Consistent with that principle,[7] SPARC recommends that NIH:

- Use any new indicators that emerge from this RFI in combination with one another and alongside qualitative, expert-driven review, rather than allowing any single indicator to become a de facto target or threshold.
- Periodically revisit and revise the indicators it adopts, with community input, recognizing that any measure, once it becomes consequential, will attract efforts to satisfy it in form rather than substance.
- Be explicit, in any resulting guidance, that the goal is a fuller and more accurate picture of rigor, reproducibility, collaboration, and impact, not a replacement scorecard.

Conclusion

Taken together, these recommendations would give NIH a more direct, accurate, and durable way to assess the rigor, reproducibility, collaboration, and real-world impact of the science it funds.

The research community has already developed and tested frameworks for assessing contribution, openness, and quality more holistically, and elements of that work are already in use inside NIH's own Intramural Research Program. What is needed is for NIH to work closely with the community to adapt and formalize these frameworks as indicators of scientific success, used holistically and in combination, across the extramural research it funds. SPARC appreciates NIH's willingness to solicit input on this topic and urges the agency to adopt the recommendations above in any resulting policy or guidance.

Respectfully submitted,
Corinna Turbes,
Senior Manager of Government Affairs
SPARC

---

[7] Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). "Bibliometrics: The Leiden Manifesto for research metrics." *Nature*, 520(7548), 429-431, https://doi.org/10.1038/520429a

224. Bayh-Dole Coalition

I am responding: On behalf of an organization

Name: Joseph P. Allen

Name of Organization: Bayh-Dole Coalition

Type of Organization: Other

Role: Other

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Bayh-Dole-Coalition-Comments-on-NIH-RFI-re_-Measuring-Scientific-Impact-.pdf

August 19, 2026

Jay Bhattacharya, M.D., Ph.D.
Director, National Institutes of Health
9000 Rockville Pike
Bethesda, Maryland 20892

Re: Request for Information on Measuring and Rewarding Scientific Impact
[NOT-OD-26-087]

Dear Director Bhattacharya,

Thank you for the opportunity to respond to the National Institutes of Health's Request for Information on Measuring and Rewarding Scientific Impact.

The Bayh-Dole Coalition is a group of innovation-oriented organizations and individuals committed to celebrating and protecting the Bayh-Dole Act, a 1980 law that allows universities, small businesses, and federal laboratories to partner with the private sector to translate federally funded research into real-world products.

We applaud NIH for undertaking the important effort of accurately measuring the impact of federally funded research. Key academic metrics like publications and citations are important measures of scientific productivity and dissemination. But such metrics only tell part of the story—a promising finding may be published and widely cited without ever becoming a therapy, diagnostic tool, or product.

To understand a discovery's true impact, the NIH must consider whether a discovery ultimately reaches patients and consumers. We therefore urge the NIH to recognize commercialization and technology transfer metrics as key indicators of success.

Lawmakers have previously recognized the importance of technology transfer and commercialization. Before 1980, when the government retained the patent rights to federally funded discoveries, research institutions regularly made important breakthroughs—but

those discoveries rarely reached consumers. The government licensed less than 5% of its 28,000 patents to private companies. Promising discoveries were left on laboratory shelves.

In response, Congress passed the Bayh-Dole Act, which empowered research institutions to patent and license those federally funded discoveries. That framework for technology transfer and commercialization has since allowed thousands of federally funded discoveries to move from universities and federal labs into the private sector for development—and eventually, to reach patients and consumers.

The NIH ought to reward and preserve this framework. It should therefore recognize metrics that capture patenting efforts and industry collaboration, including issued patents and patent quality, licensing activity, startup and spinout formation, follow-on private investment, and industry partnerships.

It should also measure the longer term impacts of discoveries. Many of the most important biomedical breakthroughs take years—or decades—to reach patients even after licensing. The NIH should incorporate measures that capture downstream outcomes, such as clinical development milestones, FDA approvals, and adoption by patients, physicians, or healthcare systems. The NIH should also include longer-term economic indicators like job creation and regional growth.

Including such metrics will help ensure that researchers do not feel that patenting, licensing, entrepreneurship, and industry collaboration is viewed as less valuable than publishing another paper.

To that end, the NIH should also pay close attention to the growing use of Other Transaction Agreements (OTAs). While OTAs are an important part of the R&D system, designed to bring innovative companies into government projects without invoking the onerous requirements of the Federal Acquisition Regulations, it appears that some agencies are misusing them as a workaround the Bayh-Dole Act—denying academic institutions the rights guaranteed them by Congress which are critical for commercializing critical government funded medical research. That is particularly wrong headed at a time when China is going all out to eclipse our lead in the life sciences—a lead we enjoy because

Bayh-Dole helps transform NIH funded research into useful products protecting and improving lives here and around the world.

The Bayh-Dole Act transformed federally funded research by creating a pathway for universities, small businesses, and federal labs to turn early-stage discoveries into real products. The NIH's evaluation framework should similarly recognize that the ultimate measure of scientific impact is not only what researchers publish, but what reaches patients and grows the economy.

Thank you again for taking on this important effort, and for the opportunity to comment.

Sincerely,



Joseph P. Allen
Executive Director
Bayh-Dole Coalition

225. Association of Public and Land-grant Universities

I am responding: On behalf of an organization

Name: Kacy Redd

Name of Organization: Association of Public and Land-grant Universities

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/APLU-NIH-NOT-OD-26-087-final.pdf

Description: The Association of Public and Land-grant Universities (APLU) appreciates the opportunity to respond to NIH's Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087). APLU is a membership organization that fosters a community of university leaders collectively working to advance the mission of public research universities. Founded in 1887, APLU is the oldest higher education association in the United States and its U.S. membership consists of more than 240 public research universities, land-grant institutions, state university systems, and affiliated organizations spanning across all 50 states, the District of Columbia, and six U.S. territories.

ASSOCIATION OF PUBLIC & LAND-GRANT UNIVERSITIES

August 19, 2026

NIH Office of Science Policy
9000 Rockville Pike
Bethesda, Maryland 20892

Re: Request for Information (RFI) on Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Submitted online at https://osp.od.nih.gov/comment-form-measuring-and-rewarding-scientific-impact/

To Whom It May Concern,

The Association of Public and Land-grant Universities (APLU) appreciates the opportunity to respond to NIH's Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087). APLU is a membership organization that fosters a community of university leaders collectively working to advance the mission of public research universities. Founded in 1887, APLU is the oldest higher education association in the United States and its U.S. membership consists of more than 240 public research universities, land-grant institutions, state university systems, and affiliated organizations spanning across all 50 states, the District of Columbia, and six U.S. territories. The association and its members collectively focus on increasing student success and workforce readiness; promoting pathbreaking scientific research; and bolstering economic and community engagement. Annually, its U.S. member campuses enroll 4.6 million undergraduates and 1.4 million graduate students, award 1.3 million degrees, employ 1.3 million faculty and staff, and conduct $70 billion in university-based research.

Supporting Foundational Scientific Exploration

APLU supports NIH's objectives to incentivize and reward rigorous research that delivers value to the American public. The research enterprise, built on curiosity-driven and use-inspired research evaluated through merit-based peer review, has produced statins, MRI technology and diagnostics, early detection of pancreatic cancer, the Human Genome Project, and more. Each of these advancements depended on decades of foundational NIH

investment before the impact on human health was realized. The next generation of cures and technologies will require the same sustained commitment to foundational science and high-risk, high-reward initiatives. The enterprise that produced these results is worth protecting, and 21st-century reforms should build on what has made it so successful.

This commitment must span the full range of health-relevant science, including behavioral and social science research methods that determine whether treatments reach people. Treatment adherence and chronic disease management across populations depend on understanding human behavior and social determinants of health. Ensuring that new approaches to measuring scientific impact recognize this full range of contributions is essential to advancing NIH's mission.

Expanding how to evaluate scientific success

NIH seeks to align how it measures scientific success with its mission to generate knowledge that "enhances health, lengthens life, and reduces illness and disability." APLU agrees that important conversations are needed about better assessing research impacts that capture the full range of what the research community and broader public value. However, the problem is not a shortage of indicators; the research community has identified approximately 250 potential indicators across 10 domains of impact (Aurbach et al, 2025). The problem is that the research community has not converged on what constitutes impact or how to evaluate it consistently across the spaces where research evaluation occurs.

This is a significant consensus and coordination challenge. Researchers serve on tenure committees, journal editorial boards, NIH study sections, and disciplinary society conference panels. They make choices based on what they believe is valuable scholarship and what the system currently rewards. Those beliefs are shaped by the incentive structures of government, non-profit, and industry funders; departments and institutions; disciplinary societies; and publishers, and change to the incentive structure in one sector without corresponding change in the others will be ineffective. If institutions reform promotion and tenure, but researchers cannot be funded or published for translation work and their graduate students cannot build the publication records needed to compete on the job market, the reform will stall. If they can get funded but their work is not legible to their department, they may not get promoted or have access to graduate students. This is an interconnected evaluation system that requires cross-sector coordination to change, and NIH is well-positioned to catalyze that coordination.

If we can align reward and incentive structures across the spaces where research is evaluated, we can build a system that more quickly moves discoveries into treatments that

advance NIH's mission to enhance health, lengthen life, and reduce illness and disability.

Reducing researcher burden

Researchers have finite time, and every new expectation competes with existing obligations. Faculty are currently evaluated on being excellent in teaching, research, and service, but NIH's RFI could be interpreted to expect all researchers to also excel at team science, entrepreneurship, data sharing, mentorship, and public impact. Therefore, reforms must account for how they increase workload, provide researchers with the training and guidance to successfully meet new expectations, and reduce burden where possible.

Institutions face some of the same constraints and finite capacity. New requirements that assume infrastructure that does not yet exist will concentrate research capacity in fewer universities and fewer regions, the opposite of what a robust national research enterprise requires. NIH has the opportunity to use its influence to move the system, to work with the other federal agencies to harmonize policies, engage the research community to define what to measure, invest in the infrastructure researchers need to meet those expectations, and coordinate across sectors to align how research is evaluated and rewarded in all the contexts that matter to researchers. NIH cannot and should not do this alone. This is a system-wide challenge.

Given these considerations and constraints, APLU offers the following suggestions. APLU draws on its work developing Public Impact Research (PIR) as an organizing concept for how research serves society (APLU, 2019; Aurbach et al., 2023; Redd et al., 2024) and its experience leading cross-sector and multi-institution convenings.

Recommendation 1. Organize and fund cross-sector convenings around the seven RFI topic areas

As the framing above makes clear, no single sector can shift the evaluation system alone. NIH should organize cross-sector convenings that bring together researchers, institutional leaders, funders, disciplinary societies, publishers, and aligned national reform efforts to define mutual roles and obligations across the seven RFI topic areas.

NIH is already engaging the community in measurement design through the RFI and the S-index challenge for data sharing. APLU recommends expanding these engagements to multi-institution and cross-sector convenings. As an example, APLU's 2024 Fostering a Public Impact Research Ecosystem Workshop, co-convened with the Pew Charitable Trusts, brought sixty participants across eight sectors to identify where cross-sector alignment was needed and what each sector could do. The workshop produced

recommended actions for federal agencies, including establishing guidelines for including public impact research in grant proposals and project reports, expanding cross-agency partnerships to scale funding, and enhancing professional development for researchers engaged in public impact work (Redd et al, 2024). Similarly, the Accelerating Public Access to Research Data (APARD) initiative, co-sponsored by NIH and NSF and led by APLU and the Association of American Universities (AAU), worked with more than 100 institutions to produce aligned policies and practices for data governance, infrastructure, training, and culture change for data sharing (Redd et al, 2022). NIH could support similar convenings for rigor and reproducibility, entrepreneurship and translation, and the other areas identified in the RFI.

Recommendation 2. Engage NIH-relevant disciplinary societies to convene researchers

Academic fields set their disciplinary norms, and disciplinary societies are the venue where researchers convene to develop those norms, codify them into standards, and share them with the next generation of researchers in their field. NIH should work with these societies on interventions that advance each of the seven RFI areas of interest.

There is a model for this. In 2007, NSF, NIH, HHMI, and AAAS launched Vision and Change in Undergraduate Biology Education (AAAS, 2011). AAAS managed a series of regional conversations with biology educators, researchers, and administrators, followed by a national conference in 2009 that brought more than 500 participants together to build consensus. The resulting report established shared standards that the community adopted. Within four years, citations of Vision and Change in NSF education proposals rose from under 1% to as high as 67%, and professional societies built training programs and implementation networks around its recommendations (Vasaly et al., 2014). A 2013 follow-up conference assessed progress and identified remaining gaps.

This model is particularly needed for public impact research. APLU has worked with institutional leaders to put supportive policies and processes in place (APLU, 2019; Aurbach et al., 2023), but institutional infrastructure and policies alone do not produce culture change. Researchers need discipline-specific guidance on what public impact means in their field and how it is evaluated. NIH could fund disciplinary societies to develop that guidance, much as NSF funded AAAS to build consensus around undergraduate biology education reform. Complementary efforts, such as the NSF-funded Promotion and Tenure Innovation and Entrepreneurship (PTIE) initiative, are demonstrating similar approaches for research translation at the departmental level (Carter et al., 2021).

Public impact is one area where this approach is needed, but the model applies across all seven RFI topic areas. Ideas that might emerge from researcher-driven conversations include uncoupling data management and community partner costs from modular grant caps, resources to develop discipline-specific guidance for each area of interest, or supplemental funding for the different areas in the RFI. When NIH funds community-engaged research, data sharing infrastructure, or translational work, it sends an important signal to researchers and departments. Additional resources are a strong motivator for changing researcher practice, and they signal NIH values these practices.

Recommendation 3. Coordinate with other federal agencies to harmonize research requirements

NIH can help lead interagency coordination. Institutions cannot build bespoke recognition structures for each funder or discipline. Broader impacts, data sharing, and public engagement are each defined differently by each agency. If NIH, NSF, DoW, DoE, and USDA each define impact differently, institutions and researchers face an impossible compliance puzzle, one that has grown substantially: sixty-six percent of federal research requirements were issued in the last decade (COGR, 2026). This growing compliance burden impedes institutional change efforts that could focus resources on strengthening the research enterprise. NIH should lead interagency harmonization of research requirements across federal funders that could free up resources for accelerating science.

Additionally, if NIH's new approaches to evaluation are STEM-centric, they will fail to capture the full range of research with public benefit and will have limited adoption. Institutions evaluate all scholarly activity, and evaluation approaches that cannot represent disciplines beyond STEM, including arts, humanities, and social sciences, will be difficult for institutions to integrate. Coordinating with NSF, NEH, and other agencies increases the likelihood that new approaches are appropriate for all disciplines and scholarly venues, making them more likely to gain widespread adoption.

Recommendation 4. Build the shared infrastructure that rigorous, reproducible science requires

NIH should engage the research community on what national infrastructure investments are needed to accelerate science and ensure it is rigorous and reproducible. These might include data stewardship and curation services, secure compute enclaves for sensitive health data, prototyping facilities and translational cores, a PubMed equivalent for datasets, and increased access to compute, data and AI infrastructure. It should include additional investment in research infrastructure through existing mechanisms: C06, G20, and S10 programs for facilities and instrumentation. APLU further discussed investment in research infrastructure and deferred maintenance in its comments on the NIH Strategic

Plan. Without these investments, new expectations around data sharing, reproducibility, and public impact will disproportionately burden institutions with fewer resources, which risks concentrating research capacity in fewer institutions.

Recommendation 5. Pilot reform internally and evaluate outcomes
NIH should pilot measurement reform internally through its intramural research program before extending new indicators to the extramural community. NIH should apply scientific rigor to assessment reform itself. Monitor whether policy changes produce the outcomes sought and be transparent with the broader community about the lessons learned.

APLU has provided some suggested readings and resources from partners it collaborates with on PIR, entrepreneurship and translation, and data sharing. APLU looks forward to ongoing conversations with NIH as it further develops its plan to support a more robust research enterprise.

Sincerely,

Kacy Redd, PhD
Associate Vice President, Research & STEM Education,
Association of Public and Land-grant Universities

Curated Reference List

Rigor and Reproducibility
NASEM Report
National Academies of Sciences, Engineering, and Medicine. (2017). Fostering Integrity in Research. Washington, DC: The National Academies Press.

Data, Software, and Model Sharing
Accelerating Public Access to Research Data (APARD)
Redd, K., Smith, T., Miller, E., & Nusser, S. (2022). Guide to Accelerate Public Access to Research Data. Washington, DC: AAU and APLU. https://doi.org/10.31219/osf.io/tjybn
A joint AAU and APLU initiative, co-sponsored by NSF and NIH, that worked with 261 campus representatives from 111 institutions to develop aligned policies and practices for research data sharing. The resulting Guide covers governance, infrastructure, training, and culture change.

AAMC Credit for Data Sharing Initiative
A joint initiative of the AAMC, the Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard (MRCT Center), and the New England Journal of Medicine (NEJM) to develop and implement a system for crediting researchers who share scientific data. Full resources at https://www.aamc.org/about-us/mission-areas/biomedical-research/data-sharing

Public Impact
APLU's Public Impact Research Reports
APLU. (2019). Public Impact Research: Engaged Universities Making the Difference. Washington, D.C.: APLU. https://www.aplu.org/wp-content/uploads/public-impact-research-engaged-universities-making-the-difference.pdf
Developed with more than 60 university leaders, this report proposes PIR as a shared umbrella term for university research with societal benefit. Covers institutional approaches to PIR governance, community engagement, promotion and tenure reform, and communications. Includes a companion PIR Activation Guide at https://www.aplu.org/our-work/2-fostering-research-innovation/public-impact-research/public-impact-research-pir-activation-guide/

Aurbach, E.L., Kennedy, H., Redd, K., Bennett, J., & Gobstein, H. (2023, November). Modernizing Scholarship for the Public Good: An Action Framework for Public Research Universities. Association of Public and Land-grant Universities. DOI: https://osf.io/uekpb/
Supplementary materials are available at https://osf.io/42753/
An eight-area action framework for institutional leaders seeking to incentivize and recognize publicly engaged research. Areas span leadership development, promotion and tenure reform, institutional infrastructure, faculty capacity building, catalytic funding, and student training. Each area includes concrete steps by role and unit and university case studies.

Redd, K., Olneck-Brown, B., & Bennett, J. (2024). Fostering the Public Impact Research Ecosystem Workshop Report. Washington, DC: APLU. https://doi.org/10.31219/osf.io/rufv2
Report from a cross-sector convening co-hosted by APLU and The Pew Charitable Trusts, with more than 60 participants across eight sectors. Identifies five coordinated action priorities: national policy advocacy, shared PIR values, redesigned funding models, quality standards, and incentive alignment with action opportunities specific to federal agencies, funders, publishers, and universities.

Pew Charitable Trusts' Presidents and Chancellors Council on Public Impact Research
Pew convenes a Presidents and Chancellors Council on Public Impact Research. The Council works alongside Pew's Impact Funders Forum, a collaborative of more than 80 research funding organizations worldwide, to develop a sector-wide roadmap and calls to action for public impact research. It represents an active, senior-leadership coalition with direct reach into institutional decision-making.

Advancing the Societal Impacts of Research (ARIS)
ARIS is an NSF-funded center at the University of Missouri that builds capacity for researchers and institutions to measure and communicate research impact in communities. ARIS's Broader Impacts Toolkit supports researchers in addressing NSF's Broader Impacts criterion. Website: https://www.researchinsociety.org.

Civic Science Fellows Program
A network and fellows program that advances public engagement and brings together journalists, scientists, public engagement organizations, and funders to build collaboration between science and diverse communities. Funded by the Rita Allen Foundation, Kavli Foundation, Burroughs Wellcome Fund, and other philanthropic organizations.

Research Impact and Indicators Toolkit (RIITool)
Aurbach, E.L., Welzenbach, R., & Smith, J. (2024). Research Impact and Indicators Toolkit: Concept Brief and Responsible Use Guide. University of Michigan, Office of the Vice President for Research. https://deepblue.lib.umich.edu/items/f13c7b11-b3ce-4bda-a08f-02a2d896a062
A three-component toolkit: a Class x Domain impact matrix (10 impact classes, 7 societal domains), a database of approximately 250 published impact indicators mapped to those classes and to evaluation phase (design, delivery, or impact), and a responsible use workbook adapted from the SCOPE Framework. The database is accessible at https://airtable.com/appkeEEwjG1xubrko/shrhIb2noGPFuiYiL/tblvidlZxsWWbtFLm.

Training and Mentorship
NASEM Reports
National Academies of Sciences, Engineering, and Medicine. (2026). Rethinking Researcher Assessment and Incentives at U.S. Academic Institutions: Proceedings of a Workshop.
The 2026 workshop proceedings are particularly relevant as they address several of NIH's areas of interest in the RFI (public impact research, mentorship, team science, open access and data sharing, research integrity, and technology transfer).

National Academies of Sciences, Engineering, and Medicine. (2019). The Science of Effective Mentorship in STEMM. Washington, DC: The National Academies Press.

Collaboration and Team Science

NASEM Report

National Academies of Sciences, Engineering, and Medicine. (2025). The Science and Practice of Team Science. Washington, DC: The National Academies Press.

CRediT (Contributor Roles Taxonomy)

Hosseini, M., Kerridge, S., Allen, L., Kiermer, V., & Holmes, K. (2024). CRediT Roles and Example Research Tasks That Could be Attributed to Them. Zenodo. https://zenodo.org/records/18421449

A 14-role taxonomy approved as an ANSI/NISO standard, which makes contributor roles on collaborative and team science projects legible and citable. Website: https://credit.niso.org

Entrepreneurship and Translation

Promotion and Tenure Innovation and Entrepreneurship (PTIE)

Carter, R., et al. (2021). Innovation, Entrepreneurship, Promotion, and Tenure. Science, 374(6563). https://www.science.org/doi/10.1126/science.abj2098

An NSF-funded initiative led by Oregon State University, with a coalition of approximately 70 universities that focuses on recognizing innovation and entrepreneurship in promotion and tenure. Full resources at https://www.ptie.org. The PTIE Action Checklist for Supporting Translation: https://www.ptie.org/s/PTIE-Action-Checklist-for-Support-Translation-2026.pdf

Other Reform Efforts and Resources

COGR

COGR (2026). Changes in Federal Research Requirements Since 1991. https://www.cogr.edu/blog/changes-federal-research-requirements-1991

DORA

San Francisco Declaration on Research Assessment

DORA is a global initiative that advances practical and robust approaches to research and researcher assessment across all scholarly disciplines. DORA has more than 25,000 signatories. It has practical guides, case studies, training, and policy resources. Website: https://sfdora.org

Coalition for Advancing Research Assessment (CORA)
More than 900 signatories, primarily European universities, funders, and research organizations, who have signed a formal agreement committing to reform researcher evaluation and reduce overreliance on journal-based metrics. Website: https://coara.eu

Leiden Manifesto
Hicks, D., Wouters, P., Waltman, L., et al. (2015). Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520, 429-431. https://doi.org/10.1038/520429a. Ten principles for responsible use of quantitative indicators in research assessment, endorsed by institutions and funders worldwide.

Modernizing Academic Appointment and Advancement (MA3) Challenge
A challenge grant inviting colleges and universities to redesign academic hiring, review, promotion, and tenure processes to reward open science, public engagement, team science, and diverse contributions to knowledge. Organized by the Open Research Community Accelerator (ORCA) in partnership with the American Academy of Arts and Sciences and the Aspen Institute. https://www.ma3challenge.org/

Vision and Change
AAAS. (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: American Association for the Advancement of Science.

Vasaly, H.L., Feser, J., Lettrich, M.D., Correa, K., & Denniston, K.J. (2014). Vision and Change in the biology community: Snapshots of change. *CBE—Life Sciences Education*, 13, 16-20.https://www.lifescied.org/doi/10.1187/cbe.13-12-0234

226. Association of American Publishers

I am responding: On behalf of an organization

Name: J. Carl Maxwell

Name of Organization: Association of American Publishers

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Senior Vice President of Public Policy

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: The Association of American Publishers (AAP) welcomes this opportunity to provide suggestions for the NIH Request for Information on Measuring and Rewarding Scientific Impact. AAP represents over 80 professional and scholarly publishers, including dozens of domestic scholarly societies. Globally, technical, engineering, and medical (STEM) publishing is a $11 billion industry, a small but important part of the nearly $3 trillion worldwide research enterprise. AAP STEM members directly employee over 27,100 employees domestically, with over 1.6 million individual members of the global scientific and medical community.

The Role and Value of Scholarly Publishing

For two hundred years, STEM publishers have played a key role in the medium through which science is communicated, scrutinized, and ultimately used to support future research and cures. We manage peer review, curating and editing research outputs, disseminating publications and making them easily discoverable, while preserving the validated scientific record over time. Each published article is a data point in a cumulative, iterative scientific process, not an endpoint. Publishers help ensure that what enters that record has been vetted, is traceable, and can be corrected when necessary. Moreover, publishers create an entire research infrastructure to incorporate metadata, indexing, and other elements of the research process to enhance discoverability.

Most major research funders expect grantees to publish their findings in peer reviewed journals, even though this is not an explicit obligation of their grants, precisely because publication in such journals confers key benefits: amplification of results, formal dissemination to the relevant communities, validation and quality control through peer review and editorial processes, and durable integration into the literature. For researchers, journals are an important signaling mechanism, allowing readers to sort through the flood of information to find the true marker of scientific discovery and information.

Scientific publishing is not a luxury add on to research; it is integral to how research is communicated, scrutinized, and built upon. For this reason it is important that adequate and sustained funding continues to be available in NIH grants for publishing costs and subscriptions so that researchers can both contribute to and benefit from “the minutes of science.” Scientific publishing, when combined with

artificial intelligence and other evolving data and research products, stands at the forefront of a scientific revolution.

Ensuring the Integrity of Gold Standard Science

AAP members continually invest to ensure the quality and integrity of the scholarly record. They deploy plagiarism and image manipulation detection, AI generated content checks, and specialized integrity teams. They refine peer review and editorial processes and support robust systems for corrections, expressions of concern, retractions, disclosure of conflicts of interests, and sharing of data.

Publishers' work supporting the integrity, transparency and reproducibility of the scientific record is resource intensive and must be adequately funded in order to ensure the highest possible standards in research and science, and in order to ensure that taxpayers can rely on the accuracy of federally supported research.
Preprints, or pre-publisher manuscripts, illustrate both opportunity and risk. They are valuable for rapid scientific dialogue and early sharing of results, but because their trustworthiness is uncertain, they are not substitutes for peer reviewed articles. To non-specialist readers and the media, preprints can appear to be as authoritative as journal articles even though they have not been vetted and may have errors or preliminary interpretations. Government and research funder policies should recognize the distinct roles of preprints and peer reviewed publications and avoid treating them as interchangeable in metrics frameworks.

Understanding Existing Journal Metrics

Journal Impact Factors and citation metrics provide important, standardized signals to help researchers, institutions, and funders navigate an increasingly expansive global scientific landscape. By quantifying how often published work is referenced and built upon by peers, citation-based measures offer an objective, scalable proxy for academic reach, community engagement, and editorial selectivity. For publishers, a high-impact journal reflects a sustained investment in rigorous peer review, diligent editorial curation, and the long-term stewardship of notable research. These metrics serve as valuable benchmarks to help signal quality and contextualize a publication's role within its broader discipline and enabling readers and institutions to identify trusted venues where significant scientific dialogue is taking place.

At the same time, responsible publishers recognize citation metrics capture only one aspect of research value. Critical scientific advances—such as specialized clinical innovations, foundational methodological tools, or niche interdisciplinary breakthroughs—take time to accumulate citations or may deliver public benefits that raw citation counts cannot capture. Consequently, publishers increasingly support using Impact Factor and citation metrics as part of a holistic evaluation framework that pairs quantitative metrics with expert qualitative peer review to assess the full merit and real-world utility of scientific research.

Policy Risks to Avoid

Well intentioned policies can have unintended consequences if they overlook how the publishing ecosystem functions. Focusing on one output, rather than looking holistically at the scientific and medical community ecosystem, jeopardizes the quality and integrity of research. For example:
1) Treating preprints or other non-reviewed products as functional equivalents of peer reviewed articles for assessment purposes risks amplifying misinformation and devaluing scientific integrity.
2) Scoring various open science practices (article deposit, open-source code, open laboratory notebooks, FAIR data practices) as a measure of research quality rather than a measure of transparency.
3) Policies treating publication and access cost recovery as "diversions" of research funding, rather than as necessary and essential components of the research process, risk pushing systems toward the lowest cost, lowest integrity publications.

At a time of rising integrity challenges, including organized fraud and AI enabled manipulation, the answer is more robust peer review and editorial oversight, not less.

Recommendations for NIH

AAP supports continued robust investment in federal research as critical to the advancement of American science. The United States should cultivate the discoveries which have led our great nation to a position of leadership in health innovation and cures. Publishing is an important part of the scientific and medical ecosystem and is necessary to ensure we can overcome the challenges of the 21st century.

AAP offers the following recommendations to NIH as the agency considers updating metrics:

1. Prioritize integrity and quality in publication and reproducibility.
• Recognize peer reviewed publication and the Version of Record (VoR) as the primary, trusted record of research findings.
• Invest and fund research into reproducibility, including metrics encouraging researchers to pursue reproducibility studies and share a broad range of research inputs as additive to a professional career path.
• Develop metrics to support and encourage study pre-registration, study protocols, and registered reports to detail and establish research methods in advance of experimentation and data collection.
• Encourage grant-funding agencies to prioritize outputs with strong integrity safeguards in grant reviews and policy design.
• Support initiatives to expand and improve peer review, including metrics to reward peer review participation, training, and support for structured peer review.

2. Support sustainable public access.
• Preference appropriately licensed and funded VoR in public repositories and clearly label material that has not been peer reviewed or formally published.
• Ensure sustained and robust funding continues to be available for publishing costs as part of NIH grants. Avoid one size fits all mandates or metrics that undermine the financial sustainability of domestic non-profit scientific and medical societies and other high integrity publishers.

3. Preserve author's choice and intellectual property.

• Allow researchers to choose where and how to publish their work, including the licenses applied, within a framework that advances public health.
• Uphold robust intellectual property protections and contractual freedom as foundational to a vibrant, competitive publishing marketplace that can innovate in support of open science and AI era needs.

4. Partner with publishers on open science and AI.
• Collaborate with publishers to design metrics to encourage public access, data sharing, and AI related policies that use existing infrastructure for provenance, corrections, retractions, and integrity checks.
• Encourage the use of licensed, verified VoR and curated datasets as preferred inputs for high-stakes AI systems, recognizing publishers as part of the trust and safety infrastructure for scientific information.

A financially sustainable and collaborative open science system that engages the publishing community can be a powerful engine for research and innovation. AAP members stand ready to work with NIH to ensure the agency's investment in science advances human health and welfare, supports high-quality jobs, and strengthens U.S. leadership in discovery and innovation

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/2026.08.19-NIH-RFI-ResearchMetrics.pdf

Description: PDF of Association of American Publishers Response to RFI

AAP | ASSOCIATION OF AMERICAN PUBLISHERS

August 19, 2026

The National Institutes of Health
9000 Rockville Place
Bethesda, MD, 20892

RE: Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

To Whom It May Concern:

The Association of American Publishers (AAP) welcomes this opportunity to provide suggestions for the NIH Request for Information on Measuring and Rewarding Scientific Impact. AAP represents over 80 professional and scholarly publishers, including dozens of domestic scholarly societies. Globally, technical, engineering, and medical (STEM) publishing is a $11 billion industry, a small but important part of the nearly $3 trillion worldwide research enterprise. AAP STEM members directly employee over 27,100 employees domestically, with over 1.6 million individual members of the global scientific and medical community.

The Role and Value of Scholarly Publishing

For two hundred years, STEM publishers have played a key role in the medium through which science is communicated, scrutinized, and ultimately used to support future research and cures. We manage peer review, curating and editing research outputs, disseminating publications and making them easily discoverable, while preserving the validated scientific record over time. Each published article is a data point in a cumulative, iterative scientific process, not an endpoint. Publishers help ensure that what enters that record has been vetted, is traceable, and can be corrected when necessary. Moreover, publishers create an entire research infrastructure to incorporate metadata, indexing, and other elements of the research process to enhance discoverability.

Most major research funders expect grantees to publish their findings in peer reviewed journals, even though this is not an explicit obligation of their grants, precisely because publication in such journals confers key benefits: amplification of results, formal dissemination to the relevant communities, validation and quality control through peer review and editorial processes, and durable integration into the literature. For researchers, journals are an important signaling mechanism, allowing readers to sort through the flood of information to find the true marker of scientific discovery and information.

Scientific publishing is not a luxury add on to research; it is integral to how research is communicated, scrutinized, and built upon. For this reason it is important that adequate and sustained funding continues to be available in NIH grants for publishing costs and subscriptions so that researchers can both contribute to and benefit from "the minutes of science." Scientific publishing, when combined with artificial intelligence and other evolving data and research products, stands at the forefront of a scientific revolution.

Ensuring the Integrity of Gold Standard Science

AAP members continually invest to ensure the quality and integrity of the scholarly record. They deploy plagiarism and image manipulation detection, AI generated content checks, and specialized integrity teams. They refine peer review and editorial processes and support robust systems for corrections, expressions of concern, retractions, disclosure of conflicts of interests, and sharing of data.

Publishers' work supporting the integrity, transparency and reproducibility of the scientific record is resource intensive and must be adequately funded in order to ensure the highest possible standards in research and science, and in order to ensure that taxpayers can rely on the accuracy of federally supported research.

Preprints, or pre-publisher manuscripts, illustrate both opportunity and risk. They are valuable for rapid scientific dialogue and early sharing of results, but because their trustworthiness is uncertain, they are not substitutes for peer reviewed articles. To non-specialist readers and the media, preprints can appear to be as authoritative as journal articles even though they have not been vetted and may have errors or preliminary interpretations. Government and research funder policies should recognize the distinct roles of preprints and peer reviewed publications and avoid treating them as interchangeable in metrics frameworks.

Understanding Existing Journal Metrics

Journal Impact Factors and citation metrics provide important, standardized signals to help researchers, institutions, and funders navigate an increasingly expansive global scientific landscape. By quantifying how often published work is referenced and built upon by peers, citation-based measures offer an objective, scalable proxy for academic reach, community engagement, and editorial selectivity. For publishers, a high-impact journal reflects a sustained investment in rigorous peer review, diligent editorial curation, and the long-term stewardship of notable research. These metrics serve as valuable benchmarks to help signal quality and contextualize a publication's role within its broader discipline and

enabling readers and institutions to identify trusted venues where significant scientific dialogue is taking place.

At the same time, responsible publishers recognize citation metrics capture only one aspect of research value. Critical scientific advances—such as specialized clinical innovations, foundational methodological tools, or niche interdisciplinary breakthroughs—take time to accumulate citations or may deliver public benefits that raw citation counts cannot capture. Consequently, publishers increasingly support using Impact Factor and citation metrics as part of a holistic evaluation framework that pairs quantitative metrics with expert qualitative peer review to assess the full merit and real-world utility of scientific research.

Policy Risks to Avoid

Well intentioned policies can have unintended consequences if they overlook how the publishing ecosystem functions. Focusing on one output, rather than looking holistically at the scientific and medical community ecosystem, jeopardizes the quality and integrity of research. For example:

1) Treating preprints or other non-reviewed products as functional equivalents of peer reviewed articles for assessment purposes risks amplifying misinformation and devaluing scientific integrity.
2) Scoring various open science practices (article deposit, open-source code, open laboratory notebooks, FAIR data practices) as a measure of research quality rather than a measure of transparency.
3) Policies treating publication and access cost recovery as "diversions" of research funding, rather than as necessary and essential components of the research process, risk pushing systems toward the lowest cost, lowest integrity publications.

At a time of rising integrity challenges, including organized fraud and AI enabled manipulation, the answer is more robust peer review and editorial oversight, not less.

Recommendations for NIH

AAP supports continued robust investment in federal research as critical to the advancement of American science. The United States should cultivate the discoveries which have led our great nation to a position of leadership in health innovation and cures. Publishing is an important part of the scientific and medical ecosystem and is necessary to ensure we can overcome the challenges of the 21st century.

AAP offers the following recommendations to NIH as the agency considers updating metrics:

1. **Prioritize integrity and quality in publication and reproducibility.**
    - Recognize peer reviewed publication and the Version of Record (VoR) as the primary, trusted record of research findings.
    - Invest and fund research into reproducibility, including metrics encouraging researchers to pursue reproducibility studies and share a broad range of research inputs as additive to a professional career path.
    - Develop metrics to support and encourage study pre-registration, study protocols, and registered reports to detail and establish research methods in advance of experimentation and data collection.
    - Encourage grant-funding agencies to prioritize outputs with strong integrity safeguards in grant reviews and policy design.
    - Support initiatives to expand and improve peer review, including metrics to reward peer review participation, training, and support for structured peer review.
2. **Support sustainable public access.**
    - Preference appropriately licensed and funded VoR in public repositories and clearly label material that has not been peer reviewed or formally published.
    - Ensure sustained and robust funding continues to be available for publishing costs as part of NIH grants. Avoid one size fits all mandates or metrics that undermine the financial sustainability of domestic non-profit scientific and medical societies and other high integrity publishers.
3. **Preserve author's choice and intellectual property.**
    - Allow researchers to choose where and how to publish their work, including the licenses applied, within a framework that advances public health.
    - Uphold robust intellectual property protections and contractual freedom as foundational to a vibrant, competitive publishing marketplace that can innovate in support of open science and AI era needs.
4. **Partner with publishers on open science and AI.**
    - Collaborate with publishers to design metrics to encourage public access, data sharing, and AI related policies that use existing infrastructure for provenance, corrections, retractions, and integrity checks.

- Encourage the use of licensed, verified VoR and curated datasets as preferred inputs for high-stakes AI systems, recognizing publishers as part of the trust and safety infrastructure for scientific information.

A financially sustainable and collaborative open science system that engages the publishing community can be a powerful engine for research and innovation. AAP members stand ready to work with NIH to ensure the agency's investment in science advances human health and welfare, supports high-quality jobs, and strengthens U.S. leadership in discovery and innovation.

With Regards,

J. Carl Maxwell
Senior Vice President, Public Policy
Association of American Publishers

227. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Rigor and reproducibility are fundamental aspect of research, and we do our best to follow these principles. However, investigators who focus on reproducibility and scientific rigor often face more challenges in securing fundings. I would recommend the following actions:

1- Add extra points to grant proposals that include independent replications of the results and are conducted by independent investigators
2- Reward team science by adding extra points to research team with multiple PIs and include at least one mPI focused on the computational work
3- Publish new RFA that are focused on secondary analysis of data

A growing and urgent challenge for U.S. investigators is the escalating cost of publishing research under mandatory open-access policies. For many federally funded scientists, including myself, the publication fees required by top-tier journals have become prohibitive. For example, Nature now charges approximately $13,000 for an open-access article. These costs are not covered by most grants, and they far exceed what individual investigators or small laboratories can absorb. As a result, many researchers are effectively excluded from publishing in journals that provide the most rigorous peer review and highest visibility. This undermines the very goals of rigor, reproducibility, and scientific impact that NIH seeks to promote. If investigators cannot afford to publish in journals with strong editorial standards, then efforts to strengthen scientific quality become aspirational rather than achievable. Sustainable solutions to open-access publishing costs are essential to ensure that high-quality science remains accessible, equitable, and competitive.

2. Data, Software, and Model Sharing: Data sharing is mandated by NIH but many investigators still manage to avoid sharing of data. I would like to suggest these actions:

1- Reward data sharing by providing extra funds to investigators who shared data and would like to maintain the data by sharing updates
2- Add extra points to grant proposals that will augment data repository
3- Publish new RFA that support generation and sharing of software

3. Training and Mentorship: Training and mentorship are important, and mentors usually do this for free. It is important to support time dedicated to mentoring with salary support

4. Collaboration: It is important to appreciate grant proposals with multiple PIs that cover every aspect of a complex project. Statisticians and bioinformaticians role should be elevated to an mPI role

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: This is all well said. The reality is that reviewers in study sections kill grant proposals that do not include preliminary data, for example R21 proposals. Reviewers scores should be redacted when they make these mistakes

7. Public Impact: Grant proposals should include a clear dissemination plan that goes beyond publications in scientific journals. It would be important to reward investigators who include panels of stakeholders and have a plan to engage the community using modern ways of communications, e.g. blogs, social media, press release, interviews with reputable journalists

8. Other Areas: N/A

228. Biophysical Society

I am responding: On behalf of an organization

Name: Leann Fox

Name of Organization: Biophysical Society

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Director, Advocacy and Public Affairs

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: The scientific research enterprise has long been built on a foundation of rigor and reproducibility, treating them as essential safeguards for ensuring public trust. Across disciplines, researchers have relied on systematic methods, peer review, transparent documentation, and replication studies to ensure that findings can be independently verified. This commitment has shaped everything from laboratory protocols to data sharing norms - reinforcing the idea that scientific progress depends not just on discovery, but on the ability of others to reliably reproduce and build upon that work.

2. Data, Software, and Model Sharing: Collaboration has been a defining strength of modern science since the post WWII era, when domestic cooperation and international partnerships became central to advancing knowledge and addressing global challenges. This collaborative principle underpins modern practices of data, software, and resource sharing, which accelerate discovery by enabling validation of results, expanding access to high value datasets, and supporting reuse that fuels new lines of inquiry. By incentivizing these shared resources, NIH reinforces a long standing pillar of science: that collective effort—across institutions, disciplines, and borders—drives innovation and maximizes the impact and society benefit of research investments.

3. Training and Mentorship: Investment in the foundational and biomedical research workforce is essential to sustaining U.S. scientific and technological leadership, and it begins with building a strong domestic pipeline that can grow and thrive over time. Like any other endeavor, U.S. prominence at scientific frontiers depends on predictable, long term funding and on creating meaningful opportunities for early career researchers to develop their skills, establish independence, and contribute new ideas. Training and mentorship are central to this effort: they strengthen the research workforce, support stability, and ensure that emerging scientists have the guidance and protected time needed to advance their careers. By deepening its commitment to these investments and improving how their impacts are measured, NIH can help secure a strong and innovative research workforce for the future.

4. Collaboration: Collaboration has been a defining feature of the scientific enterprise since WWII, when the United States and its allies built a model of cooperative research that brought together leading experts to confront global challenges. Today's increasingly interdisciplinary and convergent biomedical research continues to rely on that foundation, drawing strength from partnerships that span fields,

institutions, and national borders. Such collaboration enables teams to integrate diverse expertise, accelerate discovery, and produce cutting edge science that no single discipline or country could achieve alone. As NIH explores new ways to assess and recognize contributions to team based research, maintaining a strong commitment to collaboration in all its forms remains essential to advancing scientific innovation.

5. Entrepreneurship and Translation: Basic, foundational research remains the engine of biomedical innovation, providing discoveries and insights that make future medical advances possible. History repeatedly shows that transformative breakthroughs—from new therapies to cutting edge technologies—emerge from sustained investment in early stage fundamental scientific enquiry. As NIH works to continually accelerate translation from bench to bedside and strengthen innovation across sectors, it must also preserve robust opportunities for foundational research, ensuring that efforts to spur entrepreneurship and product development do not come at its expense. Protecting this balance is essential to maintaining a research ecosystem where curiosity driven discovery and translational progress reinforce one another and continue to deliver medical advances for the public.

6. Foundational Scientific Exploration: Foundational and high risk, high reward research are both essential to a thriving scientific ecosystem, and each depends on a commitment to curiosity driven exploration. Foundational research expands the boundaries of knowledge, while bold, high risk efforts push science toward breakthroughs that can redefine entire fields. Repeatedly, history shows that new basic knowledge can give rise to cutting edge and curative advances in unexpected ways, making it vital NIH continue to support these approaches in equal measure. Embedding both forms of inquiry within the academic enterprise ensures that scientists have the freedom to explore, experiment, and imagine—conditions that make transformative discovery possible.

7. Public Impact: NIH's focus on assessing and recognizing the broader public impact of biomedical research highlights how deeply science contributes to positive health outcomes, domestic economic growth, and advances in clinical practices and their accessibility. These benefits depend on a strong and sustained research enterprise—one that continues to earn public trust by delivering tangible improvements in people's lives. To maintain this momentum, increased and predictable funding is essential, ensuring that researchers have the stability and resources needed to drive innovation and translate discoveries into broad public benefit.

8. Other Areas: N/A

229. AAMC

I am responding: On behalf of an organization

Name: Anurupa Dev

Name of Organization: AAMC

Type of Organization: Non-Profit Research Organization

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/AAMC-Comments-re-NIH-NOT-OD-26-087.pdf

Description: This comment was submitted via email

AAMC
655 K Street, NW, Suite 100
Washington, DC 20001-2399
T 202-828-0400
aamc.org

August 19, 2026

Office of Extramural Research
National Institutes of Health
Bethesda, MD 20892

Re: Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Submitted via https://osp.od.nih.gov/comment-form-measuring-and-rewarding-scientific-impact/

The AAMC[1] appreciates the opportunity to comment to NIH on indicators and incentives for measures of scientific success that align with the agency's core mission. We strongly agree that scientific impact can be reflected in a myriad of outputs beyond publications and citation metrics and believe that the use of additional indicators can enhance our understanding of both research productivity and the effectiveness of grant funding.

To frame these comments, first we would like to stress the importance of building on and amplifying existing community efforts across the identified areas, including the development of infrastructure, standards, metrics for evaluation and assessment. NIH has a key role in uplifting these initiatives as well as lending credibility and value to these additional measures of success by integrating them throughout the grants process, including at the application stage, within progress reports, and as an internal indicator to understand the full return of NIH investment. The agency should consider how its policies, processes, and funding priorities may inhibit or promote these activities, particularly in areas such as collaboration, foundational research, and rigor/reproducibility. We also note that the effectiveness of specific indicators will vary widely, depending on the scientific discipline, career stage, and resources available to an individual investigator. The adoption of additional measures of

[1] Established in 1876, the AAMC is a nonprofit association dedicated to improving the health of people everywhere through medical education, clinical care, biomedical research, and community collaborations. Its members are all 163 U.S. medical schools accredited by the Liaison Committee on Medical Education; 12 Canadian medical schools accredited by the Committee on Accreditation of Canadian Medical Schools; more than 400 academic health systems and teaching hospitals; and more than 70 academic societies. The AAMC also collaborates with the Department of Veterans Affairs (VA) and VA medical centers through their academic affiliations with AAMC-member institutions. Through these institutions and organizations, the AAMC leads and serves America's medical schools, academic health systems and teaching hospitals, and the millions of individuals across academic medicine, including more than 220,000 full-time faculty members, 100,000 medical students, 167,000 resident physicians, and 60,000 graduate students and postdoctoral researchers in the biomedical sciences. Through the Alliance of Academic Health Centers International, AAMC membership reaches more than 60 international academic health centers throughout five regional offices across the globe.

success by NIH could facilitate a culture change in how the outcomes of biomedical research are valued and incentivized across the enterprise.

The AAMC conducted a comprehensive study in 2016[2] in collaboration with RAND Europe, on metrics which can be used to assess and communicate the value of biomedical research. We would like to recommend these findings to NIH as a key resource that responds to many of the areas which NIH cites in this RFI and identifies many potential metrics for the agency to consider, particularly in the areas of training and mentorship, entrepreneurship and translation, and public impact. The report details how each of these metrics might be obtained, as well as caveats or limitations on their use. The report was released alongside additional resources,[3] including a brief on developing a research evaluation framework and proceedings from a workshop discussing how the framework could be applied within a variety of institutional structures, missions, and disciplines across the spectrum of biomedical research. We would be pleased to engage with NIH in a discussion of how these metrics could be incorporated into this initiative.

We can provide particular insight into metrics and incentives for scientific data, model, and software sharing based on years of work in collaboration with NIH and many other stakeholders across the research enterprise. The AAMC strongly supports policies which facilitate meaningful sharing and re-use of the data resulting from NIH-funded research projects, and the development and implementation of mechanisms which would track and incentivize these research activities. To further this goal, the AAMC, along with the New England Journal of Medicine and the Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard, carried out an effort titled "Credit for Data Sharing,"[4] which describes a system to appropriately credit researchers for sharing data[5]. This multi-sector initiative convened academic institutions, nonprofits, publishers, and funders to discuss the actions needed by these partners to enable such a system.

We understand that this is long-term work which involves both technical infrastructure and standards, as well as changes to incentives and the current culture of which research outputs are valued. In order to incentivize data and software sharing, we must first provide guidance and support for its curation, storage, dissemination, and tracking. We are pleased to engage with NIH's initiative to do just that, through the S-index Challenge,[6] and engage in the follow-up efforts of socializing the resulting model(s) and metrics developed within the academic research community. The more these metrics are recognized and accepted by institutions and investigators, the more straightforward it will be to incorporate them into current measures of research success.

---

[2] 100 Metrics to Assess and Communicate the Value of Biomedical Research (2016). https://www.rand.org/pubs/research_reports/RR1606.html

[3] Assessing & Communicating the Value of Biomedical Research. https://www.aamc.org/about-us/mission-areas/biomedical-research/research-evaluation

[4] Credit for Data Sharing. https://www.aamc.org/about-us/mission-areas/biomedical-research/data-sharing

[5] Pierce, H.H., et al. Credit data generators for data reuse. *Nature* 570, 30-32 (2019) https://doi.org/10.1038/d41586-019-01715-4

[6] NIH Data Sharing Index (S-index) Challenge. https://www.nih.gov/challenges/nih-data-sharing-index-s-index-challenge

The AAMC appreciates NIH's engagement with the research community on identifying indicators for measuring and rewarding scientific impact, and we look forward to working with you as you move forward on these efforts. We would be pleased to provide additional information or clarification on any of the recommendations in these comments. For questions, please contact Heather Pierce, JD, MPH, Senior Director for Science Policy and Regulatory Counsel (hpierce@aamc.org) or Anurupa Dev, PhD, Director, Science Policy and Strategy (adev@aamc.org).

Sincerely,



Elena Fuentes-Afflick, MD, MPH
Chief Scientific Officer

cc: David J. Skorton, MD, AAMC President and Chief Executive Officer

230. United States Association for the Study of Pain

I am responding: On behalf of an organization

Name: Brittany Knight, PhD

Name of Organization: United States Association for the Study of Pain

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Executive Director & CEO

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: USASP supports NIH's efforts to strengthen rigor, transparency, reproducibility, and replicability as related but distinct dimensions of trustworthy research. Findings may be reproduced across studies that share the same biases, for example, while differences among studies may reveal meaningful variation across populations, methods, models, or settings. USASP recommends that NIH evaluate reproducibility within a broader framework that rewards rigorous design, transparent reporting, appropriate interpretation of uncertainty, independent validation, and communication of null and negative findings.

Prioritizing Replication and Validation. Replication investments should focus on findings whose failure would carry substantial scientific, clinical, or public consequences. Relevant considerations include uncertainty or conflict in the existing evidence; reliance of subsequent research or clinical decisions on the finding; burden of illness or disability; and the importance of the question and outcomes to people affected by the condition. Replication need not mean repeating a study identically. Conceptual replication, multisite validation, reanalysis, and evaluation across populations, methods, models, or real-world settings may provide greater scientific value and clarify generalizability—an especially important consideration in heterogeneous fields such as pain research.

Strengthening Review and Reporting. USASP recommends that NIH ensure reviewers have access to the methodological expertise and training needed to determine whether a study can credibly answer its research question. Review should address field-relevant considerations such as sample-size justification, safeguards against bias, measurement validity, predefined data-handling decisions, relevant biological variables, authentication of key resources, and alignment between the research question and analytic plan. For causal claims, reviewers should assess whether the design, assumptions, comparison groups, and analyses support causal inference and whether the conclusions appropriately distinguish causation from association. Existing initiatives, including the Landis reporting principles and IMI-Europain recommendations, demonstrate how consequential design decisions can be made visible while allowing standards to vary across research contexts (Landis et al., 2012; Knopp et al., 2015).

Potential Indicators. Indicators of rigor and reproducibility should make consequential research decisions and evidence of independent validation visible while preserving expert assessment of quality. Relevant

evidence could include transparent reporting of design and analytic decisions; protocol availability when appropriate; documentation of deviations and limitations; responsible availability of methods, data, software, and code; independent validation; and dissemination of rigorous null, negative, and replication findings. NIH could also identify subsequent studies that independently confirm, extend, refine, or challenge NIH-supported findings and allow investigators to document verified examples of downstream validation. Citation context may help identify these relationships, but a supportive citation should not be considered replication unless the citing study tests the relevant finding. NIH should address publication bias by supporting prospective registration and registered reports where appropriate, encouraging publication or repository pathways for rigorous null and negative findings, and recognizing complete reporting as a scientific contribution. No indicator should be treated as proof of research quality. Indicators must be interpreted in context to determine whether the methods were appropriate and the conclusions supported.

Measuring Scientific Contribution. Journal prestige, publication volume, citation counts, and researcher-level indices such as the h-index are poor proxies for methodological rigor or scientific contribution. Macleod et al. found no relationship between journal impact factor and reported safeguards against bias in preclinical research (2015). NIH should instead consider what research changes or enables, including whether it clarifies uncertainty, redirects subsequent work, improves methods or shared resources, enables discovery, influences health or practice, or prevents unproductive investment. These contributions may take years to emerge and should not be collapsed into another universal numerical "impact factor."

Consideration of impact beyond medical researchers is important. This includes ensuring that the presentation of data always strives to provide easily understood graphs (and tables) of results and careful explanation of the implications in language understandable to non-scientists and other affected communities. The inclusion of community members in the conduct, presentation, and publication of research can help to ensure that presentation of results will be more impactful on the people impacted by the findings. Quantifying impact into additional areas such as clinical, community, economic, and policy domains as described as part of the Translational Science Benefits Model (TSBM), helps to expand the influence of funded research. Researchers should be trained to be part of research training programs.

Building a Culture of Trustworthy Research. Sustainable improvement requires aligned action by NIH, research institutions and centers, journals, investigators, and professional societies. NIH should pair its expectations with support for replication, methodological expertise, research personnel, data stewardship, protected time, training, and shared infrastructure. The pain-specific ENTRUST-PE framework provides a useful lifecycle model by considering responsible governance, rigorous and transparent methods, applicability to intended populations, relevance of questions and outcomes, balanced communication, and well-documented data (O'Connell et al., 2025).

Positive incentives and field-led initiatives can translate rigor principles into practical changes in research culture. Professional societies can contribute by adapting these principles to their disciplines. For example, USASP has used NINDS-developed rigor icons at its Annual Scientific Meetings to help presenters communicate key design features. NIH could also build upon the NINDS Rigor Champions

Prize as a positive-incentive model (NINDS, 2026). Challenge prizes can recognize investigators, trainees, research staff, and teams while identifying promising practices implemented beyond routine requirements. Successful approaches should be eligible for follow-on support to evaluate, refine, disseminate, and sustain them.

NIH should partner with societies, community-based organizations, and other stakeholders to determine whether tools and funded interventions meaningfully improve research, but also of equal importance is consideration of the practices and quality of clinical care. Such approaches should be thoughtful enough to support planning, reflection, and cultural change rather than becoming certification systems or additional checklists. USASP strongly recommends that NIH collect data and feedback on the effectiveness, burden, and feasibility of rigor interventions and disseminate the findings not only to the research community but the general community on whom it will have an impact. This evidence should guide decisions about broader implementation or necessary refinements to improve feasibility and effectiveness.

2. Data, Software, and Model Sharing: USASP supports NIH's efforts to incentivize the sharing of scientific data, software, models, and other research resources. Successful sharing should be defined by whether prospective users can understand how a resource was generated, evaluate its quality and limitations, and reuse or combine it appropriately.

Supporting Curation and Infrastructure. Effective data sharing depends on expert curation and sustained infrastructure. Even within coordinated research networks, investigators may struggle to standardize metadata, reconcile variable definitions, document provenance, complete quality-control requirements, and satisfy repository expectations. These challenges are compounded when submissions are infrequent and staff turnover limits retention of specialized knowledge. We recommend that NIH provide sustained support for repository curators and data coordinators who can guide investigators through preparation and submission, preserve institutional knowledge, and improve the quality and usability of shared resources. For large consortia, each data-generating laboratory could designate a trained data steward and develop a data-management SOP in coordination with the repository or data coordinating center. Dedicated budget categories, administrative supplements, shared curation cores, and continued support beyond the original award period would prevent these activities from competing with core research costs. Together, these investments would reduce submission inefficiencies and ensure that shared resources remain well documented, interoperable, and reusable over time.

Promoting Quality and Harmonization. Harmonization should begin during study design rather than being imposed retrospectively. Common data elements, structured metadata, controlled vocabularies, data dictionaries, and transparent protocols can support cross-study comparison, pooled analyses, and machine-assisted reuse. The HEAL Data Ecosystem illustrates the value of coordinated standards when they are scientifically appropriate. NIH could encourage conformance with established standards while allowing justified variation across study designs, populations, methods, and data types. Standardization cannot repair biased sampling, inadequate power, invalid measures, or weak analysis; NIH should therefore assess the quality, provenance, documentation, and limitations of shared resources, not merely their technical compliance. Uses involving artificial intelligence also require clear provenance, licensing, versioning, validation, documentation of limitations, and appropriate protection of participant information.

Promoting Responsible Reuse. Data availability does not ensure that investigators understand how shared resources may be used. Misconceptions persist about whether public data can support grant applications, hypothesis development, pooled analyses, experimental troubleshooting, or subsequent publications. NIH should coordinate with journals, repositories, and institutional libraries to provide clear guidance and examples demonstrating that responsible reuse is legitimate and encouraged when consistent with consent, access, attribution, and data-use requirements. Addressing these misconceptions would increase the scientific return on existing data investments.

Measuring and Rewarding Reuse. NIH should evaluate shared resources based on demonstrated reuse and downstream scientific value, not merely whether they were deposited or made available. Evidence could include formal citation, verified incorporation into new studies, independent validation, external adoption, development of derivative tools, educational or clinical applications, and research that would otherwise have been difficult or costly. Evaluation should reflect the type of resource involved. Software may require documentation, versioning, testing, and maintenance, while computational and experimental models require appropriate validation, characterization, provenance, and disclosure of intended uses and limitations. To make these contributions traceable, USASP recommends that NIH work with journals and repositories to promote persistent identifiers, formal resource citation, and bidirectional links among grants, publications, datasets, software, models, and contributors.

Resources demonstrating substantial downstream value should be eligible for competitive supplements supporting continued curation, harmonization, expansion, maintenance, validation, and user assistance. Recognition and support should extend to curators, data managers, research software engineers, tool developers, and coordinating-center personnel. This combination of visible attribution and reinvestment would reward meaningful reuse while preserving high-value resources for future discovery.

3. Training and Mentorship: Training and mentorship should be evaluated by the capabilities, independence, and sustainable careers they foster, rather than simply by participation or conventional academic outputs. Publications, grants, appointments, and trainee counts provide useful context but do not fully capture growth in rigorous research practices, methodological expertise, communication, leadership, interdisciplinary collaboration, or team-science skills. USASP therefore, recommends that NIH consider a broader portfolio of outcomes, including access to professional networks and sponsorship, successful career transitions, sustained participation in research, and subsequent contributions to science, health and the community at large.

Training outcomes should be assessed over time and interpreted in relation to the program, career stage, and trainee's goals. Including training on methods for increasing the impact of science research and communication beyond the research community will substantially improve our ability to assess the importance of the work promoted by our training program. (See Haroutounian, S. et al. IMMPACT recommended considerations (2023). For interdisciplinary pain researchers, developing foundational knowledge across fields, establishing durable collaborations, or remaining engaged in pain research may represent important early outcomes before publications or independent funding emerge. Therefore, we recommend that NIH use existing data systems and follow-up surveys wherever possible to evaluate longer-term outcomes without creating duplicative reporting requirements.

Recognizing Mentorship Quality and Effort. Mentorship quality should be evaluated independently from a trainee's eventual publication or funding record, which is influenced by institutional resources, funding conditions, career goals, and other factors beyond the mentoring relationship. NIH could use validated, confidential mentee-reported assessments addressing communication, accessibility, constructive feedback, career guidance, support for increasing independence, and creation of a respectful research environment. These measures should support program improvement rather than rank individual mentors, and reporting must protect trainees when programs or mentoring relationships whenever possible.

NIH should also recognize mentorship occurring outside formal training and career-development awards, including team mentorship, peer mentorship, sponsorship, and cross-institutional mentorship. Investigators could document significant mentoring contributions through a concise narrative or structured contribution statement. Because high-quality mentorship requires time and sustained engagement, NIH should permit reasonable mentor effort on training awards and consider broader mechanisms providing protected time or salary support for substantial mentoring responsibilities. Recognition programs such as the NINDS Landis Award can highlight exemplary mentorship, but selective awards should complement—not substitute for—routine support of mentoring effort.

Expanding Access to Mentorship. NIH should assess the complete mentoring environment rather than relying solely on an individual investigator's prior mentoring record. Relevant evidence may include mentoring teams, institutional resources, plans for mentor development, and access to complementary scientific and career expertise. This approach would allow early-career investigators to develop as mentors while ensuring appropriate support for trainees, and it would recognize high-quality mentorship at institutions that have fewer formal NIH training programs.

Professional societies can complement institutional mentoring by providing cross-institutional networks, interdisciplinary engagement, and leadership development. For example, the USASP Leadership Academy pairs established pain researchers and clinicians with midcareer investigators to develop leadership and professional skills that may not be available within their home institutions. NIH could partner with professional societies to evaluate and expand approaches of this kind while preserving institutional responsibility for mentorship and career development.

Supporting Sustainable Career Pathways. Training programs should recognize successful careers in clinical practice, industry, government, nonprofit organizations, research administration, communication, and other sectors—not only tenure-track appointments or NIH-funded faculty positions. NIH could consider whether its indicators inadvertently encourage mentors to select trainees based on their predicted likelihood of achieving a narrowly defined academic outcome. NIH could also consider regularly assessing whether NRSA stipends and related support remain adequate relative to inflation, regional living costs, and institutional compensation requirements. Adequate and predictable support is necessary to retain talented researchers and realize the long-term value of NIH's training investments. Recognizing and sustaining varied career pathways would extend the benefits of NIH training across the broader research enterprise and strengthen the workforce needed to advance science and health.

4. Collaboration: USASP recommends that NIH assess collaboration by the scientific integration and added value it produces, rather than by the size or visibility of the team. Meaningful team science

combines complementary expertise, methods, perspectives, and resources to address questions that investigators could not answer as effectively on their own. This approach would reward substantive collaboration rather than the appearance of interdisciplinarity.

Funding Integrated Team Science. We recommend that NIH expand upon its opportunities for investigator-initiated multidisciplinary research through U19 and comparable multi-project or cooperative mechanisms when scientific objectives require interdependent projects, shared cores, coordinated methods, and sustained collaboration. Planning awards, collaborative pilot funding, administrative supplements, and coordinating centers could support smaller or emerging teams. Allowable costs should recognize the infrastructure required for collaboration, including project management, team-science facilitation, shared communication and data platforms, methodological consultation, standards development, evaluation, and appropriate compensation for external partners.

Policies designed around individual investigators should not inadvertently discourage genuine multiple-PI projects. If NIH limits the number or value of awards held by an investigator, it should consider the investigator's actual effort, authority, and share of project leadership rather than treating participation in an integrated MPI team as equivalent to directing the project alone. Any exception should require evidence of substantive leadership to discourage nominal MPI designations.

Integrating Training with Team Science. Large, milestone-driven programs are not automatically effective training environments. Trainees may contribute substantially to predetermined deliverables without gaining experience identifying research gaps, developing hypotheses, or designing studies. NIH should encourage team-science awards to include trainee pilot funds, exploratory subprojects, mentored proposal development, and opportunities to lead cross-disciplinary working groups or propose questions arising from consortium resources.

The HEAL PAIN Cohort T90/R90 and NIDA T32 programs provides a useful model by combining interdisciplinary mentoring, authentic collaborative projects, career development, and scientific training across the spectrum of disciplines relevant to pain research. Comparable features could be incorporated into research consortia and multi-project awards, including programs not primarily designed as training mechanisms. This would allow trainees to develop both team-science skills and intellectual independence.

Assessing Collaboration Quality. Relevant indicators could include whether team members jointly shaped the research question and design; integrated distinct methods or disciplinary perspectives; established clear governance, communication, decision-making, authorship, and conflict-resolution practices; and produced knowledge, resources, or capabilities that would have been difficult to achieve independently. NIH could also consider the durability of relationships, development of shared methods or datasets, creation of new interdisciplinary directions, and dissemination of expertise beyond the original team. Contributions to community standards, metadata, protocols, and interoperability should be recognized when they enable collaboration across a field.

Recognizing Individual Contributions. Principal-investigator status, authorship order, and consortium authorship often obscure essential contributions from data curators, software developers, statisticians,

project managers, research coordinators, standards developers, and domain experts. Structured contributor-role frameworks offer a useful foundation but may need greater specificity for metadata design, standards development, maintenance of shared tools, use-case development, and other enabling work. NIH should encourage structured, machine-readable contribution records across publications, datasets, software, models, and standards, supplemented, when necessary, by brief descriptions of the significance of each contribution.

Within consortia, funding and governance plans should prospectively recognize laboratories and investigators contributing data, validation, use cases, domain expertise, or standards development even when they do not build or publish the final resource. Appropriate subaward support, protected effort, and formal credit would reduce friction and ensure that essential collaborative work is not treated as invisible service.

5. Entrepreneurship and Translation: Defining NIH's Role. NIH can most effectively advance translation by reducing the scientific, technical, and coordination risks that prevent credible discoveries from reaching capable developers and users. Its role should be to support validation, development, training, and connections to appropriate expertise—not to presume that every discovery should become a commercial product or that NIH should operate as an entrepreneur. Support should reflect the type and maturity of the innovation and connect investigators with relevant regulatory, manufacturing, clinical, implementation, and commercialization expertise.

Matching Support to Development Requirements. Therapeutic development often requires more time and capital than conventional small-business or milestone-driven awards provide. Validation, toxicology, manufacturing, regulatory engagement, and clinical studies may be required before private investment, or an industry partnership becomes feasible. NIH should evaluate whether SBIR/STTR, UG3/UH3, and related mechanisms provide realistic budgets, timelines, and continuity for the development stages they are intended to support. Where persistent gaps exist, NIH could provide larger or longer awards, bridge support between development stages, or complementary mechanisms extending beyond conventional limits. Continued funding should remain tied to clearly defined scientific, technical, regulatory, and development milestones, with independent decisions to continue, revise, or discontinue support.

NIH should also expand access to training and technical assistance in regulatory strategy, manufacturing, clinical development, reimbursement, intellectual property, and commercialization planning. These resources may be particularly valuable for academic investigators and first-time entrepreneurs who have strong scientific expertise but limited product-development experience.

Designing for Adoption. Technical quality does not ensure that a product, tool, standard, or infrastructure resource will be used. NIH should encourage translation-focused projects to identify intended users and potential adopting organizations early enough for them to help define requirements, test usability, establish implementation milestones, and plan for the distribution of knowledge into the community of potential users. Including affected communities and users early would help to ensure that the presentation of data is made understandable to non-researchers. Joint funding or shared milestones between developers and adopting partners could create accountability for both development and implementation. Expectations should vary by project maturity so that early engagement demonstrates a credible need without requiring a fully committed commercial partner before an innovation is

sufficiently developed.

Improving University-to-Startup Translation. The time, cost, and uncertainty involved in licensing university-owned technologies can prevent promising discoveries from reaching qualified startups. NIH-funded institutions should be encouraged to use transparent licensing processes, reasonable timelines, standardized startup agreements, proportionate financial terms, and procedures for resolving stalled negotiations. More predictable terms would allow early-stage companies to invest limited resources in validation and development while preserving appropriate institutional and public interests in federally supported discoveries.

Measuring Translational Contribution. Translational impact should reflect whether an innovation responsibly advanced toward practical application, not simply the number of patents, licenses, startup companies, or dollars raised. Relevant evidence may include completion of scientific, regulatory, manufacturing, or clinical milestones; engagement of qualified development or adoption partners; integration into independent research or clinical workflows; and appropriate follow-on support. Unprompted adoption by organizations outside the originally funded program will provide especially meaningful evidence that a product, tool, or standard addressed a broader need.

Indicators should account for differences among therapeutics, devices, diagnostics, software, research tools, communication into the general community, and implementation strategies. NIH should also recognize well-conducted work demonstrating that a product is ineffective, unsafe, infeasible, or not ready for further development. Timely discontinuation based on credible evidence can prevent greater downstream losses and should not automatically be treated as failure.

6. Foundational Scientific Exploration: Basic research and high-risk, high-reward research are both essential to scientific progress but should not be treated as interchangeable. Basic research builds the knowledge, methods, and conceptual foundations upon which future advances depend; much of it is incremental and cumulative rather than inherently high-risk. High-risk research deliberately tests unconventional ideas whose probability of success is uncertain, but whose potential scientific effect is substantial. These forms of exploration require distinct funding and review strategies.

Sustaining Foundational Basic Research. Foundational research should not be required to demonstrate an immediate path to a product, intervention, or clinical outcome. Its value may become apparent only after multiple discoveries converge, sometimes many years later. Research on incretin physiology and glucagon-like peptides ultimately enabled GLP-1–based therapies, while studies of bacterial immune systems provided the foundation for CRISPR-Cas9 gene editing (NIDDK, 2021; NIGMS, 2024). Neither trajectory could have been fully predicted when the underlying observations were first investigated.

NIH should maintain stable support for investigator-initiated basic research alongside goal-directed, translational, and clinical programs. Tightly defined initiatives can mobilize research around urgent needs, but excessive emphasis on predetermined outcomes may narrow the questions being investigated and reduce opportunities for unexpected discovery. NIH should transparently monitor its investments across these categories and ensure that every institute and center sustains the foundational science needed for future progress.

Foundational science should also include careful incremental research that validates methods, resolves inconsistencies, characterizes mechanisms, refines models, or establishes the boundaries of a finding. NIH should not equate innovation solely with novelty or dramatic claims. Relevant indicators of foundational contribution may include creation of new concepts, methods, measures, models, datasets, or research directions; resolution of scientific uncertainty; or eventual uptake across disciplines. Because these effects may emerge unpredictably, investigators should not be required to forecast a specific clinical or commercial impact.

Supporting High-Risk, High-Reward Research. High-risk research requires review processes that tolerate genuine uncertainty without relaxing methodological standards. Although NIH's R21 and High-Risk, High-Reward programs state that preliminary data are not required (NIH, 2026a; NIH, 2026b), applicants may still perceive that reviewers expect conventional evidence of feasibility. NIH should examine reviewer critiques, scoring patterns, and funding decisions to determine whether preliminary-data expectations are being reintroduced in practice or whether projects described as innovative remain comparatively conventional.

Reviewer training should distinguish the uncertainty inherent in testing an unconventional idea from avoidable weaknesses in study design. Review criteria could separately consider the importance of the question, degree of conceptual departure, potential magnitude of the advance, credibility of the scientific rationale, rigor of the approach, and value of the knowledge generated even if the central hypothesis is not confirmed. NIH should recruit reviewers experienced in evaluating unconventional or cross-disciplinary research while ensuring that high-risk mechanisms remain accessible to early-career investigators and researchers entering new fields.

Funding for Learning. High-risk programs should be structured around learning and decision-making rather than the expectation that every aim will succeed. Staged awards could use learning milestones, flexible follow-on funding, and opportunities to modify direction when early findings challenge the original premise. NIH should evaluate these mechanisms as portfolios by asking whether they support ideas unlikely to succeed in conventional review, generate new concepts or methods, and occasionally produce unusually substantial advances. Projects that disprove plausible hypotheses, identify unworkable approaches, or establish unexpected limitations can still provide a meaningful return by preventing unproductive investment and informing future research.

7. Public Impact: USASP supports NIH's efforts to recognize the many ways biomedical research benefits the public. Public impact is multidimensional, often collective, and may emerge long after an award ends. Relevant evidence could include improvements in health outcomes or clinical practice; incorporation into guidelines, policy, or health-system operations; development of accessible tools and educational resources; meaningful use by people affected by health conditions, clinicians, educators, or community organizations; economic or workforce benefits; and contributions to public understanding of science. Public visibility, social-media reach, or favorable attention alone should not be treated as evidence of meaningful impact.

Building Capacity for Dissemination. Public impact rarely follows from publication alone. Researchers may lack the time, expertise, partnerships, or resources needed to communicate findings to those who could use them. We appreciate efforts like NIH ENGAGE that focus on integrating people with lived

experience into research from early stages through to translation. But with an evolution in practice, capacity building is essential. NIH should provide training and dedicated support for plain-language writing, visual communication, responsible use of social media, and dissemination to people living with health conditions, clinicians, community health workers, educators, policymakers, and other intended users. These skills should be incorporated into training and career-development programs and project dissemination plans. Grant funds should also be available to support communication specialists and compensate external partners who help adapt and disseminate findings.

Creating Accessible Scientific Resources. Well-curated repositories can extend public impact by offering accessible pathways to NIH-supported research. Where privacy, consent, and other protections permit, repositories could provide plain-language study summaries, public-facing explanations connecting data to research questions, educational materials grounded in active research, appropriately deidentified classroom datasets, and channels through which patient and community organizations can ask questions or identify areas of interest. Evidence of educational use, adoption by public-serving organizations, or incorporation into accessible tools may help demonstrate impact more meaningfully than website traffic or download counts alone.

Recognizing and Sharing Effective Practices. NIH could establish an annual recognition program honoring investigators, research teams, research staff, communicators, and community partners whose work exemplifies meaningful public impact. Awards could be named for distinguished former NIH leaders whose careers reflected public service and the application of scientific knowledge for societal benefit. Recognition need not include monetary support; an opportunity to present the work publicly and dissemination of the recipient's approach as a case study could provide meaningful visibility.

Selection should focus on documented contribution rather than institutional reputation, publication metrics, or public prominence. Publishing recipients' strategies, challenges, and lessons learned would make the program more than an award by helping NIH identify and disseminate effective approaches throughout the biomedical research community.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/@USASP-RFI-comment_081926.pdf

August 19, 2026

National Institutes of Health
Office of Science Policy
Submitted electronically via the NIH RFI response portal

Re: NOT-OD-26-087 — Request for Information on Measuring and Rewarding Scientific Impact

Dear NIH Office of Science Policy Colleagues:

On behalf of the United States Association for the Study of Pain (USASP), thank you for the opportunity to respond to NIH's Request for Information on Measuring and Rewarding Scientific Impact. USASP brings together researchers, clinicians, people with lived experience of pain, and other partners committed to advancing science that reduces the burden of pain.

Because pain research spans basic, preclinical, clinical, behavioral, rehabilitative, translational, and implementation science, measures of impact must remain appropriate to the goals, methods, and settings involved. Across all topic areas, NIH should use context-sensitive portfolios of evidence rather than universal scores or rankings. Indicators should inform, not replace, expert judgment and should be piloted before broad implementation. NIH should evaluate their validity, administrative burden, privacy implications, susceptibility to gaming, and effects across disciplines, institutions, and career stages. New expectations should also be accompanied by the funding, personnel, training, and infrastructure necessary for meaningful implementation.

Rigor and Reproducibility

USASP supports NIH's efforts to strengthen rigor, transparency, reproducibility, and replicability as related but distinct dimensions of trustworthy research. Findings may be reproduced across studies that share the same biases, for example, while differences among studies may reveal meaningful variation across populations, methods, models, or settings. USASP recommends that NIH evaluate reproducibility within a broader framework that rewards rigorous design, transparent reporting, appropriate interpretation of uncertainty, independent validation, and communication of null and negative findings.

Prioritizing Replication and Validation. Replication investments should focus on findings whose failure would carry substantial scientific, clinical, or public consequences. Relevant considerations include uncertainty or conflict in the existing evidence; reliance of subsequent research or clinical decisions on the finding; burden of illness or disability; and the importance of the question and outcomes to people affected by the condition. Replication need not mean repeating a study identically. Conceptual replication, multisite validation, reanalysis, and evaluation across populations, methods, models, or real-world settings may provide greater

scientific value and clarify generalizability—an especially important consideration in heterogeneous fields such as pain research.

Strengthening Review and Reporting. USASP recommends that NIH ensure reviewers have access to the methodological expertise and training needed to determine whether a study can credibly answer its research question. Review should address field-relevant considerations such as sample-size justification, safeguards against bias, measurement validity, predefined data-handling decisions, relevant biological variables, authentication of key resources, and alignment between the research question and analytic plan. For causal claims, reviewers should assess whether the design, assumptions, comparison groups, and analyses support causal inference and whether the conclusions appropriately distinguish causation from association. Existing initiatives, including the Landis reporting principles and IMI-Europain recommendations, demonstrate how consequential design decisions can be made visible while allowing standards to vary across research contexts (Landis et al., 2012; Knopp et al., 2015).

Potential Indicators. Indicators of rigor and reproducibility should make consequential research decisions and evidence of independent validation visible while preserving expert assessment of quality. Relevant evidence could include transparent reporting of design and analytic decisions; protocol availability when appropriate; documentation of deviations and limitations; responsible availability of methods, data, software, and code; independent validation; and dissemination of rigorous null, negative, and replication findings. NIH could also identify subsequent studies that independently confirm, extend, refine, or challenge NIH-supported findings and allow investigators to document verified examples of downstream validation. Citation context may help identify these relationships, but a supportive citation should not be considered replication unless the citing study tests the relevant finding. NIH should address publication bias by supporting prospective registration and registered reports where appropriate, encouraging publication or repository pathways for rigorous null and negative findings, and recognizing complete reporting as a scientific contribution. No indicator should be treated as proof of research quality. Indicators must be interpreted in context to determine whether the methods were appropriate and the conclusions supported.

Measuring Scientific Contribution. Journal prestige, publication volume, citation counts, and researcher-level indices such as the h-index are poor proxies for methodological rigor or scientific contribution. Macleod et al. found no relationship between journal impact factor and reported safeguards against bias in preclinical research (2015). NIH should instead consider what research changes or enables, including whether it clarifies uncertainty, redirects subsequent work, improves methods or shared resources, enables discovery, influences health or practice, or prevents unproductive investment. These contributions may take years to emerge and should not be collapsed into another universal numerical “impact factor.”

Consideration of impact beyond medical researchers is important. This includes ensuring that the presentation of data always strives to provide easily understood graphs (and tables) of results and careful explanation of the implications in language understandable to non-scientists and other affected communities. The inclusion of community members in the conduct, presentation, and publication of research can help to ensure that presentation of results will be more impactful on the people impacted by the findings. Quantifying impact into additional

areas such as clinical, community, economic, and policy domains as described as part of the Translational Science Benefits Model (TSBM), helps to expand the influence of funded research. Researchers should be trained to be part of research training programs.

Building a Culture of Trustworthy Research. Sustainable improvement requires aligned action by NIH, research institutions and centers, journals, investigators, and professional societies. NIH should pair its expectations with support for replication, methodological expertise, research personnel, data stewardship, protected time, training, and shared infrastructure. The pain-specific ENTRUST-PE framework provides a useful lifecycle model by considering responsible governance, rigorous and transparent methods, applicability to intended populations, relevance of questions and outcomes, balanced communication, and well-documented data (O'Connell et al., 2025).

Positive incentives and field-led initiatives can translate rigor principles into practical changes in research culture. Professional societies can contribute by adapting these principles to their disciplines. For example, USASP has used NINDS-developed rigor icons at its Annual Scientific Meetings to help presenters communicate key design features. NIH could also build upon the NINDS Rigor Champions Prize as a positive-incentive model (NINDS, 2026). Challenge prizes can recognize investigators, trainees, research staff, and teams while identifying promising practices implemented beyond routine requirements. Successful approaches should be eligible for follow-on support to evaluate, refine, disseminate, and sustain them.

NIH should partner with societies, community-based organizations, and other stakeholders to determine whether tools and funded interventions meaningfully improve research, but also of equal importance is consideration of the practices and quality of clinical care. Such approaches should be thoughtful enough to support planning, reflection, and cultural change rather than becoming certification systems or additional checklists. USASP strongly recommends that NIH collect data and feedback on the effectiveness, burden, and feasibility of rigor interventions and disseminate the findings not only to the research community but the general community on whom it will have an impact. This evidence should guide decisions about broader implementation or necessary refinements to improve feasibility and effectiveness.

Data, Software, and Model Sharing

USASP supports NIH's efforts to incentivize the sharing of scientific data, software, models, and other research resources. Successful sharing should be defined by whether prospective users can understand how a resource was generated, evaluate its quality and limitations, and reuse or combine it appropriately.

Supporting Curation and Infrastructure. Effective data sharing depends on expert curation and sustained infrastructure. Even within coordinated research networks, investigators may struggle to standardize metadata, reconcile variable definitions, document provenance, complete quality-control requirements, and satisfy repository expectations. These challenges are compounded when submissions are infrequent and staff turnover limits retention of specialized knowledge. We recommend that NIH provide sustained support for repository curators and data coordinators who can guide investigators through preparation and

submission, preserve institutional knowledge, and improve the quality and usability of shared resources. For large consortia, each data-generating laboratory could designate a trained data steward and develop a data-management SOP in coordination with the repository or data coordinating center. Dedicated budget categories, administrative supplements, shared curation cores, and continued support beyond the original award period would prevent these activities from competing with core research costs. Together, these investments would reduce submission inefficiencies and ensure that shared resources remain well documented, interoperable, and reusable over time.

Promoting Quality and Harmonization. Harmonization should begin during study design rather than being imposed retrospectively. Common data elements, structured metadata, controlled vocabularies, data dictionaries, and transparent protocols can support cross-study comparison, pooled analyses, and machine-assisted reuse. The HEAL Data Ecosystem illustrates the value of coordinated standards when they are scientifically appropriate. NIH could encourage conformance with established standards while allowing justified variation across study designs, populations, methods, and data types. Standardization cannot repair biased sampling, inadequate power, invalid measures, or weak analysis; NIH should therefore assess the quality, provenance, documentation, and limitations of shared resources, not merely their technical compliance. Uses involving artificial intelligence also require clear provenance, licensing, versioning, validation, documentation of limitations, and appropriate protection of participant information.

Promoting Responsible Reuse. Data availability does not ensure that investigators understand how shared resources may be used. Misconceptions persist about whether public data can support grant applications, hypothesis development, pooled analyses, experimental troubleshooting, or subsequent publications. NIH should coordinate with journals, repositories, and institutional libraries to provide clear guidance and examples demonstrating that responsible reuse is legitimate and encouraged when consistent with consent, access, attribution, and data-use requirements. Addressing these misconceptions would increase the scientific return on existing data investments.

Measuring and Rewarding Reuse. NIH should evaluate shared resources based on demonstrated reuse and downstream scientific value, not merely whether they were deposited or made available. Evidence could include formal citation, verified incorporation into new studies, independent validation, external adoption, development of derivative tools, educational or clinical applications, and research that would otherwise have been difficult or costly. Evaluation should reflect the type of resource involved. Software may require documentation, versioning, testing, and maintenance, while computational and experimental models require appropriate validation, characterization, provenance, and disclosure of intended uses and limitations. To make these contributions traceable, USASP recommends that NIH work with journals and repositories to promote persistent identifiers, formal resource citation, and bidirectional links among grants, publications, datasets, software, models, and contributors.

Resources demonstrating substantial downstream value should be eligible for competitive supplements supporting continued curation, harmonization, expansion, maintenance,

validation, and user assistance. Recognition and support should extend to curators, data managers, research software engineers, tool developers, and coordinating-center personnel. This combination of visible attribution and reinvestment would reward meaningful reuse while preserving high-value resources for future discovery.

Training and Mentorship

Training and mentorship should be evaluated by the capabilities, independence, and sustainable careers they foster, rather than simply by participation or conventional academic outputs. Publications, grants, appointments, and trainee counts provide useful context but do not fully capture growth in rigorous research practices, methodological expertise, communication, leadership, interdisciplinary collaboration, or team-science skills. USASP therefore, recommends that NIH consider a broader portfolio of outcomes, including access to professional networks and sponsorship, successful career transitions, sustained participation in research, and subsequent contributions to science, health and the community at large.

Training outcomes should be assessed over time and interpreted in relation to the program, career stage, and trainee's goals. Including training on methods for increasing the impact of science research and communication beyond the research community will substantially improve our ability to assess the importance of the work promoted by our training program. (*See* Haroutounian, S. et al. IMMPACT recommended considerations (2023). For interdisciplinary pain researchers, developing foundational knowledge across fields, establishing durable collaborations, or remaining engaged in pain research may represent important early outcomes before publications or independent funding emerge. Therefore, we recommend that NIH use existing data systems and follow-up surveys wherever possible to evaluate longer-term outcomes without creating duplicative reporting requirements.

Recognizing Mentorship Quality and Effort. Mentorship quality should be evaluated independently from a trainee's eventual publication or funding record, which is influenced by institutional resources, funding conditions, career goals, and other factors beyond the mentoring relationship. NIH could use validated, confidential mentee-reported assessments addressing communication, accessibility, constructive feedback, career guidance, support for increasing independence, and creation of a respectful research environment. These measures should support program improvement rather than rank individual mentors, and reporting must protect trainees when programs or mentoring relationships whenever possible.

NIH should also recognize mentorship occurring outside formal training and career-development awards, including team mentorship, peer mentorship, sponsorship, and cross-institutional mentorship. Investigators could document significant mentoring contributions through a concise narrative or structured contribution statement. Because high-quality mentorship requires time and sustained engagement, NIH should permit reasonable mentor effort on training awards and consider broader mechanisms providing protected time or salary support for substantial mentoring responsibilities. Recognition programs such as the NINDS Landis Award can highlight exemplary mentorship, but selective awards should complement—not substitute for—routine support of mentoring effort.

Expanding Access to Mentorship. NIH should assess the complete mentoring environment rather than relying solely on an individual investigator's prior mentoring record. Relevant evidence may include mentoring teams, institutional resources, plans for mentor development, and access to complementary scientific and career expertise. This approach would allow early-career investigators to develop as mentors while ensuring appropriate support for trainees, and it would recognize high-quality mentorship at institutions that have fewer formal NIH training programs.

Professional societies can complement institutional mentoring by providing cross-institutional networks, interdisciplinary engagement, and leadership development. For example, the USASP Leadership Academy pairs established pain researchers and clinicians with midcareer investigators to develop leadership and professional skills that may not be available within their home institutions. NIH could partner with professional societies to evaluate and expand approaches of this kind while preserving institutional responsibility for mentorship and career development.

Supporting Sustainable Career Pathways. Training programs should recognize successful careers in clinical practice, industry, government, nonprofit organizations, research administration, communication, and other sectors—not only tenure-track appointments or NIH-funded faculty positions. NIH could consider whether its indicators inadvertently encourage mentors to select trainees based on their predicted likelihood of achieving a narrowly defined academic outcome. NIH could also consider regularly assessing whether NRSA stipends and related support remain adequate relative to inflation, regional living costs, and institutional compensation requirements. Adequate and predictable support is necessary to retain talented researchers and realize the long-term value of NIH's training investments. Recognizing and sustaining varied career pathways would extend the benefits of NIH training across the broader research enterprise and strengthen the workforce needed to advance science and health.

Collaboration and Team-Based Science

USASP recommends that NIH assess collaboration by the scientific integration and added value it produces, rather than by the size or visibility of the team. Meaningful team science combines complementary expertise, methods, perspectives, and resources to address questions that investigators could not answer as effectively on their own. This approach would reward substantive collaboration rather than the appearance of interdisciplinarity.

Funding Integrated Team Science. We recommend that NIH expand upon its opportunities for investigator-initiated multidisciplinary research through U19 and comparable multi-project or cooperative mechanisms when scientific objectives require interdependent projects, shared cores, coordinated methods, and sustained collaboration. Planning awards, collaborative pilot funding, administrative supplements, and coordinating centers could support smaller or emerging teams. Allowable costs should recognize the infrastructure required for collaboration, including project management, team-science facilitation, shared communication and data platforms, methodological consultation, standards development, evaluation, and appropriate compensation for external partners.

Policies designed around individual investigators should not inadvertently discourage genuine multiple-PI projects. If NIH limits the number or value of awards held by an investigator, it should consider the investigator's actual effort, authority, and share of project leadership rather than treating participation in an integrated MPI team as equivalent to directing the project alone. Any exception should require evidence of substantive leadership to discourage nominal MPI designations.

Integrating Training with Team Science. Large, milestone-driven programs are not automatically effective training environments. Trainees may contribute substantially to predetermined deliverables without gaining experience identifying research gaps, developing hypotheses, or designing studies. NIH should encourage team-science awards to include trainee pilot funds, exploratory subprojects, mentored proposal development, and opportunities to lead cross-disciplinary working groups or propose questions arising from consortium resources.

The HEAL PAIN Cohort T90/R90 and NIDA T32 programs provides a useful model by combining interdisciplinary mentoring, authentic collaborative projects, career development, and scientific training across the spectrum of disciplines relevant to pain research. Comparable features could be incorporated into research consortia and multi-project awards, including programs not primarily designed as training mechanisms. This would allow trainees to develop both team-science skills and intellectual independence.

Assessing Collaboration Quality. Relevant indicators could include whether team members jointly shaped the research question and design; integrated distinct methods or disciplinary perspectives; established clear governance, communication, decision-making, authorship, and conflict-resolution practices; and produced knowledge, resources, or capabilities that would have been difficult to achieve independently. NIH could also consider the durability of relationships, development of shared methods or datasets, creation of new interdisciplinary directions, and dissemination of expertise beyond the original team. Contributions to community standards, metadata, protocols, and interoperability should be recognized when they enable collaboration across a field.

Recognizing Individual Contributions. Principal-investigator status, authorship order, and consortium authorship often obscure essential contributions from data curators, software developers, statisticians, project managers, research coordinators, standards developers, and domain experts. Structured contributor-role frameworks offer a useful foundation but may need greater specificity for metadata design, standards development, maintenance of shared tools, use-case development, and other enabling work. NIH should encourage structured, machine-readable contribution records across publications, datasets, software, models, and standards, supplemented, when necessary, by brief descriptions of the significance of each contribution.

Within consortia, funding and governance plans should prospectively recognize laboratories and investigators contributing data, validation, use cases, domain expertise, or standards development even when they do not build or publish the final resource. Appropriate subaward support, protected effort, and formal credit would reduce friction and ensure that essential collaborative work is not treated as invisible service.

Entrepreneurship and Translation

Defining NIH's Role. NIH can most effectively advance translation by reducing the scientific, technical, and coordination risks that prevent credible discoveries from reaching capable developers and users. Its role should be to support validation, development, training, and connections to appropriate expertise—not to presume that every discovery should become a commercial product or that NIH should operate as an entrepreneur. Support should reflect the type and maturity of the innovation and connect investigators with relevant regulatory, manufacturing, clinical, implementation, and commercialization expertise.

Matching Support to Development Requirements. Therapeutic development often requires more time and capital than conventional small-business or milestone-driven awards provide. Validation, toxicology, manufacturing, regulatory engagement, and clinical studies may be required before private investment, or an industry partnership becomes feasible. NIH should evaluate whether SBIR/STTR, UG3/UH3, and related mechanisms provide realistic budgets, timelines, and continuity for the development stages they are intended to support. Where persistent gaps exist, NIH could provide larger or longer awards, bridge support between development stages, or complementary mechanisms extending beyond conventional limits. Continued funding should remain tied to clearly defined scientific, technical, regulatory, and development milestones, with independent decisions to continue, revise, or discontinue support.

NIH should also expand access to training and technical assistance in regulatory strategy, manufacturing, clinical development, reimbursement, intellectual property, and commercialization planning. These resources may be particularly valuable for academic investigators and first-time entrepreneurs who have strong scientific expertise but limited product-development experience.

Designing for Adoption. Technical quality does not ensure that a product, tool, standard, or infrastructure resource will be used. NIH should encourage translation-focused projects to identify intended users and potential adopting organizations early enough for them to help define requirements, test usability, establish implementation milestones, and plan for the distribution of knowledge into the community of potential users. Including affected communities and users early would help to ensure that the presentation of data is made understandable to non-researchers. Joint funding or shared milestones between developers and adopting partners could create accountability for both development and implementation. Expectations should vary by project maturity so that early engagement demonstrates a credible need without requiring a fully committed commercial partner before an innovation is sufficiently developed.

Improving University-to-Startup Translation. The time, cost, and uncertainty involved in licensing university-owned technologies can prevent promising discoveries from reaching qualified startups. NIH-funded institutions should be encouraged to use transparent licensing processes, reasonable timelines, standardized startup agreements, proportionate financial terms, and procedures for resolving stalled negotiations. More predictable terms would allow

early-stage companies to invest limited resources in validation and development while preserving appropriate institutional and public interests in federally supported discoveries.

Measuring Translational Contribution. Translational impact should reflect whether an innovation responsibly advanced toward practical application, not simply the number of patents, licenses, startup companies, or dollars raised. Relevant evidence may include completion of scientific, regulatory, manufacturing, or clinical milestones; engagement of qualified development or adoption partners; integration into independent research or clinical workflows; and appropriate follow-on support. Unprompted adoption by organizations outside the originally funded program will provide especially meaningful evidence that a product, tool, or standard addressed a broader need.

Indicators should account for differences among therapeutics, devices, diagnostics, software, research tools, communication into the general community, and implementation strategies. NIH should also recognize well-conducted work demonstrating that a product is ineffective, unsafe, infeasible, or not ready for further development. Timely discontinuation based on credible evidence can prevent greater downstream losses and should not automatically be treated as failure.

Foundational Scientific Exploration

Basic research and high-risk, high-reward research are both essential to scientific progress but should not be treated as interchangeable. Basic research builds the knowledge, methods, and conceptual foundations upon which future advances depend; much of it is incremental and cumulative rather than inherently high-risk. High-risk research deliberately tests unconventional ideas whose probability of success is uncertain, but whose potential scientific effect is substantial. These forms of exploration require distinct funding and review strategies.

Sustaining Foundational Basic Research. Foundational research should not be required to demonstrate an immediate path to a product, intervention, or clinical outcome. Its value may become apparent only after multiple discoveries converge, sometimes many years later. Research on incretin physiology and glucagon-like peptides ultimately enabled GLP-1–based therapies, while studies of bacterial immune systems provided the foundation for CRISPR-Cas9 gene editing (NIDDK, 2021; NIGMS, 2024). Neither trajectory could have been fully predicted when the underlying observations were first investigated.

NIH should maintain stable support for investigator-initiated basic research alongside goal-directed, translational, and clinical programs. Tightly defined initiatives can mobilize research around urgent needs, but excessive emphasis on predetermined outcomes may narrow the questions being investigated and reduce opportunities for unexpected discovery. NIH should transparently monitor its investments across these categories and ensure that every institute and center sustains the foundational science needed for future progress.

Foundational science should also include careful incremental research that validates methods, resolves inconsistencies, characterizes mechanisms, refines models, or establishes the boundaries of a finding. NIH should not equate innovation solely with novelty or dramatic claims. Relevant indicators of foundational contribution may include creation of new concepts,

methods, measures, models, datasets, or research directions; resolution of scientific uncertainty; or eventual uptake across disciplines. Because these effects may emerge unpredictably, investigators should not be required to forecast a specific clinical or commercial impact.

Supporting High-Risk, High-Reward Research. High-risk research requires review processes that tolerate genuine uncertainty without relaxing methodological standards. Although NIH's R21 and High-Risk, High-Reward programs state that preliminary data are not required (NIH, 2026a; NIH, 2026b), applicants may still perceive that reviewers expect conventional evidence of feasibility. NIH should examine reviewer critiques, scoring patterns, and funding decisions to determine whether preliminary-data expectations are being reintroduced in practice or whether projects described as innovative remain comparatively conventional.

Reviewer training should distinguish the uncertainty inherent in testing an unconventional idea from avoidable weaknesses in study design. Review criteria could separately consider the importance of the question, degree of conceptual departure, potential magnitude of the advance, credibility of the scientific rationale, rigor of the approach, and value of the knowledge generated even if the central hypothesis is not confirmed. NIH should recruit reviewers experienced in evaluating unconventional or cross-disciplinary research while ensuring that high-risk mechanisms remain accessible to early-career investigators and researchers entering new fields.

Funding for Learning. High-risk programs should be structured around learning and decision-making rather than the expectation that every aim will succeed. Staged awards could use learning milestones, flexible follow-on funding, and opportunities to modify direction when early findings challenge the original premise. NIH should evaluate these mechanisms as portfolios by asking whether they support ideas unlikely to succeed in conventional review, generate new concepts or methods, and occasionally produce unusually substantial advances. Projects that disprove plausible hypotheses, identify unworkable approaches, or establish unexpected limitations can still provide a meaningful return by preventing unproductive investment and informing future research.

Public Impact

USASP supports NIH's efforts to recognize the many ways biomedical research benefits the public. Public impact is multidimensional, often collective, and may emerge long after an award ends. Relevant evidence could include improvements in health outcomes or clinical practice; incorporation into guidelines, policy, or health-system operations; development of accessible tools and educational resources; meaningful use by people affected by health conditions, clinicians, educators, or community organizations; economic or workforce benefits; and contributions to public understanding of science. Public visibility, social-media reach, or favorable attention alone should not be treated as evidence of meaningful impact.

Building Capacity for Dissemination. Public impact rarely follows from publication alone. Researchers may lack the time, expertise, partnerships, or resources needed to communicate findings to those who could use them. We appreciate efforts like NIH ENGAGE that focus on

integrating people with lived experience into research from early stages through to translation. But with an evolution in practice, capacity building is essential. NIH should provide training and dedicated support for plain-language writing, visual communication, responsible use of social media, and dissemination to people living with health conditions, clinicians, community health workers, educators, policymakers, and other intended users. These skills should be incorporated into training and career-development programs and project dissemination plans. Grant funds should also be available to support communication specialists and compensate external partners who help adapt and disseminate findings.

Creating Accessible Scientific Resources. Well-curated repositories can extend public impact by offering accessible pathways to NIH-supported research. Where privacy, consent, and other protections permit, repositories could provide plain-language study summaries, public-facing explanations connecting data to research questions, educational materials grounded in active research, appropriately deidentified classroom datasets, and channels through which patient and community organizations can ask questions or identify areas of interest. Evidence of educational use, adoption by public-serving organizations, or incorporation into accessible tools may help demonstrate impact more meaningfully than website traffic or download counts alone.

Recognizing and Sharing Effective Practices. NIH could establish an annual recognition program honoring investigators, research teams, research staff, communicators, and community partners whose work exemplifies meaningful public impact. Awards could be named for distinguished former NIH leaders whose careers reflected public service and the application of scientific knowledge for societal benefit. Recognition need not include monetary support; an opportunity to present the work publicly and dissemination of the recipient's approach as a case study could provide meaningful visibility.

Selection should focus on documented contribution rather than institutional reputation, publication metrics, or public prominence. Publishing recipients' strategies, challenges, and lessons learned would make the program more than an award by helping NIH identify and disseminate effective approaches throughout the biomedical research community.

Conclusion

USASP appreciates NIH's consideration of these recommendations. We encourage NIH to develop context-sensitive measures that recognize varied scientific contributions, preserve expert judgment, and are adequately resourced and evaluated before broad implementation.

Sincerely,

[redacted]

Brittany Knight, PhD
Executive Director & CEO
United States Association for the Study of Pain

Literature Cited

ENTRUST-PE Network, O'Connell, N. E., Belton, J., Crombez, G., Eccleston, C., Fisher, E., Ferraro, M. C., Hood, A., Keefe, F., Knaggs, R., Norris, E., Palermo, T. M., Pickering, G., Pogatzki-Zahn, E., Rice, A. S., Richards, G., Segelcke, D., Smart, K. M., Soliman, N., Stewart, G., … Williams, A. C. C. (2025). Enhancing the trustworthiness of pain research: A call to action. *The journal of pain*, *28*, 104736. https://doi.org/10.1016/j.jpain.2024.104736

Haroutounian S, Holzer KJ, Kerns RD, Veasley C, Dworkin RH, Turk DC, Carman KL, Chambers CT, Cowan P, Edwards RR, Eisenach JC, Farrar JT, Ferguson M, Forsythe LP, Freeman R, Gewandter JS, Gilron I, Goertz C, Grol-Prokopczyk H, Iyengar S, Jordan I, Kamp C, Kleykamp BA, Knowles RL, Langford DJ, Mackey S, Malamut R, Markman J, Martin KR, McNicol E, Patel KV, Rice ASC, Rowbotham M, Sandbrink F, Simon LS, Steiner DJ, Vollert J. Patient engagement in designing, conducting, and disseminating clinical pain research: IMMPACT recommended considerations. Pain. 2024 May 1;165(5):1013-1028. doi: 10.1097/j.pain.0000000000003121. Epub 2023 Dec 19. PMID: 38198239; PMCID: PMC11017749.

Knopp KL, Stenfors C, Baastrup C, Bannon AW, Calvo M, Caspani O, Currie G, Finnerup NB, Huang W, Kennedy JD, Lefevre I, Machin I, Macleod M, Rees H, Rice ASC, Rutten K, Segerdahl M, Serra J, Wodarski R, Berge OG, Treede RD. Experimental design and reporting standards for improving the internal validity of pre-clinical studies in the field of pain: Consensus of the IMI-Europain consortium. Scand J Pain. 2015 Apr 1;7(1):58-70. doi: 10.1016/j.sjpain.2015.01.006. PMID: 29911603.

Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature.* 2012;490(7419):187–191. doi:10.1038/nature11556.

Lalu, M. M., Presseau, J., Foster, M. K., Hunniford, V. T., Cobey, K. D., Brehaut, J. C., Ilkow, C., Montroy, J., Cardenas, A., Sharif, A., Jeffers, M. S., & Fergusson, D. A. (2023). Identifying barriers and enablers to rigorous conduct and reporting of preclinical laboratory studies. *PLoS biology*, *21*(1), e3001932. https://doi.org/10.1371/journal.pbio.3001932

Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, et al. (2015) Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. PLOS Biology 13(10): e1002273. https://doi.org/10.1371/journal.pbio.1002273

National Institute of Diabetes and Digestive and Kidney Diseases. *Story of discovery: How different medications for diabetes and obesity emerged from basic research on one pancreatic hormone.* 2021. NIDDK

National Institutes of Health. *Exploratory/Developmental Grants (R21).* NIH R21 (2026a)

National Institutes of Health Common Fund. *High-Risk, High-Reward Research.* NIH Common Fund (2026b)

National Institute of Neurological Disorders and Stroke. *Rigor Champions and Opportunities.* Updated June 25, 2026. https://www.ninds.nih.gov/current-research/trans-agency-activities/rigor-transparency/rigor-champions-and-opportunities

Smith K. *What Is CRISPR?* National Institute of General Medical Sciences Biomedical Beat. 2024. NIGMS

231. Coalition to Illuminate and Address Animal Methods Bias

I am responding: On behalf of an organization

Name: Catharine E. Krebs, PhD

Name of Organization: Coalition to Illuminate and Address Animal Methods Bias

Type of Organization: Other

Type of Organization - Other: Coalition of researchers and NGO advocates

Role: Other

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/COLAAB-Response-NOT-OD-26-087-2026-08-19.pdf

COLAAB
Coalition to Illuminate and Address Animal Methods Bias
5100 Wisconsin Ave. NW, Suite 400, Washington, DC 20016
info@animalmethodsbias.org • animalmethodsbias.org

August 19, 2026

Lyric Jorgenson, Ph.D.
Director, Office of Science Policy
National Institutes of Health

Re: NOT-OD-26-087: Request for Information on Measuring and Rewarding Scientific Impact

Submitted via comment form: https://osp.od.nih.gov/comment-form-measuring-and-rewarding-scientific-impact/

Dear Director Jorgenson:

We are writing on behalf of the Coalition to Illuminate and Address Animal Methods Bias (COLAAB) in response to the National Institutes of Health (NIH) request for information on measuring and rewarding scientific impact. Thank you for the opportunity to provide input.

The COLAAB is an international collaboration of researchers and advocates working to develop strategies to address the phenomenon of "animal methods bias," when research is unfairly evaluated because it does not include animal-based experiments.[1] As recognized by NIH's initiative to prioritize human-based research technologies and reduce animal use, nonanimal biomedical research methods like tissue chips and organoids are increasingly effective and available for disease modeling and drug development, but peer review remains one barrier to their broader uptake.[2] Members of the COLAAB have personally experienced animal methods bias, and empirical evidence gathered by our team demonstrates that it can cause delays in publication or rejection of manuscripts or grant applications, negatively impacting scientists' productivity and effectively disincentivizing the use of nonanimal approaches.[3]

[1] https://www.animalmethodsbias.org/
[2] National Institutes of Health. NIH to prioritize human-based research technologies. nih.gov. Posted April 29, 2025. Accessed August 12, 2026. https://www.nih.gov/news-events/news-releases/nih-prioritize-human-based-research-technologies; Bhattacharya, J. Advancing NIH's Mission Through a Unified Strategy. National Institutes of Health. Published August 15, 2025. Accessed August 13, 2026. https://www.nih.gov/about-nih/nih-director/statements/advancing-nihs-mission-through-unified-strategy
[3] Krebs CE, Lam A, McCarthy J, Constantino H, Sullivan K. A survey to assess animal methods bias in scientific publishing. *ALTEX*. 2023;40(4):665-676. doi:10.14573/altex.2210212; Parvatam S, Kaushik K, Mahadik K, et al. A

To incentivize and reward rigorous, translatable, impactful research that delivers value to the American public, human-based research should continue to be prioritized agency-wide, and further action should be taken to mitigate animal methods bias. As requested, below, we provide specific recommendations for measurable indicators and incentives of NIH-funded research that is collaborative, rigorous, and mission-driven.

1. RIGOR AND REPRODUCIBILITY

Recommendations to incentivize rigor and reproducibility:

A. Implement grant application review criteria to assess proposals for adherence to NIH priorities, human relevance, context of use, and translatability. Animal methods bias undermines evaluations of rigor by allowing the methodological familiarity or gaps in reviewer expertise to unfairly influence assessments of scientific rigor. We urge the agency to recognize that methodological suitability is a key factor of rigorous and reliable science. A method should be fit for its intended purpose, regardless of its historical prominence or familiarity. NIH should distinguish successful replication from evidence of external validity, human relevance, or translational value. A finding can be reproducible within a particular experimental model without that model accurately representing human biology or predicting human outcomes. Furthermore, NIH should encourage reviewers to be critical of applicants' claims about the necessity and past contribution of an animal-based method. Correlation is no measure of necessity or validity (animals' involvement in successful drug development does not mean success depended on their involvement).[4]

B. Require reviewers to provide substantive, referenced reasoning if they argue that lacking animal experiments is a weakness of any particular proposal.[5] A study should not be considered *more* rigorous because it includes animal use, nor *less* rigorous because it does not.

C. Continue to evaluate reviewer performance for evidence of methodological bias.[6]

---

survey to evaluate animal methods bias experienced by India-based researchers in the peer review of manuscripts and grant applications. *NAM Journal*. Published online August 6, 2025:100042. doi:10.1016/j.namjnl.2025.100042

[4] Matthews RAJ. Medical progress depends on animal models - doesn't it? *J R Soc Med*. 2008;101(2):95-98. doi:10.1258/jrsm.2007.070164

[5] Vinken M. Advancing toxicology through new approach methodologies. *Toxicology*. 2026;526:154546. doi:10.1016/j.tox.2026.154546

[6] Yee LM and Olufokunbi Sam D. CSR Advisory Council Reviewer Evaluation Working Group Update. *Center for Scientific Review Advisory Council Meeting.* September 23, 2024. Accessed August 13, 2026. https://public.csr.nih.gov/sites/default/files/2024-08/Update_on_CSRAC_Working_Groups_Reviewer_Evaluation.pdf

D. Incentivize replication in human-based systems, including with human cells, organoids, organ chips, computational models, clinical datasets, or convergent evidence across complementary human-based approaches.[7]
E. Encourage and reward researchers to follow best practice guidelines for reporting *in vitro* studies.[8]

Recommendations to measure rigor and reproducibility:

A. Measure the proportion of individuals with expertise in human-based and nonanimal methods within study sections, advisory committees, and review staff.
B. Track and investigate reviewer comments that judge lacking animal experiments as a weakness.
C. Track rates of methodological bias within reviewer evaluations.
D. Track funding rates of proposals with animal- and nonanimal-based methodological approaches, and evaluate the relationship of those findings with review group composition.

2. DATA, SOFTWARE, AND MODEL SHARING

Recommendations to incentivize data, software, and model sharing:

A. Incentivize and reward open science practices, such as open access publishing, systematic reviews, data sharing, and preregistration of studies. This will foster transparency, rigor, and reproducibility, helping to build confidence in nonanimal methods and mitigate animal methods bias by addressing the hesitancy of some reviewers to embrace nonanimal methods.
B. Incentivize and reward data, software, and model sharing to help researchers validate findings without conducting new animal experiments. One of the primary drivers of animal methods bias is the reviewer expectation that nonanimal findings must be "validated" using animals. On August 4th, the National Cancer Institute (NCI) published new Guidance on Applications for Studies with Animal Models and included a requirement that grant applicants should include human-based validation for animal experiment findings.[9] NIH

---

[7] Ingber DE. Is it Time for Reviewer 3 to Request Human Organ Chip Experiments Instead of Animal Validation Studies? *Adv Sci (Weinh)*. 2020;7(22):2002030. doi:10.1002/advs.202002030

[8] National Centre for the Replacement Refinement & Reduction of Animals in Research. DRIVER recommendations. Accessed July 23, 2026. https://nc3rs.org.uk/3rs-resources/driver-recommendations; Organisation for Economic Co-operation and Development. Guidance Document on Good In Vitro Method Practices (GIVIMP), Second Edition. Published online December 15, 2025. Accessed December 18, 2025. https://doi.org/10.1787/5ba6777b-en; Mohapatra R, Leist M, Aulock S von, Hartung T. Guidance for Good In Vitro Reporting Standards (GIVReSt) – A draft for stakeholder discussion and background documentation. *ALTEX - Alternatives to Animal Experimentation*. 2025;42(3):376-396. doi:10.14573/altex.2507041; Pamies D, Leist M, Coecke S, et al. Guidance document on Good Cell and Tissue Culture Practice 2.0 (GCCP 2.0). *ALTEX - Alternatives to Animal Experimentation*. 2022;39(1):1. doi:10.14573/altex.2111011

[9] Letai A. Guidance on studies with Animal Models. National Cancer Institute. August 4, 2026. Accessed August 14, 2026. https://www.cancer.gov/grants-training/nci-bottom-line-blog/2026/guidance-animal-models

should implement NCI's requirement agency-wide, encourage researchers to use clinically relevant forms of validation, and enable this by incentivizing the sharing of high-quality human datasets, computational tools, and experimental models.

C. Incentivize and reward publishing null (or inconsistent) results through measures like requirements to report negative results and de-emphasis of journal-based metrics in evaluations of investigator capabilities. Publication bias against null studies has resulted in the scientific literature being dominated by studies that reject the null hypothesis, an absence of complete data in the scientific record, and—consequently—a threat to the rigor of systematic reviews. This jeopardizes the accurate assessment of interventions and model validity. Existing scientific literature may present an inaccurately positive portrayal of experiments on animals, where studies demonstrating the failure of those experiments to answer research questions or replicate human biology have been excluded from the record. Bias against null studies can therefore distort interpretations and opinions of experiments on animals, particularly regarding their precision and translatability, thus contributing to animal methods bias in the form of reviewers' expectations or requests for animal studies—including ones that have already been carried out but never published.[10]

Recommendations to measure data, software, and model sharing:

A. Measure open access publishing and preregistration rates of NIH-funded research.
B. Track null result publications associated with grants.

3. TRAINING AND MENTORSHIP

Recommendations to incentivize training and mentorship:

A. Continue requiring bias mitigation training for all reviewers that includes information on methodological biases.
B. Incentivize training in nonanimal approaches.
C. Continue to support and promote the Early Career Reviewers (ECR) Program and make a special effort to recruit ECRs who have experience in human-based technologies.

Recommendations to measure training and mentorship:

A. Track funds toward training in nonanimal approaches.
B. Track early-career researcher participation in review groups.
C. Survey early-career researchers about any experiences where they felt pressure to conduct animal experiments—from reviewers, advisors, mentors, peers, or other sources—to understand the needs of trainees to overcome animal methods bias. The time cost of grant

[10] Trunnell ER and Krebs CE. Chapter 15 Overcoming animal methods bias and its impacts on nonanimal research assessments. In: *Moving Toward Nonanimal Approaches in Medical Research and Testing*. Elsevier, 2026 (In press). ISBN: 9780443265143

and manuscript rejection and resubmission associated with animal methods bias can have especially severe consequences on early career researchers. Furthermore, although the intersection of animal methods bias and person-based biases (including related to age or seniority) has yet to be directly studied, the hierarchical structure of academic research could mean that early career researchers may experience animal methods bias in a compounded manner. When confronted with the pressure to use animals because of institutional inertia, expectation to obtain funding and publish, or direct requests from advisors, reviewers, or others, early career researchers may be less empowered to push back on these directives or requests, making them more susceptible to their harmful effects.

4. COLLABORATION

Recommendation to incentivize collaboration:
To develop and sustain an interdisciplinary research workforce capable of advancing the development, usability, qualification, and application of nonanimal, human-based technologies, NIH should incentivize and reward interdisciplinary training, research teams, research programs, research centers, and core facilities. Skills across fields such as medicine, clinical research, bioengineering (including microfabrication and microfluidics), pharmacology, toxicology, computational biology, and AI will be crucial.[11]

Recommendation to measure collaboration:
Measure not just disciplinary diversity but also methodological diversity of collaborative science.

5. ENTREPRENEURSHIP AND TRANSLATION

Recommendation to incentivize entrepreneurship and translation:
Translatability is a key component of rigorous, reproducible, and reliable methods and should be evaluated and rewarded (see response to 1. Rigor and Reproducibility).

Recommendations to measure entrepreneurship and translation:
A. Measure translation rates and contributing factors through systematic reviews and meta-analyses.
B. Track funding, uptake, and replication of translatable nonanimal systems.
C. Track the application of nonanimal systems to preclinical, clinical, and regulatory workflows, as well as the evidence that they contributed to decisions, products, policies, or improvements in human health.

[11] Moruzzi A, Shroff T, Keller S, Loskill P, Cipriano M. Training the Next Generation of Researchers in the Organ-on-Chip Field. *Education Sciences*. 2023;13(2):2. doi:10.3390/educsci13020144

6. FOUNDATIONAL SCIENTIFIC EXPLORATION

Recommendation to incentivize foundational scientific exploration:
When innovative human-based technologies are used for foundational scientific exploration, they should be supported and stand on their own, not subjected to animal methods bias. For some, confidence in these approaches is still being established, an inherently high-risk, high-reward process. They must, therefore, be further embedded within the academic enterprise to build evidence, familiarity, and ensure continued support for bold and innovative research directions.

Recommendations to measure foundational scientific exploration:
Track the use of human-based technologies for basic research.

7. PUBLIC IMPACT

Recommendation to incentivize public impact:
Encourage the practice of researcher assessment to move away from journal-based metrics like impact factors and toward scientific content, merit, and real-world impact on medical practice, health outcomes, societal benefit, and public policy.[12] COLAAB studies have found that researchers who use nonanimal approaches feel pressure to add animal experiments to their studies to ensure publication, especially from journals with high impact factors; and when authors do not comply with reviewer requests for animal experiments, they report publishing in lower impact journals among other consequences. The narrow focus of research assessments on journal-based metrics like impact factors penalizes research that counters the status quo, including the animal experimentation paradigm that is upheld by decades of embedded preference, infrastructure, and culture, as well as entire industries with vested interests in maintaining it.

Recommendation to measure public impact:
Investigate animal methods bias by assessing reviews and funding rates of projects with animal- and human-based methods. Based on preliminary evidence from a pilot study and workshop we conducted, at least some NIH grant application reviews may be impacted by animal methods bias.[13] However, the full extent of this impact, including how it affects funding rates of

[12] Allen L, Barbour V, Borrell-Damián L, Cobey K, Dube N, Firth C, et al. A Practical Guide to Implementing Responsible Research Assessment at Research Funding Organizations. Declaration on Research Assessment (DORA). May 2026. doi:10.5281/zenodo.18402271

[13] Trunnell ER, Roe KV. Predominance of animal-based expertise may bias NIH neuroscience grant review: A pilot study with implications for non-animal methodologies. *bioRxiv*. Preprint posted online March 2, 2025:2025.02.28.640877. doi:10.1101/2025.02.28.640877; Krebs CE, Geissler S, Herrmann K, et al. Exploring

animal- and human-based projects, remains to be determined. Understanding the breadth of this issue and its specific characteristics in the NIH context will be helpful for addressing it appropriately and holistically. We therefore recommend assessing review reports and reviewer comments during study section meetings for evidence of bias toward animal use and the resulting impacts on funding rates. Having already performed similar analyses with the aid of AI methods, the COLAAB would be pleased to assist with this effort. We have found large language models (LLMs) to be very useful for screening review reports for potential animal methods bias. However, a locally hosted LLM or cloud environment with privacy and confidentiality agreements may be needed to use such an approach for grant application reviews due to confidentiality restrictions; the compute power available will determine how large a reasoning model can be queried.

Thank you for your attention to these recommendations. We welcome further collaboration to ensure their implementation.

Sincerely,

[redacted]

Catharine E. Krebs, Ph.D.
Physicians Committee for Responsible Medicine

[redacted]

Emily R. Trunnell, Ph.D.
People for the Ethical Treatment of Animals

animal methods bias in biomedical research funding: Workshop proceedings and action steps. *NAM Journal*. 2025;1:100004. doi:10.1016/j.namjnl.2024.100004

232. People for the Ethical Treatment of Animals

I am responding: On behalf of an organization

Name: Emily R. Trunnell, Ph.D.

Name of Organization: People for the Ethical Treatment of Animals

Type of Organization: Non-Profit Research Organization

Role: Bioethicist

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Experiments on animals have long been recognized as a major contributor to the reproducibility crisis in biomedical research. Former NIH leadership acknowledged that "[p]reclinical research, especially work that uses animal models, seems to be the area that is currently most susceptible to reproducibility issues" (https://doi.org/10.1038/505612a).

In addition to the harms inflicted on animals and poor translation, irreproducible experiments impose substantial costs on science and the public. Failed attempts to replicate findings can consume years of effort and millions of dollars while delaying the development of effective interventions. Importantly, irreproducibility is an indicator that certain approaches are not generating sufficiently reliable knowledge to justify continued investment.

NIH must prioritize metrics that identify, evaluate, and replace methods that fail to produce reproducible, translatable, and clinically meaningful results.

NIH should establish dedicated funding mechanisms for evidence synthesis and comparative assessments of research methods, such as that which can be generated through systematic reviews. Systematic reviews can assess whether methods are scientifically valid, translatable, and relevant to human biology (https://brill.com/edcollbook-oa/title/35072; https://ntp.niehs.nih.gov/sites/default/files/2024-03/VWG_Report_27Feb2024_FD_508.pdf; https://doi.org/10.1177/01410768221093551).

Extensive data already exist without the need to reproduce every experiment at the benchtop. Researchers who identify ineffective approaches and facilitate their replacement should be recognized and rewarded for these important scientific contributions.

Reproducibility is also influenced by the quality and consistency of the biological reagents used in research. Antibodies and cell culture media can be significant sources of experimental variability when they are derived from irreproducible sources or are poorly characterized. NIH should support efforts to improve the adoption and characterization of reliable research reagents, including sequence-defined non-animal recombinant antibodies and chemically-defined cell culture media. NIH already recognizes the importance of this issue through its requirements that grant applicants address the authentication of key biological and chemical resources used in proposed research. However, beyond requiring researchers

to describe authentication plans, NIH should establish targeted funding opportunities, incentives, and reporting metrics that promote the development, validation, characterization, and adoption of more reliable reagents. Measurable indicators could include increased funding and use of sequence-defined antibodies and chemically-defined cell culture medium.

2. Data, Software, and Model Sharing: In addition to the sharing of data and software, NIH should reward investigators and institutions for sharing information that enables the scientific community to identify superior research approaches and discontinue ineffective ones. This could include creating databases that catalog human-based research methods, sharing data on the predictive performance and validation of human-based models, or developing data, tissue, cell, and other model-related repositories that facilitate the replacement of experiments on animals. NIH's Standardized Organoid Modeling Center is an excellent start, and efforts like this must be expanded for other human-based technologies.

NIH should incentivize the creation of openly accessible resources that help researchers identify and adopt non-animal methods. Making data on model performance publicly available would enable more informed funding decisions, facilitate meta-research, and accelerate the replacement of experiments on animals, which fail to advance human health.

3. Training and Mentorship: NIH should reward mentors and institutions that prepare researchers to use the most rigorous, reproducible, and human-relevant approaches available. Metrics such as publication output and grant success do not assess whether trainees are equipped to advance biomedical research using methods that are predictive of human biology.

NIH should incentivize training programs that support the development and adoption of non-animal methods, including organoids, organs-on-chips, computational models, artificial intelligence tools, and other human-based approaches. Indicators of successful training and mentorship could include participation in cross-disciplinary programs, acquisition of skills in emerging technologies, and transition from experiments on animals to human-based methods.

Mentors who help trainees critically evaluate the strengths and limitations of existing research methods should also be recognized for advancing scientific rigor and training future scientists in the important discipline of meta-research. Likewise, institutions should receive credit for providing researchers with the knowledge, infrastructure, and support needed to replace experiments on animals with modern human-based approaches.

As part of efforts to build institutional capacity, NIH should support educational opportunities for individuals involved in research oversight, including Institutional Animal Care and Use Committee (IACUC) and Institutional Review Board (IRB) members. Online training programs focused on non-animal methods, emerging human-relevant technologies, and the scientific and regulatory limitations of animal models would help to better equip IACUCs to evaluate animal use proposals in light of available non-animal methods and advances in biomedical science.

NIH should also support retraining opportunities for established investigators whose careers have been built around animal experimentation. Providing scientists with the resources and incentives needed to adopt human-relevant methods would strengthen the biomedical workforce and accelerate the

transition away from experiments on animals.

For more information on this topic, see “Developing a global education hub for animal-free innovation,” by Janssens et al. (https://doi.org/10.14573/altex.2411251).

4. Collaboration: Replacing experiments on animals will be most successful when we combine expertise that spans multiple disciplines, including cell biology, omics, bioengineering, materials science, computational biology, data science, and artificial intelligence. Advances such as organoids, organs-on-chips, digital twins, and AI-enabled drug discovery depend on the integration of knowledge, tools, and datasets across these fields. NIH should therefore recognize and reward collaborative projects that bring together diverse scientific expertise to further develop and apply human-based approaches. Particular recognition should be given to collaborative efforts that generate shared resources or produce evidence demonstrating that human-based methods can replace experiments on animals and facilitate the transition away from them.

5. Entrepreneurship and Translation: Translation should be measured by whether a research approach improves prediction of human outcomes.

Drug candidates typically progress through a development pipeline in which experiments on animals remain the norm. Basic and translational research often involve animal experiments, the most commonly used models of human disease are not human, and preclinical development has historically required extensive animal testing. The result of this status quo has been 95% of novel drugs failing in clinical trials (https://ncats.nih.gov/research/research-activities/ntu). After decades of dependence on animal use, such a failure rate cannot reasonably be described as a successful translational paradigm. The failure of this system to deliver translational impact, and the need for it to change, are clear.

Entrepreneurial and translational impact should be assessed based on the development, validation, and adoption of better methods that improve the prediction of human responses. Researchers, institutions, and companies that develop approaches capable of replacing experiments on animals should be recognized for advancing more effective pathways from discovery to clinical application. NIH should also reward innovations that reduce the cost, time, and high attrition rates that have long stymied drug development.

6. Foundational Scientific Exploration: Human-based technologies such as organoids, organs-on-chips, advanced imaging, computational modeling, and AI are enabling researchers to study human biology directly, capture patient-specific variation, and answer questions that experiments on animals cannot. These approaches have the potential to transform basic research itself by providing unprecedented insight into human development, disease, and physiology and should be prioritized when considering foundational research applications.

Expanding the scientific community's ability to study human biology directly should also be recognized as a paradigm-shifting contribution to basic research and the biomedical research enterprise. NIH should therefore consider projects that develop, improve, or apply these technologies as essential to foundational scientific exploration. Scientific impact should be measured not only by discoveries made using existing research tools, but also by the creation of new capabilities that expand what biomedical research can achieve.

7. Public Impact: Public impact should include transparency, accountability, and alignment with public expectations regarding how taxpayer-funded research is conducted. The public reasonably expects that federally funded research adheres to the public's views on morality, uses the most scientifically advanced approaches available, and that research investments are regularly evaluated to ensure they continue to serve the public interest. In terms of the public's views on morality, opposition to the use of animals in experiments has increased steadily. In 2026, Gallup reported that only 45% of Americans felt that medical testing on animals was morally acceptable, down from 65% in 2001 (https://news.gallup.com/poll/710981/moral-acceptability-falls-several-behaviors.aspx; https://news.gallup.com/opinion/polling-matters/694550/trends-adults-acceptance-moral-values-behaviors.aspx).

NIH must recognize that improving public impact requires discontinuing practices that no longer meet scientific and ethical standards. Research that identifies areas where experiments on animals can be ended and provides evidence-based pathways for modernization can help ensure that public funds are directed toward approaches most likely to benefit human health.

NIH should also recognize efforts to increase transparency regarding the methods it funds and the outcomes those methods produce. This should include public reporting on research portfolios, assessments of the scientific value and human relevance of funded approaches, and measures of progress toward reducing experiments on animals, including quantitative tracking of the number of animals used in NIH-supported research and the purposes for which they are used.

Public impact should also encompass activities that increase public confidence in the biomedical research enterprise. Scientists who openly evaluate the strengths and limitations of existing research methods, engage with public concerns, and/or facilitate the transition from experiments on animals to human-based methods provide a valuable service to both science and society.

A meaningful indicator of public impact would be whether NIH-funded research becomes more transparent, more accountable to taxpayers, and more responsive to scientific advances. Efforts to replace experiments on animals with modern human-based approaches can advance all three goals by aligning research practices with both scientific progress and growing public expectations.

8. Other Areas: Animal methods bias occurs when research is judged less favorably because it does not include animal experiments, regardless of whether animal use is scientifically necessary. As human-based approaches such as organoids, organs-on-chips, computational models, and AI become increasingly capable of addressing biomedical research questions, researchers continue to report pressure from reviewers, funders, and mentors to conduct experiments on animals to secure funding.

Animal methods bias can distort scientific evaluation by conflating the inclusion of animal experiments with rigor, validity, or translational relevance. A research method should be assessed based on its fitness for purpose and ability to answer the research question at hand, not on its familiarity or historical prominence. Bias toward animal use may disadvantage innovative human-based approaches and discourage researchers from pursuing methods that are directly relevant to human biology.

NIH should recognize animal methods bias as a potential barrier to scientific progress and research

modernization. Understanding the extent to which bias influences peer review, funding decisions, publication outcomes, and career advancement would help ensure that researchers are rewarded for scientific merit and impact rather than adherence to established methodologies. Mitigating animal methods bias would support innovation, accelerate the adoption of human-relevant approaches, and better align research incentives with NIH's goal of advancing rigorous and impactful science. On this matter, NIH should consult the Coalition to Illuminate and Address Animal Methods Bias (www.animalmethodsbias.org).

233. American Chemical Society

I am responding: On behalf of an organization

Name: John P. Ochs

Name of Organization: American Chemical Society

Type of Organization: Other

Type of Organization - Other: Scholarly Society

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: The American Chemical Society (ACS) is a congressionally chartered nonprofit organization representing a global community of more than 262,000 individuals across academia, industry, and government. Guided by its vision of a world built on science, ACS advances scientific knowledge and research impact through programs and activities that serve educational and government institutions, businesses, and the broader public.

As a socially responsible organization deeply rooted in the scholarly community, ACS appreciates the opportunity to respond to NIH's request for information on broadening the ways scientific impact is measured and rewarded. We support efforts to identify and incentivize the characteristics of modern scientific research that most effectively advance discovery and societal benefit.

One of the most important features of 21st-century science that should continue to be measured, recognized, and incentivized is the publication of research findings as a clear, measurable output of scientific activity. We are concerned, however, that the RFI suggests publication and citation counts may not be well aligned with NIH's core mission to generate knowledge that enhances health, lengthens life, and reduces illness and disability.

We agree that no single metric can fully capture scientific impact, and publication and citation counts are no exception. Practices such as fragmenting research findings into multiple publications solely to increase publication counts should not benefit researchers and are explicitly discouraged by the ACS Ethical Guidelines to Publication of Chemical Research. However, it would be a disservice to the research

community to allow concerns about isolated misuse to diminish the value of publications as evidence of sustained scientific productivity and knowledge generation.

Similarly, citation counts have certain limitations. In some instances, citations may reflect narrow networks of scholarly acknowledgement. More commonly, however, they provide a meaningful indication that research findings are being used, evaluated, and extended by others. Citations often signal that new knowledge is influencing subsequent advances that contribute to human health and welfare, economic growth, and innovation.

Publication and citation metrics remain important because publication itself is fundamental to the scientific enterprise and directly supports NIH's mission to generate and disseminate knowledge in multiple ways.

First, scholarly journals provide a trusted and effective forum for the vetting, exchange, and dissemination of scientific ideas and discoveries. They enable researchers across academia, industry, and government to share knowledge, evaluate emerging findings, and identify new directions for research. Journals also serve as one of the primary mechanisms through which the data underlying scientific advances become accessible to the scientific community, American businesses, policymakers, and the public. These functions accelerate future discoveries, support innovation, and fuel U.S. scientific and economic leadership.

The value of research publications is already recognized by both federal agencies and the private sector. Articles published in ACS journals were downloaded 5.6 million times during 2024 and 2025 by federal agency staff in support of agency missions and evidence-based decision-making. ACS articles acknowledging federal funding were downloaded more than 293,000 times by U.S. corporations during the same period, demonstrating the role of published research in driving innovation and economic competitiveness. Researchers at U.S. academic institutions downloaded these articles more than 4.8 million times, further illustrating the importance of broad access to high-quality scientific information.

Second, rigorous peer review remains a cornerstone of scientific quality and integrity. The peer review process helps ensure that published research findings meet recognized standards of validity, transparency, and reliability. This provides scientists with a trusted body of knowledge upon which to build, reducing unnecessary duplication of effort and enabling more efficient progress. Peer-reviewed literature also plays an important role in fostering public confidence in science and in supporting evidence-based policymaking in areas of national importance, including those funded by NIH.

Third, scientific publications serve as a critical source of research and development knowledge for small and medium-sized enterprises that often lack substantial internal R&D capabilities. Access to rigorously reviewed scientific literature enables these organizations to benefit from advances generated across the research ecosystem and to translate scientific discoveries into products, services, and innovations that improve human health, create jobs, and contribute to economic growth.

For these reasons, publication and citation metrics should not be discounted as indicators of scientific impact. While they should not be the only standards used, they provide valuable signals regarding the

production, dissemination, and use of new knowledge. NIH should therefore adopt a balanced and nuanced approach that considers publication and citation data alongside other measures of impact when evaluating whether federally funded research is achieving its intended outcomes.

Such complementary measures may include indicators related to individual articles, journals, policy influence, data sharing, public engagement, technology transfer, clinical implementation, or other pathways through which scientific knowledge creates value.

Importantly, many of the additional indicators referenced in the RFI are already enabled by and embedded within the scholarly communication ecosystem. The competitive scholarly publishing community plays a critical role in ensuring that scientific information is rigorously evaluated, corrected when necessary, validated, preserved, interrelated, and disseminated. Scientific progress depends on a robust ecosystem that supports expert peer review, access to the scientific record, discovery of relationships between studies, and effective communication of research findings.

Knowledge cannot improve health, lengthen life, or reduce illness unless it is communicated, preserved, and made discoverable. The scholarly publishing ecosystem creates an organized, durable, and accessible scholarly record that enables current and future research and supports the scientific discoveries and interrelationships that drive societal progress. While publication and citation metrics should not be viewed as the sole indicators of impact, they provide objective evidence that knowledge is being generated, shared, utilized, and built upon by the broader research community.

Efforts to expand the measurement and recognition of scientific impact should build upon this foundation. NIH should work collaboratively with other federal funding agencies, research institutions, investigators, and the scholarly publishing community to develop additional standards that complement existing measures while preserving appropriate recognition of the central role that publication and scholarly communication play in advancing science and delivering public benefit.

8. Other Areas: N/A

234. Information Technology and Innovation Foundation

I am responding: On behalf of an organization

Name: Meghan Ostertag

Name of Organization: Information Technology and Innovation Foundation

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/2026-comments-nih-scientific-impact.pdf

ITIF | INFORMATION TECHNOLOGY & INNOVATION FOUNDATION

700 K Street NW Suite 600
Washington, DC 20001

COMMENTS OF ITIF

to the

National Institutes of Health

Washington, D.C.

In the Matter of: Request for Information on Measuring and Rewarding Scientific Impact	)	Notice No. NOT-OD-26-087

August 19, 2026

CONTENTS

Introduction and Summary 2
New Drug Development 3
Industry Funding of University R&D 4
Translational Research 6
Expand and Modify SBIR and STTR 9
International Scientific Collaboration 10
Production of STEM Talent 11
A Comprehensive Competitiveness Assessment For America's Biopharmaceutical Industry 11
Conclusion 12

INTRODUCTION AND SUMMARY

The Information Technology and Innovation Foundation (ITIF) is pleased to submit the following comments in response to the National Institutes of Health's (NIH's) Request for Information on Measuring and Rewarding Scientific Impact. ITIF is an independent, nonprofit, nonpartisan research and educational institute focusing on the intersection of technological innovation and public policy.

The biomedical industry is one of the most critical advanced industries in the world, supporting the development of biologics, pharmaceuticals, and medical devices to improve human health. The United States has been the world leader in the biotechnology industry, with the highest levels of new drug development, medical device production and exports, and the largest share of global output in the industry. But, as ITIF has written extensively, the People's Republic of China (PRC) is rapidly closing the gap with the United States, with the goal of becoming the world leader in life sciences, specifically in areas such as innovative drug development.[1] China has utilized every policy lever possible to eclipse the United States as the global leader in biomedical innovation and development, including legitimate practices such as research and development (R&D) investments, tax incentives, and talent development programs; as well as a number of illegitimate or illegal tools, according to international trade law, all of which have aided in China's goal of becoming a genuine competitor in the industry.

As the United States attempts to fight back against China's growing push for dominance in this industry (and many other high-tech sectors), much needs to change. But to understand which policy levers must be pulled and where the United States needs to advance, the ways the United States stimulates and measures innovation must change as well.[2]

---

[1] Stephen Ezell, Meghan Ostertag, and Sandra Barbosu, "China's Burgeoning Biopharmaceutical Competitiveness Demands a US Response" (ITIF, June 29, 2026), https://itif.org/publications/2026/06/29/chinas-burgeoning-biopharmaceutical-competitiveness-demands-us-response/.

[2] Robert D. Atkinson, "Testimony to the US Senate Subcommittee on Science, Manufacturing, and Competitiveness on "Measuring What Matters: Science, Standards, and Strategic Competition" (ITIF, July 21, 2026),

ITIF commends the administration and the NIH for seeking to improve the American scientific enterprise. In the United States, research productivity has continued to decline since the 1980s, with studies showing that scientific research has become increasingly consolidated while papers and patents have become less disruptive since 1980.[3] Moreover, though budgets in areas such as biomedical and pharmaceutical sciences have increased in the past several decades, progress has not been commensurate, with new drug approvals in the United States slowing.[4]

While these issues are concerning and demand change, ITIF first urges the administration not to make any rash decisions that could negatively impact the state of American science. One such adverse action would be the Office of Management and Budget's (OMB's) Proposed Rule to Revise the Guidance for Federal Financial Assistance. This rule would greatly expand the authority of OMB and the executive branch, giving them and agencies the authority to terminate grants, or entire classes of awards, mid-award if they are viewed as inconsistent with the goals of the agency or program that is funding them. Moreover, the language used in the proposed rule is incredibly vague, including a provision that allows the termination of a grant if it "is in the interest of the Federal agency," giving agencies near-unfettered power to revoke grants at will.[5] This proposed rule should be withdrawn from consideration to avoid the unintended consequences that would surely be realized under its implementation.

These comments will present a series of options for how the NIH and other agencies throughout the government can accurately measure scientific advancement and innovation, including measuring new drug development, total industry funding of R&D, translational research, international scientific collaboration, and bolstering the American scientific workforce over the next generation. It will also present several policy proposals for how agencies can increase scientific advancement and innovation in the United States based on these new metrics.

NEW DRUG DEVELOPMENT

Undoubtedly, the most important and relevant indicator of success in biomedical research impact is the total number of new medicines a country creates. Over the past two decades, the United States has been the leader in new drug development, as measured by the number of new chemical and biological entities produced. (See Figure 1.) However, this measure also shows how rapidly China is rising to compete with the United States. Between 2019 and 2023, China produced 71 new chemical or biological entities, a booming increase compared to the four it produced between 2014 and 2018. Over that same period, the number of new entities produced by the United States increased from just 125 to 148.[6] And China has continued to increase

---

https://itif.org/publications/2026/07/21/testimony-regarding-science-standards-and-strategic-competition/.

[3] Michael Park, Erin Leahey, and Russell J. Funk, "Papers and patents are becoming less disruptive over time," *Nature*, 613 (January 2023), 138–144, https://www.nature.com/articles/s41586-022-05543-x.

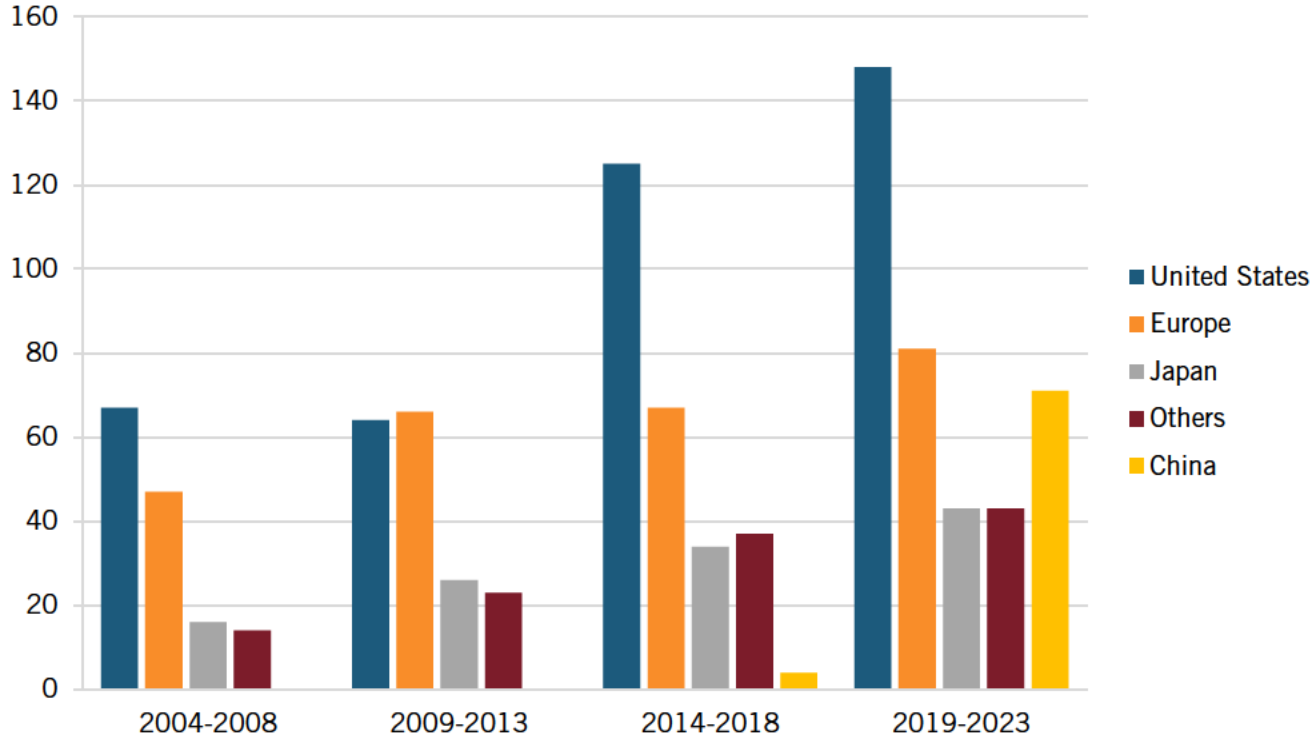
[4] The White House, "Remarks by Director Kratsios at the National Academy of Science," May 19, 2025, https://www.whitehouse.gov/briefings-statements/2025/05/remarks-by-director-kratsios-at-the-national-academy-of-sciences/.

[5] Meghan Ostertag, L. Val Giddings, and Stephen Ezell, "Comments to OMB Regarding Its Proposed Rule to Revise the Guidance for Federal Financial Assistance" (ITIF, July 13, 2026), https://itif.org/publications/2026/07/13/comments-to-omb-regarding-guidance-for-federal-financial-assistance/.

[6] European Federation of Pharmaceutical Industries and Associations, "Number of New Chemical and Biological

its development of new drugs in the years after 2023, casting aside any doubt that China is not a legitimate competitor in this industry.

Figure 1: Number of new chemical and biological entities produced, select countries



INDUSTRY FUNDING OF UNIVERSITY R&D

A key feature of the American innovation and research engine is the strong interconnection and collaboration between government, private enterprise, and universities. Universities have historically been one of the innovation labs for private enterprise, conducting the basic research that industry builds upon in applied research. Unfortunately, the private sector contributes just a small percentage of research funding to universities. For example, in 2023, the business sector contributed only $3.3 billion, or just 5 percent, of the higher education sector's total R&D spending on basic research.[7] And this rate differs dramatically from state to state. For example, North Carolina, Georgia, and Kansas led in the share of university research funded by industry, and Nebraska, Rhode Island, and Nevada were last. Among institutions, Duke, MIT, and Ohio State led, while University of Maryland, Michigan State, and Pitt were last.[8]

Robust industry funding for universities pays off economically. There are small, positive correlations between the share of a state's university research supported by industry and its strength on key innovation variables

---

Entities (2004-2023)," accessed August 12, 2026, https://www.efpia.eu/publications/data-center/the-pharma-industry-in-figures-rd/new-chemical-or-biological-entities/.

[7] Trelysa Long, "Why University Research Is Crucia to US Competitiveness" (ITIF, April 22, 2025), https://itif.org/publications/2025/04/22/why-university-research-is-crucial-to-us-competitiveness/.

[8] Robert Atkinson, "Industry Funding of University Research: Which States Lead?" (ITIF, January 2018), https://itif.org/publications/2018/01/08/industry-funding-university-research-which-states-lead/.

such as venture capital (a correlation coefficient of 0.28), scientists and engineers (0.19), high-tech startups (0.15), and high-tech jobs (0.14).[9] Industry research funding is also associated with stronger university technology output, with a correlation of 0.26 between industry investment and academic patents.[10] Moreover, interaction with industry as a STEM graduate student is associated with a significantly greater likelihood of producing intellectual property ("IP", e.g., patents, invention disclosure, etc.).[11]

Congress can and should make the following changes to incentivize and catalyze industry investment into university R&D.

- Require programs such as NSF's Engineering Research Center (ERC) to have at least some matching funds from industry as a condition of being awarded funding. In practice, for instance, this could require the University of Minnesota's ERC for Advanced Technologies for Preservation of Biological Systems to receive at least 10 percent of its funding from firms in the biotechnology or biomedical industry.[12]
- Require a number of actions in the tax code to stimulate industry support of university R&D.
    - Expand the R&D credit for companies that fund university research and eliminate language that restricts the definition of basic research to projects "not having a specific commercial objective."[13]
    - Broaden the tax credit for energy research consortia to include all research consortia.
    - Make all business funding of university and federal lab research eligible for a 40 percent flat tax credit.
    - Restore the Orphan Drug Tax Credit (ODTC) to 50 percent. To incentivize R&D of drugs for rare diseases, which are those that typically affect fewer than 200,000 people in the United States, in 1983 Congress created an orphan drug tax credit equal to 50 percent of qualified clinical trial costs and also offered a seven-year period of orphan drug exclusivity.[14]

---

[9] These variables were taken from: Robert D. Atkinson and John Wu, "The 2017 State New Economy Index" (ITIF, June 2017), https://itif.org/publications/2017/11/06/2017-state-new-economy-index. See also: John Wu and Robert D. Atkinson, "How Technology-Based Start-Ups Support U.S. Economic Growth" (ITIF, November 2017), https://www.itif.org/publications/2017/11/28/how-technology-based-start-ups-support-us-economicgrowth.

[10] National Science Board "Science & Engineering Indicators 2016" (National Science Foundation, 2016), Table 8-48, "Academic Patents Awarded per 1,000 Science, Engineering, and Health Doctorate Holders in Academia," https://www.nsf.gov/statistics/2016/nsb20161/uploads/1/13/tt0848.pdf.

[11] Jennifer Shields Schneider, "A Multivariate Study of Graduate Student Satisfaction and Other Outcomes within Cooperative Research Centers" (Raleigh: North Carolina State University, 2007), http://www.lib.ncsu.edu/resolver/1840.16/52.

[12] U.S. National Science Foundation, "Engineering Research Centers," accessed July 29, 2026, https://www.nsf.gov/eng/engineering-research-centers#current-ercs-adhttps://www.nsf.gov/eng/engineering-research-centers#current-ercs-adbb.

[13] Matthew Stepp and Robert Atkinson, "Creating a Collaborative R&D Tax Credit," (ITIF, June 2011), https://itif.org/publications/2011/06/09/creating-collaborative-rd-tax-credit/.

[14] U.S. Food and Drug Administration, "Medical products for rare diseases and conditions," (FDA, October 2024), https://www.fda.gov/industry/medical-products-rare-diseases-and-conditions.

A 2015 study by the National Organization for Rare Disorders found that at least one-third fewer new orphan drugs would have been developed to treat rare diseases over the preceding 30 years had the act not been implemented.[15] Unfortunately, the 2017 Tax Cuts and Jobs Act halved the ODTC to just 25 percent.[16] This should be restored to 50 percent.

TRANSLATIONAL RESEARCH

Since the end of WWII, U.S. science policy has followed a path of disciplinary neutrality and openness, allowing principal researchers to conduct curiosity-driven research with generous levels of federal funding and little overt incentive for turning discoveries into marketable IP.[17] This strategy was successful in leading to the development of several critical technologies, including the touchscreen and the Internet, and when coupled with a focus on international competitiveness and complemented by policies such as the Bayh-Dole Act, supported U.S. economic and industrial strength for many decades.[18]

But never in the history of the United States has the country faced a threat to global techno-economic leadership as significant as China poses, requiring a responsive shift in U.S. science policy. Several changes are of the highest importance.

1. U.S. science and research funding should not be cut under any circumstances. Unfortunately, the Trump administration's proposed FY 2027 budget calls for a 12 percent proposed cut to NIH funding, this coming after an effort to cut NIH funding by 40 percent in FY 2026.[19] If the goal of this RFI is to understand how to maximize scientific impact from federal funding of biomedical research, then policymakers should recognize that the American public enjoys a tremendous return from the federal government's investments in basic biomedical scientific research. In fact, each NIH dollar invested yields roughly $2.50 in short-term economic returns and stimulates an additional $8.30 in long-term private-sector R&D investment, underscoring the strong multiplier effect of public science funding.[20]

---

[15] National Organization for Rare Disorders, "Impact of the Orphan Drug Tax Credit on Treatments for Rare Diseases" (June 2015), https://rarediseases.org/assets/files/white-papers/2015-06-17.nord-bio-ey-odtc.pdf.

[16] Zachary Brennan, "Senate, House Agree to Cut Orphan Drug Research Credit in Half in Tax Bill," *Regulatory Affairs Professionals Society,* news release, December 18, 2017, https://www.raps.org/news-and-articles/news-articles/2017/12/senate-house-agree-to-cut-orphan-drug-research-cr.

[17] Robert D. Atkinson, "US Science Policy at a Crossroads" (ITIF, June 23, 2025), https://itif.org/publications/2025/06/23/us-science-policy-at-a-crossroads/.

[18] Association of American Universities, "Federally Funded Research Drives American Innovation," accessed August 4, 2025, https://www.aau.edu/sites/default/files/%40%20Files/Research%20and%20Scholarship/Why%20University%20Research%20Matters/Infographics/BASIC_RESEARCH_IPAD.pdf.

[19] Kritika Agarwal, "White House Once Again Proposes Massive Cuts to Scientific Research and Education," Association of American Universities, April 3, 2026, https://www.aau.edu/newsroom/leading-research-universities-report/white-house-once-again-proposes-massive-cuts.

[20] United for Medical Research, "UMR Releases Annual NIH Economic Impact Report: 2025 Update" (UMR, March 2025), https://www.unitedformedicalresearch.org/statements/umr-releases-annual-nih-economic-impact-report-2025-update/.

2. U.S. research funding can no longer be technology- or discipline-agnostic. Instead, it must be directed toward the critical and emerging technologies most essential for economic and national power. This means more focus on biology, physics, engineering, and computer science.[21]

This can be measured through the number of discoveries transferred to the commercial marketplace, similar to the statistics reported by the Association of University Technology Managers (AUTM) every year.[22] AUTM publishes data on the economic output, number of patents issued, licenses executed, and start-ups formed from academic technology transfer since 1996, giving a sense of how successful R&D investment has been over the past several decades. It also publishes the number of drugs and vaccines developed through public-private partnerships since the enactment of the Bayh-Dole Act, a critical indicator of success for the biomedical industry.[23]

The Bayh-Dole Act has had a significant impact on technology transfer and commercialization in the United States since its enactment in 1980.[24] The legislation, which allows universities, small businesses, and nonprofit institutions to take ownership of IP rights stemming from discoveries made from federally funded research. *The Economist* has called the Bayh-Dole Act "Possibly the most inspired piece of legislation to be enacted in America over the past half-century" noting "more than anything, this single policy measure helped to reverse America's precipitous slide into industrial irrelevance" as it made impact from "all the inventions and discoveries that had been made in laboratories throughout the United States with the help of taxpayers' money."[25] In the aftermath of its passage, the Bayh-Dole Act had an almost immediate impact on invention and discovery, with the number of patents awarded to universities ballooning from 390 in 1980 to 3,088 in 2009 and 6,680 in 2015.[26]

Given the profound success of the Bayh-Dole Act in catalyzing technology transfer and economic growth, it's critical that the legislation be maintained and not misused. Namely, policymakers should refrain from using the Act's march-in rights to control or ensure "reasonable prices" for drugs. March-in rights were included in the law to ensure that patent owners commercialized their inventions, with the intent to never use the provision to manipulate the pricing of a product.

Despite this, some policymakers have proposed otherwise. For instance, Senator Angus King (I-ME) has proposed legislation that would require the Department of War to issue compulsory licenses under Bayh-Dole "whenever the price of a drug, vaccine, or other medical technology is higher in the U.S. than the median price charged in the seven largest economies that have a per capita income at least half the per capita income

---

[21] Atkinson, "US Science Policy at a Crossroads."

[22] AUTM, "AUTM Surveys: Sharing Trends and Insights," https://autm.net/surveys-and-tools/surveys.

[23] "Driving the Innovation Economy," AUTM, https://autm.net/AUTM/media/SurveyReportsPDF/AUTM-2025-US-Infographic.pdf.

[24] Stephen Ezell, "The Bayh-Dole Act's Vital Importance to the U.S. Life-Sciences Innovation System" (IITF, March 2019), https://itif.org/publications/2019/03/04/bayh-dole-acts-vital-importance-us-life-sciences-innovation-system/.

[25] "Innovation's Golden Goose," The Economist, December 12, 2002, http://www.economist.com/node/1476653.

[26] Stephen Ezell, Meghan Ostertag, and Leah Kann, "The Bayh-Dole Act's Role in Stimulating University-Led Regional Economic Growth" (ITIF, June 16, 2025), https://itif.org/publications/2025/06/16/bayh-dole-acts-role-in-stimulating-university-led-regional-economic-growth/.

of the U.S."[27] The Trump administration should act to protect Bayh-Dole and its original intent by affirmatively declaring that price is not a legitimate basis for the exercise of Bayh-Dole march-in rights.[28]

U.S. universities should also collect better data on the new business start-ups coming out of U.S. universities to better understand which universities are excelling at commercializing research. This could be done through a directive from Congress directing NSF to develop a metric by which universities report such information annually. In addition, the Department of Commerce could use data available through the ES-202 form (unemployment insurance tax records), which tracks how many employees an establishment has every quarter. If the form noted the university that the founder of the organization attended, it could reveal which colleges and universities have graduates who are founding and running high-growth businesses.[29]

Federal funding agencies, such as the NIH, can take several further steps to aid and incentivize greater technology transfer and commercialization at universities.

- Federal agencies should use data on indicators of technology transfer commercialization at universities, including the number of start-ups founded or the extent of technology licensing, to reward universities that exhibit high levels of entrepreneurship. For example, agencies could give these universities bonus points on research grant proposals.
- Agencies should begin allocating a small share of university R&D funding—about 5 percent—based on how well universities commercialize research, including the number of start-ups founded from universities or the extent of technology licensing. (Here, agencies could refer to the 11 federal agencies that provide Small Business Innovation Research funding.)[30] This is similar to what other countries have done, including Finland, which allocates 25 percent of the research budgets of Finnish universities based on "quality and efficacy," including the universities ability to attract research investment from businesses. Similarly, Sweden's national government allocates 10 percent of regular research funds to universities based on performance indicators.[31]

  Similarly, Australia's engagement and impact assessment system scores universities on industry engagement, income from industry, and impact case studies. The United Kingdom's Research Excellence Framework includes an "impact" component that assesses commercial and societal impact of research, including patents, spinouts, and industry partnerships. And Chinese university funding and researcher evaluation increasingly weight patents, technology transfer, and enterprise partnerships.

---

[27] Joseph Allen, "Proposal from Senator King Won't Reduce Drug Prices, Just Innovation," IPWatchdog, July 17, 2017, https://www.ipwatchdog.com/2017/07/17/senator-king-reduce-drug-prices-innovation/id=85702/.

[28] Stephen Ezell, "Comments to the National Institutes of Health on "Maximizing NIH's Levers to Catalyze Technology Transfer" (ITIF, August 2023), https://itif.org/publications/2023/08/18/maximizing-nih-levers-to-catalyze-technology-transfer/.

[29] Stephen Ezell and Scott Andes, "Localizing the Economic Impact of Research and Development: Policy Proposals For the Trump Administration and Congress" (ITIF, December 2016), https://www2.itif.org/2016-localizing-economic-impact.pdf.

[30] U.S. Department of the Interior, "Small Business Innovation Research Programs (SBIR)," https://www.doi.gov/pmb/osdbu/small-business-innovation-research-programs-sbir.

[31] Ibid.

- The NIH should establish an entrepreneur-in-residence (EIR) program at universities that receive NIH research funding. Universities and academic medical centers that receive funding from NIH often follow the narrow and traditional path to commercializing research that revolves around patenting and licensing. The technology transfer office at the universities/medical centers then take these patents and licenses to biotechnology and pharmaceutical firms for the development of products. While this model can be an appropriate vehicle for commercialization, it often lacks strong connections between firms and research organizations. Successfully scaling a life-sciences startup requires social and capital networks, mentorship, public-private partnerships, and access to both scientific and managerial talent. The EIR program would help universities identify, support, and grow the research efforts best positioned to become high-growth companies.
- Congress should provide financial incentives to universities that develop technology accelerators/incubators to commercialize faculty and student research. For example, Stanford University created StartX and Johns Hopkins created Fast Forward, both of which assist university students and faculty in establishing entrepreneurial ventures seeking to move university-developed discoveries and inventions into the commercial sector. These types of programs are increasingly proliferating throughout the U.S. university system, but additional support from the federal government would support the development through a wider set of universities and colleges, particularly those without large endowments or wealthy alumni to self-fund such programs.

Finally, the federal government needs to prioritize commercialization activities at federal labs and research institutes. America's federal laboratories are insufficiently incentivized to invest time, energy, and resources in facilitating technology transfer. For instance, Department of Energy (DOE) lab planning guidance includes a dedicated "Technology Transitions, Commercialization, and Partnership Strategy" section asking labs to describe their Cooperative Research and Development Agreement (CRADA), Agreements for Commercializing Technology (ACT), and Strategic Partnerships Projects activities and how these tie into lab strategy. But that's a planning-narrative requirement, not a scored performance criterion on par with the eight goals**.** It needs to be an enforced performance requirement tied to funding. Further, the federal government should require that applications for federally funded research at labs, research institutes, and universities include a one-page summary of the potential commercial applications of the research being funded.

EXPAND AND MODIFY SBIR AND STTR

Billed as "America's Seed Fund," the Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs provide over $2 billion per year to qualified small businesses to fund R&D activities through multiple federal agencies, acting as a key conduit to brining scientific innovations to market. While SBIR accounts for only 3.4 percent of federal extramural research funding, the program punches well above its weight, with as much as 22 percent of America's top innovations (as reflected by studies of previous winners of R&D Magazine's R&D 100 innovation awards) coming from companies that received SBIR grants at some point in their history.[32]

[32] Fred Block and Matthew Keller, "Where Do Innovations Come From? Transformations in the U.S. National Innovation System, 1970-2006" (Washington: Information Technology and Innovation Foundation, 2008), http://www.itif.org/files/Where_do_innovations_come_from.pdf.

Yet SBIR's impact could be even greater, particularly if certain facets of the program were slightly more strongly geared toward commercialization.

- SBIR awardees should be permitted to expend up to 5 percent of their award funds for commercialization-oriented activities. Awardees are currently limited in using grant funds to support critical commercialization activities such as building product or service prototypes, acquiring commercial customers, attracting private capital, or accelerating market entry. These activities span the gamut of important commercial activities, including IP development and prosecution, marketing and market development, and the recruitment of key team members associated with customer acquisition, thereby preventing awardees from taking important steps to commercialize innovations.[33]
- All participating agencies in the SBIR and STTR programs consider commercialization potential and plans in their grant funding decisions. However, agencies differ in the weight or emphasis they place on commercialization. In agencies where the intended customers are external, a greater portion of the merit review evaluation criteria and scoring should include commercialization factors, such as the company's understanding of market opportunity, product development timelines, and needed resources. Further, to evaluate these important criteria, the composition of SBIR/STIR review panels at these agencies should include industry experts, investors with relevant industry or technology expertise, and/or representatives from commercialization intermediary organizations.

INTERNATIONAL SCIENTIFIC COLLABORATION

The China challenge is so large that the United States cannot hope to succeed without partnering with allies. Congress and federal agencies need to expand joint research initiatives with allies in critical technology areas, such as biomedical research, and develop a shared network of advanced-industry centers in which firms from both countries can participate in each other's programs. International cooperation among national technology agencies, such as the NIH, presents an opportunity for technological innovation and information sharing between the United States and allied countries at potentially lower cost. Such programs have been undertaken by the National Science Foundation's (NSF) Technology Innovation and Partnership (TIP) Directorate, which conducts both basic and translational, market-ready research.[34]

For example, the Verticals-enabling Intelligent Network Systems (VINES) is a funding opportunity that supports research to enhance wireless communication networks, specifically to advance next-generation network systems. A global network of agencies and industry partners supports VINES, including NSF, Department of Homeland Security, Qualcomm, Ericsson, Intel, and international partners from Finland, India, Japan, and Sweden.[35] Congress should encourage federal agencies to seek out and invest in initiative such as this. NSF TIP has stated that it intends to invest $50 million into VINES when it is allocated its full funding from Congress.[36]

---

[33] Ezell and Andes, ""Localizing the Economic Impact of Research and Development."

[34] Robert D. Atkinson and Meghan Ostertag, "Congress Should Fully Fund NSF's TIP Directorate to Make America More Competitive Versus China" (ITIF, June 30, 2025), https://itif.org/publications/2025/06/30/congress-should-fully-fund-nsf-tip-directorate/.

[35] "Verticals-enabling Intelligent Network Systems (VINES)," U.S. National Science Foundation, June 2025, https://www.nsf.gov/funding/opportunities/vines-verticals-enabling-intelligent-network-systems.

[36] Robert D. Atkinson and Meghan Ostertag, "Congress Should Fully Fund NSF's TIP Directorate to Make America

The federal government should not only measure the amount of international collaboration between international partners and federal agencies, such as the NIH, and their outcomes, but also actively encourage these partnerships. One way it can facilitate these is through reducing the bureaucratic barriers to collaboration. Navigating diplomatic and congressional bureaucracy can slow down or even fully prevent these partnerships from materializing. The Department of State should identify countries that are automatically approved for scientific engagement, thus streamlining the path to partnership. The nations identified would be countries with which the United States is closely allied and those that possess the human and infrastructural resources necessary to make the partnership mutually beneficial.

PRODUCTION OF STEM TALENT

NIH plays a central role not only in funding research but also in sustaining the nation's biomedical workforce. It supports graduate education, postdoctoral training, and career pathways for more than 300,000 researchers through nearly 50,000 competitive grants across 2,500 universities and research institutions.[37] In FY2024, 34,000 R01-equivalent investigators applied for NIH funding, with about 9,655 receiving awards.[38]

This talent pipeline represents a strategic national asset. A steady flow of skilled biomedical researchers is essential to respond to pandemics, biothreats, and health security challenges. Without sustained federal investment, this talent will be lost to other countries or diverted to different fields, weakening U.S. capacity for innovation and response. Initiatives such as France's Choose France for Science campaign and the European Union's Horizon Europe Strategic Plan 2021–2024, which have been explicitly designed to attract top international talent, especially from the United States, pose a risk to the United States' ability to attract and retain talent.[39] In this competitive environment, sustained NIH support is critical to maintaining the United States as an attractive global hub for world-class biomedical science and to ensuring that the nation can meet both public health and national security needs.

A COMPREHENSIVE COMPETITIVENESS ASSESSMENT FOR AMERICA'S BIOPHARMACEUTICAL INDUSTRY

These comments contain several policy recommendations for how the NIH and other federal agencies can better record and incentivize innovation and scientific success. But ultimately, any effective policy needs to be backed by ongoing institutional, analytical, and intergovernmental capacity, which appears to be lacking in the United States. The NIH should establish a committee of industry experts and government officials to assess the competitiveness of the U.S. biopharmaceutical industry and issue a biannual report on its state of

---

More Competitive Versus China."

[37] National Institutes of Health, "Budget, Research for the People," June 13, 2025, https://www.nih.gov/about-nih/organization/budget.

[38] National Institutes of Health "RePORT NIH Data Book,", January 2025, https://report.nih.gov/nihdatabook/category/14.

[39] "Choose France for Science: launch of the dedicated platform for applications to host international researchers," Republique Francaise, April 17, 2025, https://anr.fr/en/latest-news/read/news/choose-france-for-science-launch-of-the-dedicated-platform-for-applications-to-host-international-r/; Horizon Europe Strategic Plan 2021 – 2024, Directorate General for Research and Innovation, 2021, https://www.eeas.europa.eu/sites/default/files/horizon_europe_strategic_plan_2021-2024.pdf.

competitiveness. The report should include all of the metrics we've outlined in these comments as well as others, including but not limited to:

- Domestic and global industry share of value-added output;
- Share of active pharmaceutical ingredients and key starting materials produced in the United States;
- Trade balances in the sector;
- Foreign direct investment in the sector;
- Productivity per worker and total factor productivity;
- Annual number of clinical trials for each type of drug;
- Global and domestic sales;
- Location quotients.

CONCLUSION

As global competition in biomedical innovation intensifies, particularly from China's state-directed push for dominance, the metrics the United States uses to gauge scientific progress can no longer rely on outdated measures. NIH and other federal agencies must adopt a more comprehensive framework for measuring scientific impact, one that captures industry outputs such as new drug development, industry investment in university research, the translation of discoveries into commercial and clinical outcomes, the depth of international scientific partnerships, and the strength of the biomedical talent pipeline. These diagnostics will enable policymakers to better understand where U.S. policy is currently insufficient and where targeted actions are warranted.

Pairing smarter metrics with the policy recommendations outlined above, the United States can ensure that its biomedical research enterprise, and greater scientific enterprise as a whole, remains positioned to maintain the economic competitiveness of the United States for years to come.

Thank you for your consideration.

Meghan Ostertag
Policy Analyst

Stephen Ezell
Executive Vice President and Director, ITIF Center for Life Sciences Innovation

235. American Physiological Society

I am responding: On behalf of an organization

Name: Alissa Hatfield

Name of Organization: American Physiological Society

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Science Policy Manager

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Rigorous and reproducible research begins at the experimental design phase of each grant application. Researchers carefully plan experiments so that they have adequate statistical power to accurately answer the scientific questions at hand. When research involves either human or animal research subjects, it is essential that experiments are designed to include a sufficient number of participants to ensure that the data generated is statistically significant. When requested funds are reduced, investigators are forced to amend their experimental design from the original rigorous research design. This may involve reducing or eliminating the number of human or animal subjects that would otherwise be needed to test the influence of sex on outcome variables with adequate power. To support rigorous and reproducible research, and to ensure that the investigators have the resources to meet the NIH mandated focus on sex as a biological variable, grants must be fully funded at the requested level.

2. Data, Software, and Model Sharing: Rapidly changing requirements for research security have left researchers with questions about how to meet data sharing requirements while also complying with research security policies. As research security policies change, NIH should review and update data sharing expectations. Ensuring that data sharing expectations are compatible with security concerns is an essential part of promoting data sharing within and outside of the NIH community.

NIH should recognize or reward demonstrated adoption or reuse of shared resources by the broader scientific community. Activities related to developing, maintaining, and using shared data resources involve a significant investment of time and resources that may not be adequately reflected in traditional measures of research output. One way that NIH might enable this is by allowing grant funds to be used for costs related to shipping or sharing samples or data. Additionally, NIH could create a data depository, similar to PubMed. If NIH hosts the data, then the barriers are lowered on the investigators. Investigators could then more easily ensure that data are deposited in a timely fashion.

3. Training and Mentorship: To maximize training and mentorship opportunities, NIH must understand what their trainees are seeking. While it is often assumed that postdoctoral training is intended to prepare scientists for tenure-track faculty positions, only 1 in 4 postdocs will be in a tenure-track academic position after 5–6 years post-graduation (Denton et al., 2022). NIH should assess the quality of mentorship, not merely trainee outcomes. For example: trainee progression and independence,

individualized development plans, successful transition into scientifically relevant careers, opportunities for authorship and intellectual leadership, mentorship of trainees from different career stages, sustained mentoring relationships, etc. This is particularly important because simply counting how many trainees become faculty could create the wrong incentive. Training and mentorship opportunities for postdocs vary widely at the discretion of the mentor. Essential elements of this training include grant writing, publishing, experimental design, scientific communication, etc. If all elements of training are adequately addressed, NIH supported graduate students and postdocs should be well positioned for a variety of careers both inside and outside of academia. NIH should examine current trainee offerings and ensure that they provide and require experience in skills that will translate well into both academic and industry-related careers. NIH should periodically review what are considered "successful" outcomes on training grants to fully encompass and reward alternative/non-tenure track careers.

NIH should consider incorporating public-private partnerships into their mentoring activities to ensure that trainees who pursue industry careers have adequate access to mentorship in those areas of interest.

NIH should recognize the significance of the commitment to mentoring by allowing effort for this mentoring to be an allowable cost.

Denton, M., Borrego, M., & Knight, D. B. (2022). U.S. postdoctoral careers in life sciences, physical sciences and engineering: Government, industry, and academia. PLoS ONE, 17(2 February). https://doi.org/10.1371/journal.pone.0263185

4. Collaboration: Science has always been a collaborative effort. Diversity, innovation, and creativity have long been essential elements in team-based science.

The recent proposal to cap Research Project Grants (RPGs) per Principal Investigator (PI), if finalized in its current form, will likely have a negative impact on collaboration by disincentivizing multi-PI grants.

In May of 2026, NIH reaffirmed definitions of foreign components. One of these components is "collaborations with investigators at a foreign site that result in co-authorship." This policy disincentivizes collaborations that might benefit the health of the American people etc. by increasing administrative burden on relationships with foreign entities, some of whom do not receive NIH funding. Existing policies and export control regulations are sufficient to protect against intellectual property theft and to ensure that sensitive data are not misused.

5. Entrepreneurship and Translation: Over the past 18 months NIH has emphasized its goal of prioritizing "human-based research technologies." The newly formed Office of Research Innovation, Validation, and Application (ORIVA) embraces this as an essential goal. We would caution NIH that "human-based" research does not exclude animal models. Animal models have been fundamental to modeling human physiology and disease and have consistently laid the scientific foundation leading to therapeutic advances. NIH fosters the maturation of discoveries from bench to bedside, and these opportunities for innovation would stall by excluding any type of model, including animal models. As new technologies are developed that reduce reliance on traditional animal models, we encourage NIH not to discount the role that traditional animal models of human disease play in the development of scientific concepts and the testing of hypotheses and therapies. Instead, NIH should look for fit-for-

purpose experimental systems. NIH should evaluate scientific impact based on whether investigators employ the most appropriate and rigorously validated model system for the biological question, including human-based models, computational approaches, organoids, animal models, or integrated combinations of these approaches.

6. Foundational Scientific Exploration: Transformative scientific advancements often arise from curiosity-driven, investigator-initiated research. Many of the most important breakthroughs in physiology and medicine have emerged from efforts that were not originally directed toward a specific translational goal. We encourage NIH not to lose sight of the foundational, basic exploratory science. These discoveries, which do not always have an immediate translational goal, lay the groundwork for the future of scientific advancement. We encourage NIH to avoid overemphasizing short-term or immediately measurable outcomes at the cost of high risk endeavors, and to value contributions that enable future discoveries, even when direct applications are not yet apparent. It can take decades before basic science observations can be linked to others to form a transformational discovery. This interconnected network of basic observations forms the basis of science and should not be discounted. A balanced research portfolio requires support for both basic and translational science. These initiatives would also benefit from periodic training and retraining of the NIH staff (e.g. SROs) as well as the reviewers on study sections to make a clear distinction between high-risk high-reward applications and the ones that also address key knowledge gaps, but which may not fit the high-risk high-reward criterion.

7. Public Impact: APS encourages NIH to explore existing resources regarding the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit. Scientific societies and advocacy organizations have long used these resources, including United for Medical Research's report on "NIH's Role in Sustaining the U.S. Economy"

NIH should consider recognizing contributions related to public engagement and science communication, dissemination of research findings to non-scientific audiences, educational outreach activities, and contributions that enhance public trust in science and scientific institutions. These activities should complement, not replace, measures of scientific excellence and rigor. NIH may consider how these activities, and others related to civic science, might be considered at the funding level. NIH could also acknowledge that some grant funds can or should be used in civic outreach.

8. Other Areas: APS encourages NIH to pilot and evaluate any proposed metrics before widespread implementation. New measures should each be assessed with the involvement of the scientific community for feasibility, reliability, and potential unintended consequences. NIH should specifically aim to minimize burden on investigators, maintain flexibility across disciplines and career stages, and continue to incorporate expert peer review alongside other measures.

236. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: '-Create dedicated funding streams or high-status journal tracks specifically for publishing well-powered replication studies and negative results.
-Treat shared datasets, software code, and open-source hardware as first-class research outputs with distinct Digital Object Identifiers (DOIs) that must be cited.
-Federal open-access mandates have created a cost barrier that threatens scientific equity. Top-tier journals now charge prohibitive fees that grants rarely cover and individual labs cannot absorb. Many researchers are effectively excluded from the highest-impact venues, undermining rigorous peer review and visibility. Sustainable solutions to open-access costs are essential to ensure equitable access to premier publication venues.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: '-Structure major grants so that multi-disciplinary co-PIs receive equal institutional and funding credit, removing the "sole leader" prestige bias.
-Track and reward scientists who build or maintain vital research infrastructure, biobanks, or community data repositories used by thousands of other researchers.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: '-Quantify contribution to public literacy by measuring community engagement events, activities, and workshops and media contributions.

8. Other Areas: N/A

237. Zheng Wang

I am responding: On behalf of myself

Name: Zheng Wang

Name of Organization: W2 BioSolutions

Type of Organization: Not Applicable

Role: Other

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Gene expression level is one of the most important biomedical data as it links to whether a gene's function is robust in a living body or not. Messenger RNA (mRNA) profiling has long been used as biomarkers in diagnostic, prognostic, and predictive assays. Four major platforms have been developed for RNA profiling. Listed in time order, they are RT-qPCR, microarray, next-generation sequencing (NGS) and NanoString nCounter. Digital PCR (dPCR) has been developed based on RT-PCR mechanism, but separately amplify each gene copy and detect the result. All those four technologies are matured and have received clearances from the Food and Drug Administration (FDA) for some of their products although none of them are perfect. In order to further improve the rigor and reproducibility of RNA expression data, we must understand the differences between two types of quantification: relative quantification and absolute quantification.

Relative quantification measures relative change in RNA expression levels. These data determine the changes in steady-state RNA levels of a gene across multiple samples and express it relative to the levels of another RNA (1) or a group of RNAs, such as housekeeping genes (2). However, there is a problem with this assumption, because it is nearly impossible to find a reference gene expression that does not change when the biological condition changes. Without a standard normalization reference(s), the data of relative quantification are literally unable to be integrated together and have difficulties to freely perform comparisons among different samples/experiments. This is because the normalization reference itself of relative quantification is NOT a fixed value, but also variables. Unfortunately, relative quantification is the "default" quantification method in gene expression measurement. Despite NGS's data can be output as absolute number of reads, without proper normalization, it is no different from microarray's data of fold change in the term of relative quantification. Thus, data generated in different experiments, no matter by microarray or NGS, cannot be directly integrated together or compared with each other. Although Nanostring nCounter's counts may equal RNA copy numbers during ideal experimental conditions, practically, this method also requires normalization to adjust the data integrity (3,4). The spiked-in RNA is suggested to be a normalization method for NanoString nCounter too (5). NanoString nCounter does not immune from random errors and systematic errors, and the accuracy and reliability of its gene expression results is also dependent on the proper normalization of the data against reference genes (3-7). Consequently, the counts provided by current Nanostring nCounter are data of relative quantification as well since those counts are not normalized to standard known copy numbers.

Absolute quantification measures the exact number of molecules, for mRNA is copy number because it

fits the First Principles for gene expression quantification. Such measurement ensures precise and reproducible results because “copy number” to RNA expression is equivalent to “gram” to weight and “liter” to liquid volume. All measures are universal definitions and never change. Scientific researchers have been trying to establish better normalization methods for gene expression profiling. However, most of the researchers have not pursued innovation of technology to output absolute quantification data, but have depended on pure mathematical methods to solve the problem of the incomparable data. The mathematical “deconvolution methods” ended up in the impossibility of obtaining absolute measurement with universal value (8). Therefore, it is the time to pursue technology innovation to output absolute quantification data to provide gene expression data with rigor and reproducibility.

For technology innovation, please see the “Entrepreneurship and Translation” section.

References:

1. Bustin SA. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. J Mol Endocrinol. 2002;29(1):23-39. doi: 10.1677/jme.0.0290023.
2. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29(10):569-74. doi: 10.1016/j.tig.2013.05.010.
3. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol. 2008;26(3):317-25. doi: 10.1038/nbt1385.
4. Brumbaugh CD, Kim HJ, Giovacchini M, Pourmand N. NanoStriDE: normalization and differential expression analysis of NanoString nCounter data. BMC Bioinformatics. 2011; 12:479. doi: 10.1186/1471-2105-12-479.
5. Prokopec SD, Watson JD, Waggott DM, et al. Systematic evaluation of medium-throughput mRNA abundance platforms. RNA. 2013;19(1):51-62. doi: 10.1261/rna.034710.112.
6. Molania R, Gagnon-Bartsch JA, Dobrovic A, Speed TP. A new normalization for Nanostring nCounter gene expression data. Nucleic Acids Res. 2019;47(12):6073-83. doi: 10.1093/nar/gkz433.
7. Eastel JM, Lam KW, Lee NL, Lok WY, Tsang AHF, Pei XM, et al. Application of NanoString technologies in companion diagnostic development. Expert Rev Mol Diagn. 2019;19(7):591-8. doi: 10.1080/14737159.2019.1623672.
8. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. Cell Rep. 2019;26(6):1627-1640.e7 doi: 10.1016/j.celrep.2019.01.041.

2. Data, Software, and Model Sharing: Besides a software is needed to be developed to perform the calculation of measured RNA copy numbers, a centralized database and comprehensive software may be developed to fully utilize the power that standardized absolute RNA copy number data can provide. Both software and mathematics are very advanced and have no problem crunching big data. One of major obstacles that drag epigenetics research is gene expression profiling data are relative quantification data. Feeding big data with relative quantification data, the result is sort of “garbage in, garbage out.”

Absolute quantification data, mRNA copy number / nanogram RNA, are clean/clear data with a standard unit. Thus, absolutely quantified mRNA copy number of different experiments can be freely integrated together no matter from different times or different locations. Therefore, data sharing, reuse and accumulation become realistic. Which in return will dramatically increase data power and accelerate

biomedical research discovery only if there are enough researchers participating in contributing their quality controlled absolute quantification data to establish a central database.

3. Training and Mentorship: Training is needed. Mentorship is great. I'd like to applaud, particularly, NIH also practice Request for Information. If it is okay, may I boldly further suggest that a two-way communication is feasible in uncharted territory? Either training/mentorship or information submission often end up in a one-way direction. I believe that constructive dialogue and creative discussion may elevate one-way information flow to a higher and more fruitful level. If such a two-way communication mechanism is available, please forgive my ignorance.

4. Collaboration: Traditionally, the model of research and discovery depends on experimental design that may be conducted in a laboratory or collaborated among many groups. Research results are concluded within the designed experiments. This model may be fundamentally shaped when switching data from relative quantification to absolute quantification. Absolute quantification data stored with character descriptions will provide researchers the opportunity to "collaborate" with all data contributors because all accumulated data can be freely reused to answer new questions. For example, there are 3000 sets of mouse gene expression profiling data stored historically from 3000 different labs of different experiment purposes. By chance, they used the same mouse strain, but different ages. Let's assume a scientist who is studying the relevance of gene X and development. The scientist may easily pull the related data from the database to perform further analysis. In that way, organized absolute quantification data of gene expression profiling may provide unlimited scientific collaboration globally.

5. Entrepreneurship and Translation: If we can upgrade the matured gene expression profiling platforms from current relative quantification to absolute quantification, it would be an ideal innovation. The way to fix the problem is to establish a normalization method that includes defined calibration curve(s) or other standards - the key attribute of absolute quantification (1, 2). Standard curve is the traditional and most used method to perform absolute quantification. In this approach, a series of known concentrations of the target molecule is prepared to generate a standard curve. The unknown sample is then compared to this curve to determine its absolute quantity. This method is used in quantitative PCR (qPCR), where the cycle threshold (CT) values of the unknown are matched against the standard curve to calculate the exact copy number. However, qPCR is unable to meet the needs of thousands gene expression profiling. Microarray and NGS are the two most powerful high-throughput platforms, which one should be our chosen for the innovation?

The comparative study on different platforms revealed that the reproducibility of microarrays can be higher than NGS (3). Moreover, it is my understanding that it is not feasible to include RNAs at known abundances for use as quantification standards during manufacturing process based on the platform mechanisms of NGS and NanoString nCounter. It is, however, experimentally proved that standard curves can be generated on microarray platform (4,5). The next question is how we can add a reliable "scale" on microarry platform to measure gene copy number? To answer this question, let's first have a brief review of the basic mechanism of microarray technology. Microarray technology is based on nucleic acid hybridization. Sample RNA, called target, binds to probe that is immobilized on microarray chip. Labeled target RNA binds to probe that is immobilized on microarray chip. Labeled target's signal intensity bound on probe is scanned to estimate the RNA amount as it is illustrated in Figure 1 (See uploaded file).

While microarray is a dependable technology, it cannot output standard absolute quantification data

because of the labeling and hybridization variations of each experiment and the sequence-dependent labeling and hybridization efficiencies of RNAs (4, 6-10). To upgrade the microarray platform from relative to absolute quantification, two challenges must be overcome --

The number one challenge is to identify reference(s) to normalize the labeling and the hybridization variations of each experiment. The spike-in method was developed to satisfy the need of canceling biases related to labeling and hybridization of each experiment (4, 10). The spike-in is the most notable and widely used method. It is suggested to be a fundamental need for virtually all genome-wide analyses, including sequencing (11-13). One of the milestones of the spike-in is the National Institute of Standards and Technology (NIST) hosted project of the industry-initiated External RNA Controls Consortium (ERCC) (14) . ERCC partners from industry, government, and academia, were attempting to develop RNA spike-in controls and establish analytical methods for bringing reproducible gene expression measurements into routine, high-quality practice. Developed for microarray platform, ERCC was used in RNA-seq to estimate RNA copy numbers as well (15). Should we fully utilize the valuable ERCC metrology products, Standard Reference Material (SRM) 2374 (16), we can be on a fast track to achieve the goal of absolute quantification on gene expression profiling with an innovative approach.

The second challenge is to resolve sequence-dependent labeling and subsequent hybridization efficiencies of RNAs. The traditional spike-in has two shortcomings. The first shortcoming is incredibly difficult to spike-in precise amounts across all the samples (17,18). Thus, very few people have used spike-in to perform absolute quantification. Another shortcoming is that one spike-in control sequence only generates one intensity data point on an array. A standard curve can only be made on a combination of control sequences as illustrated in Figure 1. The red circle on the top indicates that a set of RNAs are needed to be spiked-in to generate different signal intensities on the chip (the red circle at the bottom). Representing individual RNA character using a standard curve can be challenging.

The innovation and mechanism of absolute quantification microarray technology:
Figure 2 (See uploaded file) shows an innovative spike-in, named S-pike-in. The core of this invention (19,20) is the fundamental principle of 1:1 hybridization ratio between probe and target. This hybridization principle is the theory that allows us to advance the current spike-in method to be able to manufacture built-in standards on a chip via controlling immobile probe(s) instead of spike-in free target(s). Fixed probe concentrations/copy numbers on a chip allow us to use one control sequence to generate one standard curve. In Figure 2, the red circle on the top indicates that one S-spike-in sequence can generate different signal intensities on the chip (the red circle at the bottom), which can generate a standard curve. So, the amount of S-spike-in does not affect the accuracy of standard curve as spike-in does. This method erases the variations caused by individually added different spike-in concentrations. This method also allows for many standard curves to be built on one array. It also means that different standard curves may be established to represent different RNA characters. Thus, the S-spike-in provides potential solutions to overcome the shortcomings of traditional spike-in, while maintaining the function of canceling biases related to labeling and hybridization intact.

Once the inherited microarray platform problems are solved and external standard copy number curves are able to build on microarray chips, the following computational work, including determining original copy number that is before amplification, are much easier. The signal intensities obtained from the

arrays will be subjected to regular background subtraction and quality inspection. Linear regression analysis on each S-spike-in reference probe can be used to generate algorithms of linear correlations of the signal intensities and mRNA copy numbers for standard curves. Finally, the linear regression formula generated by the standard curve will be used to translate signal intensities of target mRNAs into copy numbers.

References:

1. Bustin SA. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. J Mol Endocrinol. 2002;29(1):23-39. doi: 10.1677/jme.0.0290023.
2. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29(10):569-74. doi: 10.1016/j.tig.2013.05.010.
3. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nat Methods. 2014;11(8):809-15. doi: 10.1038/nmeth.3014.
4. Carter MG, Sharov AA, VanBuren V, Dudekula DB, Carmack CE, Nelson C, et al. Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. Genome Biol. 2005;6(7): R61. doi: 10.1186/gb-2005-6-7-r61.
5. Yauk CL, Williams A, Boucher S, Berndt LM, Zhou G, Zheng JL, et al. Novel design and controls for focused DNA microarrays: applications in quality assurance/control and normalization for the Health Canada ToxArray. BMC Genomics. 2006;7:266. doi: 10.1186/1471-2164-7-266.
6. Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, et al. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. RNA. 2010;16(5):991-1006. doi: 10.1261/rna.1947110. (19)
7. Baker M. MicroRNA profiling: separating signal from noise. Nat Methods. 2010;7(9):687-92. doi: 10.1038/nmeth0910-687. (20)
8. Leshkowitz D, Horn-Saban S, Parmet Y, Feldmesser E. Differences in microRNA detection levels are technology and sequence dependent. RNA. 2013;19(4):527-38. doi: 10.1261/rna.036475.112. (22)
9. Novoradovskaya N, Whitfield ML, Basehore LS, Novoradovsky A, Pesich R, Usary J, et al. Universal Reference RNA as a standard for microarray experiments. BMC Genomics. 2004;5(1):20. doi: 10.1186/1471-2164-5-20. (35)
10. Ryden P, Andersson H, Landfors M, Naslund L, Hartmanova B, Noppa L, et al. Evaluation of microarray data normalization procedures using spike-in experiments. BMC Bioinformatics. 2006; 7:300. doi: 10.1186/1471-2105-7-300. (36)
11. Chen K, Hu Z, Xia Z, Zhao D, Li W, Tyler JK. The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. Mol Cell Biol. 2015;36(5):662-7. doi: 10.1128/MCB.00970-14. (37)
12. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011;21(9):1543-51. doi: 10.1101/gr.121095.111. (38)
13. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. Nucleic Acids Res. 2017;45(4):e23. doi: 10.1093/nar/gkw984. (39)
14. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, et al. The External RNA Controls Consortium: a progress report. Nat Methods. 2005;2(10):731-4. doi: 10.1038/nmeth1005-731.

15. Owens ND, Blitz IL, Lane MA, et al. "Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development". Cell Reports. 2016;14 (3): 632–647. doi: 10.1016/j.celrep.2015.12.050.
16. https://tsapps.nist.gov/srmext/certificates/2374.pdf
17. Zhou X, Zhu Q, Eicken C, Sheng N, Zhang X, Yang L, et al. MicroRNA profiling using microParaflo microfluidic array technology. Methods Mol Biol. 2012; 822:153-82. doi: 10.1007/978-1-61779-427-8_11.
18. Benes V, Muckenthaler M. Standardization of protocols in cDNA microarray analysis. Trends Biochem Sci. 2003;28(5):244-9. doi: 10.1016/S0968-0004(03)00068-9
19. Wang, Zheng. Methods and systems for using built-in standard curve to measure molecular numbers of biological components. US Patent No. 10190156, filed April 14, 2014, and issued January 29, 2019.
20. Wang, Zheng. 内置性标准曲线检测物质分子数之方法系统(Methods and systems for using built-in standard curve to measure molecular numbers of biological components). China Patent #: ZL 201510118285.3, filed March 18, 2015, and issued May 18, 2018.

6. Foundational Scientific Exploration: It is quite encouraging that NIH is committed to foundational scientific exploration and willing to take risk of failure but the potential for paradigm-shifting breakthroughs. In my humble opinion, a paradigm-shifting is needed to speed up epigenetics research/discovery/clinical application. The current relative quantification of gene expression modules satisfies the conclusion that answers a specific question between control and treatment. Relative quantification data generated within one experiment are a closed set of data. Such data are not open to freely interact with each other, therefore, Big Data to Knowledge is difficult to reach. While absolute quantification data determined in RNA copy number/ng RNA allow data freely to interact with each other to speed up Big Data to Knowledge, another beauty of the upgraded absolute quantification microarray platform is that it can keep the function of the relative quantification. This is because the innovation described in the “Entrepreneurship and Translation” section is through adding a “scale” – standard curve on the original chip and upgraded computational software to achieve. Thus, the original relative quantification function can be intact on the absolute quantification platform. Which will provide a smooth process of paradigm-shifting. Researchers may continue their research in the way they are familiar with, simultaneously, the data can also be extracted as RNA copy number/ng RNA. Should the NIH lead the new direction of absolute quantification, there is no abrupt interruption to the current gene expression field, but open unlimited potential for the future.

Another advantage of absolute quantification is whenever a better method emerges to replace an old one, any data generated by absolute quantification including the microarray method can be easily transferred and integrated into another absolute quantification data pool generated by another platform(s). This is because each absolute quantification data is generated on a fixed value. The difference between two fixed values can be easily determined by a ratio.

7. Public Impact: Since gene expression of absolute quantification of copy number in a defined amount of biological material is a quantification with universal value, the impact of its success cannot be overemphasized. First of all, there is no doubt that it can dramatically speed up research discovery. Secondly, research results can fit clinical application seamlessly because the gene expression defined as “X gene copy number/ng RNA” is equivalent to blood pressure defined as “mmHg.” Thirdly, it is economical and cost-saving as the absolute quantification does not need the concurrent control for each

test as the relative quantification does. Fourthly, an absolutely quantified gene expression database will be a powerful tool when facing unexpected health emergency situations, pandemic COVID-19 is used as an example to explain how significant such a database would be below.

The world-wide response to COVID-19 was quite passive with struggle from diagnosis to treatment. Let's reshape the thinking model from targeting diagnosis and treatment separately to focus on one goal: restore health conditions. Health depends on each gene function normally. A pathogenic invasion damages health through altering gene expression profile. This is not only a theory, but published data showed that COVID-19 patients' transcriptomic pattern demonstrated responses to the viral infection (1). Another study, based on the swine experiment, showed that the response of transcriptome profiles of peripheral blood mononuclear cells could be detected 24h after porcine reproductive and respiratory syndrome (PRRS) vaccination. PRRS is caused by a positive-sense single-stranded RNA virus as well (2). These findings demonstrate that measuring transcriptome can be performed as a quick check on whether a person is infected and whether a person's body has generated a good immune response. Now, the question is how do we translate transcriptome data into knowledge in fighting COVID-19? It would be of tremendous value if we were to develop method(s) that are ready to test any unknown pathogenic invasion and provide essential information for treatment. Is such a goal achievable? The answer "Yes" lies in an established standardized gene expression profile database. If there was such a database available, gene expression profiles of COVID-19 patients could be compared to database stored data. Which would provide the guidance and information needed to design/select/evaluate treatments. The principle is similar to the diagnosis and treatment evaluation of anemia. The diagnosis and treatment evaluation of anemia depend on the hemoglobin range of grams per deciliter (g/dL). Changes in gene expression profiles can serve as a reference for the diagnosis and treatment assessment of COVID-19.

Needless to say, whoever owns such an absolute quantification database would have a dynamic molecular base for data-driven decision-making for health-related issues and be in an advanced position on functional genomics and precision medicine.

References:

1. Xiong Y, Liu Y, Cao L, Wang D, Guo M, Jiang A, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerg Microbes Infect. 2020;9(1):761-770. doi: 10.1080/22221751.2020.1747363

2. Islam MA, Neuhoff C, Aqter Rony S, Große-Brinkhaus C, et al. PBMCs transcriptome profiles identified breed-specific transcriptome signatures for PRRSV vaccination in German Landrace and Pietrain pigs. PLoS One. 2019;14(9):e0222513. doi: 10.1371/journal.pone.0222513.

8. Other Areas: I'm aware that I'm submitting a huge project and unpopular idea. Once an entrepreneur colleague said to me "you are against the mainstream." Besides relative quantification is the "default" quantification method in gene expression measurement, sequencing technology will replace microarray has been a common view for more than a decade. However, microarray is still being widely used today (1-7). Regarding those two platforms' strength and weakness, three groups published solid comparison studies (7-9). The FDA group's recent conclusion clearly stated microarray is a viable method for transcriptomic applications (7). Moreover, researchers even innovatively combined microarray and sequencing technologies together for scalable spatial transcriptomics (6).

I am not interested in being in or against any stream, but having a strong desire to solve problems. I submitted the similar material in last year's Replication and Reproducibility tract 1 and tract 2 submission without success. I could not obtain the reviewers' review. Therefore, I do not know what the reviewers' opinions are. Nevertheless, I believe that careful argumentation and thorough consideration are the first steps toward the success of a project, especially aimed at paradigm-shifting. If absolute quantification of gene expression profiling is an interest of NIH, I'd like to suggest that NIH summit some top scientists in related field, let's sit down to freely discuss, debate, and brainstorm together –

1) Is absolute quantification of RNA expression profiling worth pursuing? If "Yes" –

2) Is it acceptable to add the function on an old technology platform? If "Yes" –

3) Which high-throughput platform we should choose, NGS or microarray? It would be fair that let scientists of both technology platforms to present their expert opinions on: a) the possibility of mechanism that allow absolute quantification on the two platforms; b) the comparison of strengths and weaknesses of the two platforms with considering that how much money has been poured into sequencing technology and the possible improving room of microarray; c) the cost effectiveness of the two platforms including data size and storage space. I only have very limited knowledge on sequencing technology. I would not be envious if the conclusion is that sequencing is the choice. But if microarray is the choice –

4) It would be ideal to have experts in industry join the endeavor to explore the opportunity to improve the microarray platform, such as whether 3D printing technology can be used to better fit a higher requirement to the microarray chip? It is also wise to have biomedical researchers' input on quality control such as RNA purity and accuracy of RNA amount of each experiment.

5) Biological experts and software engineers may work together to design the frame of the database storage and link to various functions for external research use.

All the best.

References:

1. Hashemi Karoii D, Bavandi S, Abroudi AS, Djamali M, Azizi H, Skutella T. Microarray and Single-Cell RNA Sequencing Reveals G-Protein Gene Expression Signatures of Spermatogonia Stem Cell. Stem Cell Rev Rep. 2025;21(7):2136-2156. doi: 10.1007/s12015-025-10942-4.
2. Hu Y, Cai ZR, Huang RZ, Wang DS, Ju HQ, Chen DL. Circular RNA circPHLPP2 promotes tumor growth and anti-PD-1 resistance through binding ILF3 to regulate IL36γ transcription in colorectal cancer. Mol Cancer. 2024;23(1):272. doi: 10.1186/s12943-024-02192-8.
3. Guo X, Hou P, Zhang S, Xu Q, Zhou M, Tang W, et al. Enhancer-Associated LncRNA-ITGA2 Promotes Vascular Remodeling Through ITGA2.
Circ Res. 2025;136(12):1610-1628. doi: 10.1161/CIRCRESAHA.124.325443.
4. Wu DP, Wei YS, Hou LX, Du YX, Yan QQ, Liu LL, et al. Circular RNA APP contributes to Alzheimer's

disease pathogenesis by modulating microglial polarization via miR-1906/CLIC1 axis. Alzheimers Res Ther. 2025;17(1):44. doi: 10.1186/s13195-025-01698-7.
5. Harada K, Sakamoto N, Kitaoka T, Nakamura Y, Kondo R, Morisue R, et al. PI3 expression predicts recurrence after chemotherapy with DNA-damaging drugs in gastric cancer. J Pathol. 2025;265(4):472-485. doi: 10.1002/path.6400.
6. Cipurko D, Ueda T, Mei L, Chevrier N. Repurposing large-format microarrays for scalable spatial transcriptomics. Nat Methods. 2025;22(1):145-155. doi: 10.1038/s41592-024-02501-5.
7. Gao X, Yourick MR, Campasino K, Zhao Y, Sepehr E, Vaught C, Sprando RL, Yourick JJ. An updated comparison of microarray and RNA-seq for concentration response transcriptomic study: case studies with two cannabinoids, cannabichromene and cannabinol. BMC Genomics. 2025;26(1):392. doi: 10.1186/s12864-025-11548-3.
8. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nat Methods. 2014;11(8):809-15. doi: 10.1038/nmeth.3014.
9. Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, et al. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. RNA. 2010;16(5):991-1006. doi: 10.1261/rna.1947110.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Figures.pdf

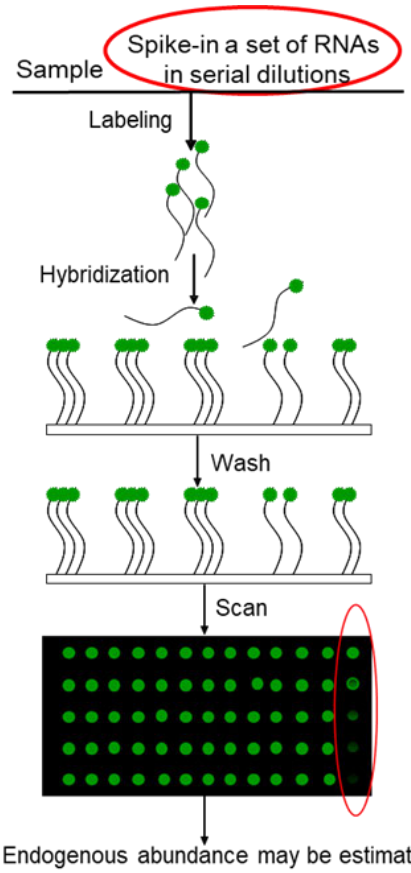


Figure 1. Scheme of spike-in. The reference of different copy numbers is based on spike-in a set of RNAs at each experiment.

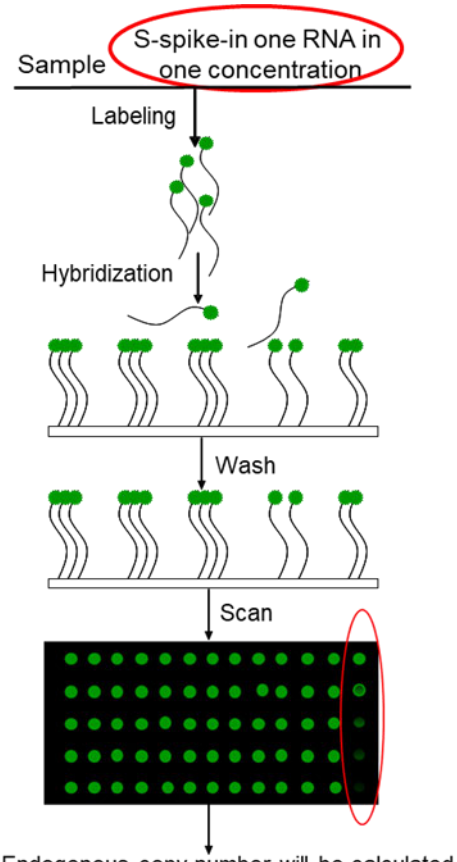


Figure 2. Scheme of S-spike-in. The reference of different copy numbers is printed on the glass via manufacturing.

238. Federation of Associations in Behavioral and Brain Sciences

I am responding: On behalf of an organization

Name: Lindsey Leary

Name of Organization: Federation of Associations in Behavioral and Brain Sciences

Type of Organization: Non-Profit Research Organization

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Behavioral and brain scientists have and continue to actively work to develop practices and norms to improve replicability. The field has scrutinized their journals and standard tools and made significant improvements to the statistical analyses and methodology. The Center for Open Science (COS) has provided critical infrastructure to enable rigorous and open practices. Scientific society publishers have changed journal standards and incentivized open science (OS) practices, including preregistration and sharing data/materials. These can improve rigor by making replication research easier and requiring scientists to think very deliberately about their research choices and how they might affect study findings. The American Psychological Association (APA), which publishes nearly 90 peer-reviewed journals, developed its Journal Article Reporting Standards (JARS), offering guidance on what information should be included in articles reporting quantitative, qualitative, and mixed methods research.

NIH might also lean on existing journals that explicitly welcome replication studies (e.g., Collabra: Psychology) or use Registered Replication Reports (RRRs; e.g., Advances in Methods and Practices in Psychological Science). In addition to looking at the publication count, it is important to consider where a paper was published.

According to COS, over 300 journals are tackling publication bias by offering a two-stage review process through the Registered Reports (RR) format. This approach prioritizes research quality over results in publishing decisions, which are made before the research is conducted. The Journal of Articles in Support of the Null Hypothesis only publishes research that does not have positive results.

In order to incentivize these changes and institutionalize replication and reproducibility efforts at the agency, NIH needs to fund these activities and tools with new calls. Some of this work, such as establishing new standard practices, may be best accomplished by scientific societies. By way of example, the FABBS Open Science Statement articulates norms for sharing data and research findings while protecting privacy of research participants. When evaluating grant proposals, NIH should reward researchers who not only follow these rigorous practices, but also researchers that devote their time and energy to reproducing the work of others.

2. Data, Software, and Model Sharing: In concert with COS, many journals use badges to signal when researchers participate in OS practices, including sharing data, study protocols, and analysis strategies. PubMed currently does not use the badge system itself; adopting it would make it easier to see which researchers have a history of data, software, and model sharing—something that could be considered in a grant application review. PubMed could track how often a dataset or protocol is reused, a clear

additional value.
Depositing data is only one step in strong open science practice. In order to be truly usable by colleagues, researchers must provide clear and detailed information for others to be able to reproduce or replicate their findings. FABBS urges NIH to carefully consider when certain datasets might not be feasible to share due to privacy concerns.
Science Tracking Across the Project Lifespan (STAPLE) is another example of the behavioral science community proactively improving open science practice. STAPLE is a science focused project management tool that not only helps with the unique challenges of project management of research, but includes open and transparent documentation of data and metadata.

3. Training and Mentorship: FABBS applauds the inclusion of training and mentorship as NIH considers broadening the measures of impact, albeit complicated. Behavioral and cognitive scientists have contributed to an evidence base for understanding and developing effective training and mentorship. We urge NIH to consider NASEM's The Science of Effective Mentorship in STEMM (2019); Chapter 6 is specifically devoted to assessment and evaluation.
The quality and impact of training and mentorship efforts can be measured in a number of ways. Primary investigators (PIs) could be required to submit detailed mentoring plans for any grants that involve recruiting research assistants, post-docs, or similar trainees who might benefit from the research experience, even if they are not funded by the grant. Following this, those trainees could report on their mentoring experiences with their PIs. Relying on just the number of mentees and trainees could provide incentives to maximize the number reported over the quality of the engagement.
Additionally, NIH grant PIs can track their mentees in several ways: replication efforts, grants awarded, and career outcomes. FABBS encourages NIH to consider not just mentees who stay in research, but also those who go into other fields and, nevertheless, have an important impact on public health.

4. Collaboration: FABBS applauds NIH's efforts to measure and reward collaboration. Team-based science is essential for maximizing the cumulative nature of science and the potential impact of the research. Furthermore, this research can be time consuming and not as conducive to acceptance in peer review journal articles that tend to be discipline specific. Indeed, a whole field of research exists on team science. In 2025, the National Academy of Sciences published Research and Application in Team Science. FABBS also encourages the agency to take into account a scientist's community-engaged research, considering how the program involves non-scientists in the research process.

5. Entrepreneurship and Translation: Pharmaceuticals and medical devices have clear pathways from basic research to product development. Industry is motivated to bring NIH-funded discoveries to scale when they are able to recoup their expenses and gain a financial profit. Unfortunately, the NIH-funded research and discoveries regarding prevention and reducing chronic diseases do not have the same financial incentives for being taken to scale.
FABBS encourages NIH to consider how to best measure the physical and financial benefits of prevention efforts — how can we measure the money saved and the lives improved when individuals avoid disease and disability thanks to healthy lifestyles and decisions?
NIH can also consider the impact a research program has had on clinical practice. For example, when an oversight body such as APA makes recommendations for how therapists treat certain patients or the American Psychiatric Association revises the Diagnostic and Statistical Manual of Mental Disorders (DSM), what research are they relying on? Can researchers be credited with contributing to clinical practice guidelines if their work is cited by oversight organizations?

NIH has made important advances with implementation science, with credit to behavioral scientists. FABBS encourages NIH to support additional work in implementation science as a key to effectively translating knowledge into health benefits.

6. Foundational Scientific Exploration: NIH recently released several notices of funding opportunities (NOFOs) for high-risk, high-reward (HRHR) projects. FABBS noted a significant change in the NOFOs: the reference to behavioral and social science had been systematically removed throughout, including in the routing code used to determine reviewers. This removal narrows the disciplinary range of scientists likely to respond to these NOFOs. FABBS encourages NIH to restore 'behavioral and social' science throughout the NOFOs and most definitely within the list of routing codes.

7. Public Impact: To date, the federal health ecosystem has fallen short in translating fundamental knowledge about health into practices and policies that reduce the burdens of illness and disease. NIH is only one part of the federal health infrastructure, yet plays a critical role. NIH could better incentivize researchers to consider and actively engage with the practitioners and policymakers in positions to incorporate knowledge and improve outcomes. Other federal agencies in the health delivery space who benefit greatly from leveraging insights and knowledge made possible by NIH-funded research. FABBS encourages NIH to recognize, measure, and value the impact of its research, as it serves the mission of sister agencies.
New tools exist to capture how governments, think tanks, and policymakers use relevant research to inform their decisions:
Sage Policy Profiles allows researchers to look up their own research and discover how their work has been used in policy decisions.
Altmetrics tracks online engagement.
There are also journals dedicated to the policy impact of science (e.g., Policy Insights from the Behavioral and Brain Sciences). We urge NIH to consider how other federal agencies use NIH-funded research in their work.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/FABBS-Comment-on-NIH-RFI-Measuring-Impact.pdf

FABBS Response to NIH Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Date: August 19, 2026

The Federation of Associations in Behavioral & Brain Sciences (FABBS) represents 35 scientific societies and 60 academic departments in the psychological, cognitive, and behavioral sciences. We advance the sciences of mind, brain, and behavior; promote evidence-based policymaking; and advance the integrity and impact of the federal scientific enterprise. With support from federal agencies, including NIH, FABBS scientists study critical health challenges such as maternal mortality, youth mental health, and chronic diseases.

FABBS values the opportunity to provide feedback on measuring and rewarding scientific impact. We agree that current indicators of scientific success, mainly publication counts and citation metrics, do not fully capture the scope of the impact of NIH funded research. In particular, these limited measures fail to assess the degree to which behavioral and cognitive research impacts well being, clinical practice, and policies. Furthermore, reliance on publications and citations can focus researchers on these measures and actually disincentivize meaningful scientific engagement with the potential to provide the greatest societal benefits. Below, we offer seven suggestions for how to improve measures of scientific rigor and impact.

(1) Rigor and Reproducibility.

Behavioral and brain scientists have and continue to actively work to develop practices and norms to improve replicability. The field has scrutinized their journals and standard tools and made significant improvements to the statistical analyses and methodology. The Center for Open Science (COS) has provided critical infrastructure to enable rigorous and open practices.

Scientific society publishers have changed journal standards and incentivized open science (OS) practices, including preregistration and sharing data/materials. These can improve rigor by making replication research easier and requiring scientists to think very deliberately about their research choices and how they might affect study findings. The American Psychological Association (APA), which publishes nearly 90 peer-reviewed journals, developed its Journal Article Reporting Standards (JARS), offering guidance on what information should be included in articles reporting quantitative, qualitative, and mixed methods research.

NIH might also lean on existing journals that explicitly welcome replication studies (e.g., *Collabra: Psychology*) or use Registered Replication Reports (RRRs; e.g., *Advances in Methods and Practices in Psychological Science*). In addition to looking at the publication count, it is important to consider where a paper was published.

According to COS, over 300 journals are tackling publication bias by offering a two-stage review process through the Registered Reports (RR) format. This approach prioritizes research quality over results in publishing decisions, which are made before the research is conducted. The

Journal of Articles in Support of the Null Hypothesis only publishes research that does not have positive results.

In order to incentivize these changes and institutionalize replication and reproducibility efforts at the agency, NIH needs to fund these activities and tools with new calls. Some of this work, such as establishing new standard practices, may be best accomplished by scientific societies. By way of example, the FABBS Open Science Statement articulates norms for sharing data and research findings while protecting privacy of research participants. When evaluating grant proposals, NIH should reward researchers who not only follow these rigorous practices, but also researchers that devote their time and energy to reproducing the work of others.

(2) ***Data, Software, and Model Sharing****.*

In concert with COS, many journals use badges to signal when researchers participate in OS practices, including sharing data, study protocols, and analysis strategies. PubMed currently does not use the badge system itself; adopting it would make it easier to see which researchers have a history of data, software, and model sharing—something that could be considered in a grant application review. PubMed could track how often a dataset or protocol is reused, a clear additional value.

Depositing data is only one step in strong open science practice. In order to be truly usable by colleagues, researchers must provide clear and detailed information for others to be able to reproduce or replicate their findings. FABBS urges NIH to carefully consider when certain datasets might not be feasible to share due to privacy concerns.

Science Tracking Across the Project Lifespan (STAPLE) is another example of the behavioral science community proactively improving open science practice. STAPLE is a science focused project management tool that not only helps with the unique challenges of project management of research, but includes open and transparent documentation of data and metadata.

(3) ***Training and Mentorship****.*

FABBS applauds the inclusion of training and mentorship as NIH considers broadening the measures of impact, albeit complicated. Behavioral and cognitive scientists have contributed to an evidence base for understanding and developing effective training and mentorship. We urge NIH to consider NASEM's The Science of Effective Mentorship in STEMM (2019); Chapter 6 is specifically devoted to assessment and evaluation.

The and impact of training and mentorship efforts can be measured in a number of ways. Primary investigators (PIs) could be required to submit detailed mentoring plans for any grants that involve recruiting research assistants, post-docs, or similar trainees who might benefit from the research experience, even if they are not funded by the grant. Following this, those trainees could report on their mentoring experiences with their PIs. Relying on just the number of mentees and trainees could provide incentives to maximize the number reported over the quality of the engagement.

Additionally, NIH grant PIs can track their mentees in several ways: replication efforts, grants awarded, and career outcomes. FABBS encourages NIH to consider not just mentees who stay in research, but also those who go into other fields and, nevertheless, have an important impact on public health.

(4) ***Collaboration.***

FABBS applauds NIH's efforts to measure and reward collaboration. Team-based science is essential for maximizing the cumulative nature of science and the potential impact of the research. Furthermore, this research can be time consuming and not as conducive to acceptance in peer review journal articles that tend to be discipline specific. Indeed, a whole field of research exists on team science. In 2025, the National Academy of Sciences published Research and Application in Team Science.

FABBS also encourages the agency to take into account a scientist's community-engaged research, considering how the program involves non-scientists in the research process.

(5) ***Entrepreneurship and Translation.***

Pharmaceuticals and medical devices have clear pathways from basic research to product development. Industry is motivated to bring NIH-funded discoveries to scale when they are able to recoup their expenses and gain a financial profit. Unfortunately, the NIH-funded research and discoveries regarding prevention and reducing chronic diseases do not have the same financial incentives for being taken to scale.

FABBS encourages NIH to consider how to best measure the physical and financial benefits of prevention efforts — how can we measure the money saved and the lives improved when individuals avoid disease and disability thanks to healthy lifestyles and decisions?

NIH can also consider the impact a research program has had on clinical practice. For example, when an oversight body such as APA makes recommendations for how therapists treat certain patients or the American Psychiatric Association revises the Diagnostic and Statistical Manual of Mental Disorders (DSM), what research are they relying on? Can researchers be credited with contributing to clinical practice guidelines if their work is cited by oversight organizations?

NIH has made important advances with implementation science, with credit to behavioral scientists. FABBS encourages NIH to support additional work in implementation science as a key to effectively translating knowledge into health benefits.

(6) ***Foundational Scientific Exploration.***

NIH recently released several notices of funding opportunities (NOFOs) for high-risk, high-reward (HRHR) projects. FABBS noted a significant change in the NOFOs: the reference to behavioral and social science had been systematically removed throughout, including in the routing code used to determine reviewers. This removal narrows the disciplinary range of scientists likely to respond to these NOFOs. FABBS encourages NIH to restore 'behavioral and social' science throughout the NOFOs and most definitely within the list of routing codes.

(7) ***Public Impact.***

To date, the federal health ecosystem has fallen short in translating fundamental knowledge about health into practices and policies that reduce the burdens of illness and disease. NIH is only one part of the federal health infrastructure, yet plays a critical role. NIH could better incentivize researchers to consider and actively engage with the practitioners and policymakers in positions to incorporate knowledge and improve outcomes. Other federal agencies in the health delivery space who benefit greatly from leveraging insights and knowledge made possible by NIH-funded research. FABBS encourages NIH to recognize, measure, and value the impact of its research, as it serves the mission of sister agencies.

New tools exist to capture how governments, think tanks, and policymakers use relevant research to inform their decisions:

- Sage Policy Profiles allows researchers to look up their own research and discover how their work has been used in policy decisions.
- Altmetrics tracks online engagement.

There are also journals dedicated to the policy impact of science (e.g., *Policy Insights from the Behavioral and Brain Sciences*). We urge NIH to consider how other federal agencies use NIH-funded research in their work.

In closing, FABBS strongly supports expanding efforts to measure and reward scientific impact and appreciates the opportunity to comment on these initial categories. As NIH moves forward developing new measures for impact, we encourage the agency to carefully consider how — given the breadth and depth of NIH-funded research — there may be considerable differences across disciplines and research questions. Accordingly, FABBS encourages NIH to continue to engage with the scientific community. We look forward to seeing your progress and providing feedback on the next step in the process.

Thank you again for the opportunity to comment. FABBS would be delighted to be a resource.

Sincerely,



Juliane Baron, Executive Director

239. Springer Nature

I am responding: On behalf of an organization

Name: Kaia Motter

Name of Organization: Springer Nature

Type of Organization: Publisher

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Springer Nature welcomes NIH's agency-wide focus on strengthening rigor, reproducibility, and replication as the foundation for gold standard biomedical science. NIH has emphasized that replication and reproducibility are essential to scientific progress, public trust, and the ability of research findings to benefit patients and communities. Springer Nature supports this commitment with infrastructure and peer review services, open access, and insight into researcher motivations and community norms.

Springer Nature already has the necessary infrastructure to ensure that NIH-funded findings are rigorously evaluated, clearly reported, openly accessible where possible, and linked to the underlying research objects. By publishing the results of NIH funded research in our journals, researchers ensure that they are properly archived and preserved as part of a trustworthy scholarly record that can be scrutinized, reused, and built upon over time. We offer analytical capabilities to help NIH identify trends in the literature and measure impact. Public access to NIH funded research, including all research outputs, allows for broader replication, reproducibility, and re-use, and is enabled with centralized open access funding models and proper licensing and attribution. Building on the lessons learned and best practices from fields like social and behavioral science, that have been focused on the replication, robustness and reproducibility of results for many years, saves time and effort. Learning from researchers directly about their motivations and barriers to participation can help shape the NIH's incentive structure and ensure compliance.

Infrastructure
Springer Nature's publishing infrastructure directly supports trustworthy, reproducible biomedical research through editorial oversight, independent peer review, research integrity screening and post publication correction and retraction processes, standardized competing interest and funding disclosure, adherence to the relevant biomedical reporting standards, and mandatory publication of data availability statements. Publication of the Version of Record (VoR) supports NIH's goals because it provides the context necessary to validate the results and ultimately accelerates future replication efforts.

The ability to reproduce and replicate results is enabled by access to clear and informative protocols and methods. In addition to Springer Nature's journal publishing infrastructure, we host the world's large collection of protocols and methods, available on SN Experiments (https://experiments.springernature.com/), which provides a resource for any researcher engaged in

replicating and reproducing research across all subject areas. NIH funded researchers can also benefit from recording their own protocols on protocols.io, which provides institutional plans for funders who wish to collate their efforts in a dedicated repository.

The flagship Nature Portfolio has a history of pioneering reporting initiatives to improve the reproducibility and replication of the research it publishes. In 2013, following a wide-reaching community consultation, that included workshops organized by institutes within the NIH, Nature Portfolio journals mandated authors in the life sciences to provide a checklist (https://www.nature.com/articles/496398a) that reported the key technical and statistical information necessary for referees and readers to assess the reproducibility of their work. Independent analysis by Macleod et al (https://access.portico.org/Portico/auView?auId=ark:%2F27927%2Fphzq8gmxdp1) concluded that this had an immediate and measurable effect on the reproducibility of published research. These requirements were subsequently extended (https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards) to the behavioral and social sciences, evolution and environmental sciences, ecology, and to specific claims in the physical sciences regarding the performance of materials and devices such as lasers and solar cells.

Springer Nature has long advocated for preregistration and the use of the Registered Reports format that prioritize rigorous experimental design over the specific implications of the results those experiments lead to. This approach seeks to address concerns about so-called publication bias in which papers that report seemingly interesting results are more likely to be published than papers whose methodologies are rigorous. The use of this format has historically been limited to behavioral and social sciences as well as the life sciences. Recently, Nature expanded the scope of the format (https://www.nature.com/articles/d41586-026-01629-y) to allow them to be submitted from any scientific discipline.

As NIH considers strategies for identifying new areas to expand replication and reproducibility efforts, the Agency can look to evidence from the increased adoption of reporting tools such as registered reports or checklists beyond the behavioral and social sciences to review any discipline specific modifications they may need to be made.

Public Access
53% of the primary research published at Springer Nature is now open access (https://stories.springernature.com/oa-report-2025/index.html). Open access publication continues to deliver impact through unrestricted engagement and licensed re-use of our machine-readable research publications. Open access delivers higher rates of citations (policy, patent, and article citations) and expanded media references. Public access is an essential component of all reproducibility and replication work and increasingly AI tools can be applied to investigate citations that have been supported or contradicted by later research, e.g. tools like scite.ai.

NIH can accelerate progress towards greater reproducibility and replication by treating scholarly communication, research integrity workflows, and long-term stewardship of the scientific record as a necessary part of the national biomedical research infrastructure and making results publicly accessible. However, peer review services are an essential component of the validation of results and without the

necessary funding, trust and integrity are compromised. Grant policies and budgets should allow researchers to plan and pay for high-quality peer reviewed publication, data and code curation, protocol sharing, and preservation in ways that advance rigor, transparency, and reuse.

Behavior and Incentives

To create incentives, NIH should recognize replication studies, null hypotheses, preregistration, reusable datasets, software, protocols, and other open science contributions as valued research outputs. Springer Nature already provides publication platforms and infrastructure to support many of these practices, including journals that publish null hypotheses and replication studies, dedicated data journals such as Scientific Data and BMC Research Notes, as well as data article formats. DOI assignment to a broad range of research outputs ensures NIH researchers receive the credit and acknowledgement they deserve. These mechanisms can help NIH reward researchers who strengthen the reliability of the biomedical evidence base, not only those who produce novel positive findings.

Springer Nature publishes over 3,000 peer reviewed scholarly research journals, working with 180,000 + academic editors to publish over 500,000 articles per year. For nearly two centuries, Springer Nature has supported the communication and dissemination of American science and over time we have learned a lot about author behavior. Springer Nature has always been at the forefront of open science innovation and has piloted many open science initiatives. These initiatives have helped us to better understand the impact of funder policy on compliance, the motivation of researchers to adopt the open sharing of research artefacts across disciplines, and barriers to participation and adoption. In addition, we periodically conduct large scale surveys of the researchers who publish with us to better understand their attitudes related to open science policy, research assessment, and research integrity training and practices. If NIH wants to develop incentives to support the reproduction and replication of existing findings, there are valuable insights to be gained from collaboration with the publishers who support researchers in the dissemination of the research outputs. Recent results from our State of Research Assessment survey (https://stories.springernature.com/state-of-research-assessment/index.html) and our 2026 Research Leaders survey (https://www.nature.com/nature-index/about-research-leaders-survey-2026) as well as our State of Open Data report (https://stateofopendata.com/) and State of Null Results report (https://stories.springernature.com/the-state-of-null-results-white-paper/index.html), contain insights into what motivates researchers, what they value, but also about what is preventing them from fully participating in the advancement of science and sharing of results. This could help inform the way NIH structures its approach to incentives.

In terms of community norms and disciplinary trends, NIH could learn from work that has been done within key subject disciplines that advance the Agency's understanding of the impact of policy interventions and best practices. Springer Nature explored progress in replication and reproducibility in the field of social and behavioral science earlier this year, publishing a collection of articles that shed insight on the challenges that exist and how we can address them. The journal Nature Human Behaviour recently partnered with the Institute for Replication (https://www.nature.com/articles/s41562-024-01818-7) to attempt the reproduction and replication of papers published in the journal from 2023 onward, at scale. This effort is currently underway, but we would be happy to share the results as soon as they're available.

We recognize the benefits of highlighting replication efforts, which are often buried amongst other research findings, to amplify individual contributions and enhance recognition and reward. We do this by aggregating research findings and hosting them on an accessible, discoverable platform. Our most recent example can be found here. Platforms that recognize individual efforts encourage researchers to share their results which is a powerful incentive.

Springer Nature is well positioned to partner with NIH to drive, incentivize, and institutionalize replication and reproducibility across the biomedical research enterprise. Through sustained investment in the infrastructure that supports trustworthy scholarly communication, Springer Nature can help NIH reduce administrative burden, scale good practices, and embed reproducibility throughout the research lifecycle.

A partnership with NIH would build on existing infrastructure and experience rather than duplicating it, helping NIH maximize the reliability, reach, and public impact of the research it funds.

2. Data, Software, and Model Sharing: Springer Nature is committed to providing a variety of high-quality, peer-reviewed journals that offer immediate, free public access to the final published article—the Version of Record (VoR).

Access to the VoR enables access to the research artefacts generated throughout a study and strengthens confidence in research results and conclusions through transparency and openness—the cornerstones of Gold Standard Science. We are well equipped to support immediate access to the full spectrum of NIH-funded research results in the following ways:

- Open access article publication in our hybrid and fully open access journals provides immediate public access, impact, and discoverability for the VoR, while enabling immediate deposit of the VoR to Agency repositories.
- Automated submission integrations with Research Square (preprints), Figshare (data), protocols.io (protocols), and Code Ocean (code) connect research outputs from every stage of the research lifecycle to the VoR and allow researchers to receive credit for their work. Linking the VoR to related data, code, protocols, and other outputs supports trust and effective reuse.
- Introduction of a standard Research Data policy for journals and books that requires publication of a Data Availability Statement. We are currently piloting cutting-edge automation to help authors develop statements that include the context needed to support reproducibility and replication. In addition, our AI-assisted data deposition workflow, currently being piloted on Nature Communications and Scientific Reports, will remove a significant administrative burden for authors who need data deposition support.
- Outlets for “sound science” publication outputs, including null hypotheses and preliminary results, through the BMC Series, Scientific Reports, and the recently launched Discover and Cureus journals.
- Publication and citation of high-value research data through dedicated data journals and data articles, including Scientific Data and BMC Research Notes, help address the credit gap for data sharing.

Together, these integrations, tools, and publication platforms provide resources that facilitate reproducibility and replication, where provenance and trust indicators must be reliable, interlinked, and machine-readable.

Springer Nature provides the infrastructure and services that enable NIH-funded researchers to share

scientific results with the necessary context and validation. Persistent identifiers, uniform data and software citation practices, and metadata standards set by industry bodies like NISO, RDA, and others allow NIH to track compliance consistently while expanding opportunities for attribution and recognition of individual research contributions. Repository and database integrations, along with bidirectional linking through the VoR, support version control and uphold trust in the integrity of the scientific record. Publishers have spent years investing in infrastructure to support interlinked research objects, as well as technologies that automate deposition activities and reduce administrative burdens. We are committed to FAIR principles and to licensing agreements that accelerate discovery while ensuring that necessary legal and ethical guardrails are in place.

Springer Nature's State of Open Data report outlines many of the barriers that prevent researchers from sharing all of their research outputs. NIH could help remove some of these obstacles by directing its attention to the administrative burdens that policy mandates impose on researchers and authors across the system, while also supporting appropriate incentive structures.

3. Training and Mentorship: In 2019, Nature hosted a meeting of stakeholders from across the Australian research community—including representatives from business, government bodies, universities, research institutes, and funding organizations—to discuss research integrity and good research practices. One of the most striking outcomes of the meeting was the recognition that little was known about researchers' level of understanding of research integrity, the training available to them, or the data needed to assess that training. This led Springer Nature to launch a series of surveys of researchers in different parts of the world to address this knowledge gap. These surveys were also intended to provide a baseline for the level of support researchers currently have access to at a time when improving research integrity is an increasing priority for stakeholders across the research ecosystem.

Between 2021 and 2024, Springer Nature surveyed (https://doi.org/10.6084/m9.figshare.29100596.v1) 7,871 researchers across Australia (993), the UK (1,078), the U.S. (1,962), India (659), Japan (1,190), China (1,008), and Brazil (981). Respondents in China (79%), Japan (73%), and Australia (68%) reported the greatest access to research integrity training. In Brazil, the proportion of respondents reporting access to training was substantially lower (27%). Just over half of respondents in the U.S. (56%), India (53%), and the UK (51%) reported having access to research integrity training.

Across countries, the overwhelming majority of respondents (84–94%) supported mandatory research integrity training at some point in a researcher's career. In many countries—particularly Australia, the U.S., the UK, and Brazil—researchers expressed interest in additional training on the curation and sharing of research data. Many respondents also expressed concern that the training they received was not tailored to the immediate needs of their research practice and functioned more as a box-ticking exercise for their institutions.

Springer Nature has used these results to open conversations with stakeholders across the research community about how to better support researchers in understanding best practices and conducting research with integrity. This work has led to the development of a suite of courses offered at no cost to editors, referees, and researchers, including a course on the fundamentals of research integrity (https://www.springernature.com/gp/authors/campaigns/research-integrity-course) and a course in publication ethics (https://www.nature.com/masterclasses/publication-ethics/52046828).

Since 2005, the annual Nature Awards for Mentoring in Science (https://www.nature.com/nature/awards/mentorship) have recognized and celebrated outstanding contributions by those who support early-career researchers in their scientific development.

NIH might consider the results of our survey as the Agency evaluates its current training program, especially as they to research integrity and data curation, and supplement its internal training program with externally available resources.

4. Collaboration: Not all collaborations result in publications, and not all publications include every contributor to a collaboration. However, publications provide one approach for assessing and recognizing both individual and team-based scientific contributions.

Springer Nature journals support both consortia authorship and team-based scientific collaborations, as well as individual author contribution statements, in ways that ensure roles and responsibilities are clearly and transparently acknowledged, and credit is assigned where credit is due.

The CRediT (Contributor Roles Taxonomy) offers another way to list author contributions across various research outputs, including preprints, data sets, and peer-reviewed journal articles. There are numerous tools that can help organize and record the methods and results of collaboration, including electronic lab notebooks and protocols.io, which allow researchers to be more productive, improve collaboration and recordkeeping, and foster a culture of transparency and accountability. NIH should encourage the use of these tools and taxonomies to track and acknowledge research collaborations resulting from NIH funding.

New forms of cross-sector and community-engaged research will drive local and regional public impact, workforce development, and commercial innovation. The results of these collaborations should be documented, preserved in the scholarly record, and made publicly accessible for the benefit of society. These collaborations may take more time to develop and execute to ensure that all participants endorse the terms of the collaboration and plans for disseminating the results, but documenting this process and its outcomes will ultimately uphold trust in the scientific enterprise.

NIH can encourage and reward team science and other forms of community-engaged research by creating funding models that specifically incentivize this type of research and require publication of results. This would help build a corpus of work that supports best practices and smooths the pathway to replication and reproducibility.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: NIH has an important opportunity to broaden how it assesses and recognizes the public impact of biomedical research. Rather than relying primarily on traditional impact metrics, such as the Impact Factor, NIH could adopt a more holistic, evidence-based framework that captures how research improves health, informs clinical practice, strengthens public trust, supports economic growth, and delivers societal benefit. This is not an easy task, as many of these outcomes are difficult to measure

quantitatively. However, journal and article-level impact analysis can offer useful insights.

For NIH-funded research to have public impact, findings must be rigorous, trusted, discoverable, accessible, reusable, and connected to the broader evidence base. Public benefit is realized not simply when research is completed, but when validated findings can be linked to real-world outcomes, such as improved quality of life, health policy, successful drug development, and informed patients. NIH should assess impact throughout every stage of the research lifecycle, including post-publication application and reuse.

The published Version of Record (VoR) is the foundation of trust on which impact is built. It allows the public to access and verify research results, understand the peer-review model, examine competing interests and sources of funding, and confirm that appropriate ethical and reporting standards were followed. These features are critical trust signals for clinicians, policymakers, researchers, patients, journalists, AI systems, and the public. Without them, it becomes harder to distinguish validated biomedical knowledge from unverified claims, increasing the risk that inaccurate or incomplete information will be amplified as evidence.

Metrics and Measurement

NIH should develop an impact framework that combines quantitative indicators with qualitative impact narratives. Traditional citation measures remain useful for understanding scholarly uptake, but they should be complemented by indicators that better reflect real-world impact. Quantitative measures might include replication and reproducibility outcomes; citations in clinical guidelines, patents, or policy documents; clinical trial registrations; media references; or other citation-based metrics that can be measured in a standardized way across the published scholarly record. Qualitative measures, or real-world outcomes, are more difficult to assess and fall outside the realm of our expertise, but we would encourage NIH to explore options with local and community leaders, patient advocacy groups, and other key stakeholders who can help expand qualitative measurement of NIH-funded research. NIH should look for evidence of downstream health outcomes, economic activity, workforce development, community benefit, or improved public understanding. At the same time, NIH should recognize that public impact is not always immediately quantifiable. Some impacts are immediate, such as public access to findings or reuse of a dataset. Others may take years to emerge, such as changes in clinical practice or policy, development of a therapeutic application, or measurable improvements in population health. NIH should therefore avoid a one-size-fits-all metric and instead use field-appropriate indicators, transparent evaluation criteria, and flexible narrative reporting that allows investigators to describe the pathway from research output to public benefit.

Traditionally, research impact in academic environments has been measured through journal citation metrics. However, Springer Nature recognizes the need for a more holistic approach to impact analysis, one that includes measures of engagement, patent and policy citation, and other metrics that allow funders and institutions to quantify socio-economic impact.

Springer Nature is well positioned to support NIH in developing and piloting expanded measures of impact. As a publishing partner for biomedical and related research, Springer Nature's experience with impact analysis—including approaches that examine engagement, patent and policy citation, open data

sharing, and impact narratives—could help NIH develop practical models for assessing returns on public research investments without relying narrowly on publication counts or journal-level metrics.
Our Impact Solutions team is developing bespoke impact narrative reporting and tools for funders and institutions to help them measure the impact of the research they have funded and their return on investment (ROI). An example of a recent project measuring socio-economic impact through article-level analysis can be found in our "From Publications to Policy" report (https://stories.springernature.com/sdg-impact-report/executive-summary/index.html), produced in collaboration with Overton. In 2024, we conducted a similar exercise (https://doi.org/10.6084/m9.figshare.27337476.v2) in partnership with Digital Science, using analysis of open data sharing to quantify policy impact.

To better align with DORA principles, Springer Nature publishes up-to-date article-level metrics, including citations and other engagement metrics. Authors can reference this data in grant applications, and funders can use it to track research impact.

Science Communication
Clear, targeted scientific and medical communication is essential to achieving public impact. When research is communicated appropriately to the relevant audience, it is more likely to be understood and trusted. Peer-reviewed scientific and medical literature is often not accessible, or easily understandable, to the public, patients, journalists, policymakers, or even non-specialist healthcare professionals who could benefit from the information. Springer Nature has recognized this challenge and taken steps to make research more accessible to the broader community. We have developed several models that improve the communication of biomedical results by synthesizing key findings and making them accessible to patients, practitioners, and members of the public who wish to better understand the science. Within our medical and health science journals, including the Adis journals, we encourage the publication of Plain Language Summaries (PLS) and Summaries of Research Articles (SRA) (https://link.springer.com/collections/bfgfdabjig). PLSs are peer reviewed as part of the source article and are subsequently indexed in PubMed. SRAs are standalone summaries of the source article that are assigned a DOI and indexed in PubMed. They are intended to be read quickly and understood by a wider, non-specialist audience, primarily healthcare professionals (HCPs). These publications are curated in collections that increase discoverability, access, and visibility for the authors who produce them, which encourages participation and broader recognition. SRAs also increase engagement with the source article. In addition, we have been experimenting with AI-generated video summaries that ensure scientific accuracy through an essential author validation process. Our experience thus far is that these initiatives have had a strong impact on article engagement—usage, citations, and Altmetric references—all meaningful indicators of broader impact.

Many researchers and clinicians are experimenting with alternative modes of communication and article types that translate complex scientific results and make them accessible through podcasts, vodcasts, blog posts, and social media posts. Many Springer Nature publications support these forms of communication, including targeted Patient/Physician Perspectives. Several of our journals include patient authors who incorporate evidence of public impact through real-world, lived patient experience. Policymakers are another important stakeholder in the research funding ecosystem, but they may require a narrative style that differs from communication intended for journalists or patients, for

example. With this in mind, several Nature journals have developed a document type called the Policy Brief, which is specifically designed for policymakers and contains the key information needed to communicate effectively and clearly with that audience. Examples are available here in Nature Energy. Springer Nature has also developed new AI tools to make science more accessible to non-specialist readers. The Nature Research Assistant (https://natureresearchassistant.com/), a reading assistant currently in beta, allows readers to quickly understand material, chat in real time, and obtain detailed descriptions of each figure. We are also piloting a "Chat with Article" feature across some of the Discover journals (open access), which allows users to ask questions, explore key concepts, and summarize findings directly alongside the full-text source article.

While there is no replacement for the peer-reviewed journal article as the foundation of scientific communication, NIH could consider recognizing and rewarding researchers who successfully communicate their research beyond the biomedical community and achieve public impact through expanded modes of communication. These may include approaches that incorporate the perspectives of patients, for example, or tailor communication to specific target audiences. NIH-funded researchers would receive credit and acknowledgement through DOI assignment, which enables article types such as SRAs to be linked to VoR and the author's profile. NIH can also easily track these outputs through indexing in PubMed and the use of persistent identifiers.

Our research assessment survey results indicate that researchers would like greater recognition for work that has societal and public impact. NIH is well positioned to recognize and reward these activities through the proper incentive structure.

NIH's replication and reproducibility efforts should be treated as central to public impact assessment. Reliable biomedical research is foundational to public trust and to ensuring that findings can ultimately benefit patients and communities.

NIH should also ensure that public impact measures account for the range of outputs from biomedical research. Foundational, preclinical, translational, clinical, population health, behavioral, data science, and implementation research may produce different forms of impact over different time horizons.

Finally, NIH should align recognition and funding mechanisms with the behaviors it seeks to encourage. If NIH wants publicly funded biomedical research to be trusted, accessible, reproducible, and useful, then grant policies, review criteria, progress reporting, and career incentives should recognize the full range of contributions that make impact possible. This includes not only novel and groundbreaking discoveries, but also data stewardship, replication studies, null hypotheses, methods development, and community or patient engagement. By treating these activities as essential to public impact rather than as ancillary dissemination costs, NIH can strengthen public trust, improve the reliability of biomedical evidence, accelerate translation into practice, and maximize the return on taxpayer investment in medical research.

8. Other Areas: Springer Nature appreciates the opportunity to comment on NIH's Request for Information on Measuring and Rewarding Scientific Impact. Research assessment reform and expanded measures of impact are key priorities for us, and we are eager to support NIH in this effort.

We encourage NIH to consider the following three high-level recommendations as the Agency takes steps to address the areas outlined in this RFI:

• Springer Nature urges NIH to partner with academic institutions, publishers, and community stakeholders, rather than acting in isolation. We encourage NIH to learn from best practices and feedback from the research community and to take advantage of existing infrastructure, analytical capabilities, and industry standards that can ensure consistency and reduce administrative burden.
• We encourage NIH to develop incentive structures that reward a broader range of research outputs in addition to rewarding behaviors and practices that support the ethical sharing of results, stakeholder consultation, openness, transparency, and cross-border and cross-sector collaboration with trusted allies and industry partners. These collaborations should respect copyright frameworks.
• The NIH recognizes the benefits of public access to taxpayer funded results, but this should come with a commensurate recognition of the need for adequate funding to support peer review, quality control services, and trust in the scientific record. Publication of the peer-reviewed Version of Record helps mitigate version-control issues and provides a hub for interlinked research outputs across the research lifecycle.

Cultivating a healthy and productive research culture requires a shared and unified commitment from stakeholders across the U.S. research ecosystem. We look forward to partnering with NIH and working with NIH-funded researchers who publish in our books and journals to help achieve the Agency's goals.

240. The Alliance for Longevity Initiatives (A4LI)

I am responding: On behalf of an organization

Name: Brenda Eap

Name of Organization: The Alliance for Longevity Initiatives (A4LI)

Type of Organization: Non-Profit Research Organization

Role: Research Participant/Patient/Advocate/Caregiver

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: A. Key Bottlenecks
1) Publication Saturation vs. Translational Reality: Standard metrics reward publication volume in hyper-explored pathways (>1,000 papers in 5 years) while actively penalizing underexplored, high-risk research questions and deep mechanistic characterization.

2) Systemic Bias Against Null/Negative Results: Academic publishing imposes a substantially higher barrier to publication for null findings in aging biology and disease-modifying assays compared to positive outcomes. Researchers face demands for disproportionate secondary validation, lower impact venues, and prolonged review cycles for negative data—creating a strong structural disincentive to report null outcomes and leading to the costly, wasteful duplication of unviable experiments across labs.

B. Measurable Indicators & Recommendations

1) Mandatory Public Negative Results Archive: Establish a central, publicly accessible, and searchable NIH repository requiring grantees to archive negative and null experimental outcomes. To incentivize adoption, NIH must formally recognize documented repository submissions as valid, citable scientific contributions equivalent to traditional journal publications. Additionally, instruct the Center for Scientific Review (CSR) to award positive scoring modifiers under "Investigator" and "Environment" criteria for PIs who routinely deposit pre-registered protocols, negative findings, and open-access datasets.

2) RPPR Modernization for Non-Traditional Scientific Outputs: Update annual Research Performance Progress Report (RPPR) reporting guidelines—specifically under Section C (Products) and Section B (Accomplishments) to explicitly instruct grantees on how to report pre-registered protocols, published null/negative results, and open-access dataset depositions. Program Officers should be directed to evaluate these documented rigor outputs as formal metrics of scientific success during annual progress reviews and grant renewals, ensuring PIs receive official credit even before or in lieu of traditional journal publication.

3) Standard Pre-Registered Verification Tracks: Create explicit, ring-fenced funding mechanisms—such as a dedicated "R-Verify" grant track (R21 or U-series variants)—for independent protocol replication. Expand multi-site animal and cellular intervention pipelines (modeled after the NIA ITP, CITP, and Nathan Shock Centers of Excellence) that evaluate studies based on the translational importance of the original

claim and statistical power rather than novelty.

4) T32 "Hands-On Replication Modules": Modernize institutional training grants by incorporating mandatory "Hands-On Replication Modules," requiring trainees to execute and archive pre-registered replication studies of core field findings. Evaluate T32 renewal applications based on program open-data deposition rates and trainee-led null/replication publications to align institutional incentives with scientific integrity.

5) Career Development for Rigor Specialists: Establish specialized career transition awards (K99/R00 alternatives) dedicated to early-career researchers who specialize in quantitative validation, cross-lab verification, methodology development, and open-dataset curation.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: A. Key Bottleneck: Academic training pathways over-emphasize traditional publication metrics while under-preparing trainees for the full career ecosystem of biomedical translation—including science policy, regulatory affairs, technology transfer, public health, and entrepreneurship.

B. Measurable Indicators & Recommendations

1) Translational & Policy Summer Experiential Grants: Expand institutional training and short-term research education grants (e.g., R25, T35, or administrative supplements to T32s) to fund 8- to 12-week off-site summer immersion fellowships. Modeled on successful programs like the NIA Summer Training in Aging Research (STAR) program, these funded tracks would place graduate and postdoctoral trainees directly inside federal policy offices (e.g., OSTP, FDA, Congressional committees), venture funds, regulatory consulting firms, or biotech startups to gain hands-on translational competencies outside traditional academic labs.

2) Intramural Research Program (IRP) Modernization: Restructure the Intramural Research Program by shifting from indefinite tenure models toward fixed-term, high-impact appointments. IRP scientists should be measured on landmark, risk-tolerant scientific advances and trainee placement into interdisciplinary translational leadership roles.

4. Collaboration: A. Key Bottleneck: Institutional & Review Barriers to Team Science: NIH scoring frameworks remain optimized for isolated single-PI or parallel sub-award models. These mechanisms discourage true multi-institutional integration, penalize complex interdisciplinary designs, and offer no structural financial incentives for senior PIs to co-lead grants with junior researchers.

B. Measurable Indicators & Recommendation
1) Cross-Disease & Cross-IC Synergy Score: Reward proposals that demonstrate cross-cutting mechanisms (e.g., targeting cellular senescence or metabolic dysfunction common to cancer, neurodegeneration, and cardiovascular disease simultaneously).

2) IDeA State Satellite & Pilot Project Requirements for Center Grants: Mandate or heavily weight collaborative satellite/pilot project funding within major center and program awards (e.g., P01s, U54s,

ADRCs). Requiring these large infrastructure awards to formalize sub-award or pilot partnerships with institutions in Institutional Development Award (IDeA) states will drive geographic funding distribution away from coastal concentration and build sustainable regional capacity.

3) MPI Protection & ESI-Co-MPI Expansion: Ensure that any proposed PI grant limits explicitly exclude or weight co-MPI awards that pair an established senior PI with an Early Stage Investigator (ESI). To incentivize genuine co-leadership without penalizing overall PI grant capacity, provide an expanded direct-cost ceiling (e.g., an additional $50,000–$100,000/year) for applications where the junior PI manages an independent sub-budget (minimum 30% of direct costs).

5. Entrepreneurship and Translation: A. Key Bottleneck: The SBIR/STTR application process remains excessively slow and structurally misaligned with private capital deployment cycles. While NIH has introduced Entrepreneur-in-Residence (EIR) programs and mentorship networks (e.g., SEED), these figures serve in purely advisory capacities without budget authority, direct grant oversight, or selection power. As a result, commercialization awards are still evaluated and managed by traditional academic study sections and career program officers who lack direct venture, industry, or regulatory execution experience—leading high-potential biotech spinouts to view NIH non-dilutive funding as inefficient compared to private capital.

1) SBIR/STTR Centralization & Venture-Experienced Program Directors: Shift EIRs and venture experts from peripheral advisory roles into formal Program Director/Manager positions with explicit budget authority and milestone-based award power (modeled after DARPA/ARPA-H PMs). Centralize SBIR/STTR program administration into a unified office staffed by leaders with proven private-sector, venture, or biotech drug development experience.

2) FDA-NIH Joint Taskforce on Preventive Therapeutics: Mandate direct structural coordination between NIH translational divisions and the FDA to formalize approval pathways, surrogate biomarkers, and composite clinical trial endpoints for therapies targeting common age-related disease drivers and multi-morbidity mechanisms.

3) Expedited Review for Trained Commercializers: Provide streamlined grant review considerations and expedited turnaround times for academic investigators who complete NIH-recognized entrepreneurship accelerator modules (e.g., I-Corps or institute-backed commercialization bootcamps) before applying for commercialization funding (SBIR/STTR).

6. Foundational Scientific Exploration: A. Key Bottleneck: Study sections systematically penalize high-risk, paradigm-shifting proposals in favor of incremental extensions of established findings. This risk aversion is compounded by the NIH's organ- and disease-specific IC structure, which lacks dedicated mechanisms to fund foundational biology that spans multiple chronic disease pathways simultaneously.

B. Measurable Indicators & Recommendations
1) Dual-Stage "Anonymized Concept" Review for High-Risk Junior Awards: Establish a mandatory two-stage review process for high-risk seed mechanisms targeting Early Stage Investigators (ESIs). In Stage 1, study sections evaluate and score a 3-page conceptual grant proposal anonymously—strictly stripped of PI name, institutional affiliation, and publishing history—focusing solely on the transformative potential

and logic of the hypothesis. Only applications scoring in the top percentile advance to Stage 2, where feasibility and environment are assessed, effectively removing the "trust factor" penalty from initial concept evaluation.

2) Expansion of "Preliminary-Data-Free" Junior Mechanisms: Expand preliminary-data-exempt mechanisms (modeled on the Director's New Innovator Award - DP2) specifically tailored for early-career researchers testing paradigm-shifting hypotheses. Reviewers for these tracks must be formally instructed to evaluate logic, feasibility of design, and potential impact rather than demanding extensive preliminary datasets that junior labs lack the capital to generate.

3) Cross-Institute Geroscience & Foundational Consortia: Establish formal, co-funded Cross-Institute Consortia (jointly funded by ICs such as NIA, NHLBI, NCI, and NINDS) equipped with dedicated, ring-fenced grant budgets and specialized review panels. These consortia would explicitly evaluate and fund foundational research targeting common biological mechanisms of aging and tissue dysfunction across multiple disease silos, insulating cross-cutting, high-risk hypotheses from standard organ-specific study section bias.

4) Structural Realignment via NIHLAR: Establish a dedicated National Institute for Healthy Longevity and Aging Research (NIHLAR) to concentrate geroscience, healthspan extension, and foundational aging biology into a unified institute. Establishing NIHLAR decouples foundational preventive biology from organ-specific disease silos, ensuring bold, transformative hypotheses receive focused peer review, targeted appropriations, and dedicated leadership.

7. Public Impact: A. Key Bottleneck: Grants are rarely re-evaluated dynamically during their funding lifecycle, resulting in sustained funding for stagnated or redundant programs.
B. Measurable Indicators & Recommendations

1) Standardized Grant Evaluation Criteria (Allocation Scoring): Evaluate incoming proposals using a strict 1–9 scoring matrix across 5 actionable dimensions:
a. Publication Saturation vs. Translational Gap: Penalizes over-explored, low-yield pathways; prioritizes underexplored targets.
b. Novelty in Mechanism or Concept: Evaluates cross-disease relevance and non-canonical pathways.
c. Technology & Translational Innovation: Assesses tech integration (AI, digital phenotyping, novel delivery).
d. Collaboration & Reproducibility: Requires multi-institutional rigor and open science commitments.
e. Clinical Trial / Adoption Path: Requires realistic regulatory and public health implementation milestones.

8. Other Areas: N/A

Description: Submitting on behalf of the Alliance for Longevity Initiatives in response to NIH RFI NOT-OD-26-087. Our comments offer concrete policy recommendations to modernize NIH funding incentives, eliminate structural biases and siloed research for understanding the complexity of human health and disease. Our goal is to align NIH’s mission with more novel funding mechanisms to support next generation scientists, policymakers, and entrepreneurs so they can scale real-world outcomes.

241. Youyou Wu (1), Yang Yang (2), Brian Uzzi (3)

I am responding: On behalf of myself

Name: Youyou Wu (1), Yang Yang (2), Brian Uzzi (3)

Name of Organization: (1) University College London, (2) University of Notre Dame, (3) Northwestern University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: • Replication findings must be discoverable to deliver value.
The value of replication research depends on its discoverability. Replication evidence that cannot be found, by researchers, systematic reviewers, knowledge curators, or increasingly by AI systems, cannot correct the record or inform future work. Our work shows there's a gap in surfacing replication information across multiple knowledge channels. At the same time, AI shows promise as a new route for bringing replication evidence to users even without being explicitly prompted to look for it.

We propose following metrics to measure the discoverability and practical impact of replication studies:

• Retrievability rate for replication studies
The "retrievability rate" simply asks: if you go looking for the replication evidence, can you actually find it? Concretely, it would be measured by taking a sample of original findings subject to NIH-funded replication attempts, searching for each one the way an ordinary user would (search the claim or the study in Google, Google scholar, and AI assistants), and checking whether the replication evidence surfaces in the results. The retrievability rate is the percentage of cases where it does. If NIH funds 100 replications and only 30 can be found this way a year later, the rate is 30%, and NIH knows 70% of its investment is effectively invisible.

• Knowledge correction rate for *failed* replications:
For failed replications, it is especially important to track whether new replication evidence is incorporated into both academic and public knowledge.
Does citation to the original finding decline, in academic, policy, and media publications?
Are replication findings always co-cited with the original finding, in academic, policy, and media publications?
For live public knowledge channels (wikipedia, textbooks and AI etc.) that mention the original finding, what % updates in response to new replication information?
Detailed methods are provided in the attached paper of ours.
Additional information:
• Simple mechanisms for make replication discoverable: broad internet presence
We found that replication records presented as simple text on a single webpage (e.g., FORRT Reversals & Replications) are much more likely to be surfaced than those available only through complex databases

or journal articles. Such webpages are easier for search engines and AI systems to index, increasing the chances that replication evidence will eventually reach relevant researchers, practitioners, and educators. By extension, even researchers simply posting their replication findings on a personal webpage may make them more visible than leaving them buried in databases or journal publications. In addition, large-scale replication projects (e.g., the Reproducibility Project: Psychology) that test multiple phenomena are especially vulnerable to discoverability failure. Individual replication outcomes are subsumed within aggregate statistics and a single publication, making them invisible to anyone searching for evidence on a specific original finding.

• Action: funders should require dissemination beyond journal publication
NIH should mandate and fund dissemination for replication studies it supports. Grantees conducting replications should be required to disseminate findings beyond journal publications. Preferred methods include publishing accessible summaries on project websites and depositing results in curated replication databases like FORRT. NIH should also require that multi-study replication projects report each constituent replication as individual records (e.g., one entry per original claim), so that evidence about any given finding can be located directly.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/draft_public_perception_of_replication_outcomes_watermarked.pdf

Description: associated research paper

Title: Are replication failures reflected in public knowledge?
Authors: Wu Youyou, Yang Yang, Brian Uzzi
Keywords*:* replicability, science communication, AI for science

Abstract:
Behavioural science now routinely replicates and verifies published findings. Replication is a powerful mechanism for science to self-correct, but it is only the first step. New evidence must also reach researchers, students, and the public and update their understanding. Within academic scholarship, there is some evidence that failed replications have led to reduced citations of the original studies. Far less is known about how many non-replicable findings continue to circulate through mainstream channels through which students and the public access knowledge. This study tracked research findings with well-established replication failures and assessed whether three public knowledge channels revised their content in response. We also queried two AI platforms about the original findings to assess whether their responses reflected new replication evidence. Traditional channels including Wikipedia, textbooks, and university course syllabi, acknowledged replication failures roughly one-third of the time, which was higher than researchers' predictions in a preregistered survey. AI systems performed substantially better: ChatGPT acknowledged replication failures for up to 80% of findings, rising to 95% among findings with high online visibility. We also discovered that such knowledge revision was more likely to occur when replication records appeared on a simple, openly accessible webpage, and less likely when buried within large-scale replication projects. These findings suggest that scientific self correction depends not only on producing replication evidence, but also on making that evidence easy to find. We recommend that researchers and funders invest in dissemination to help replication evidence reach the public.

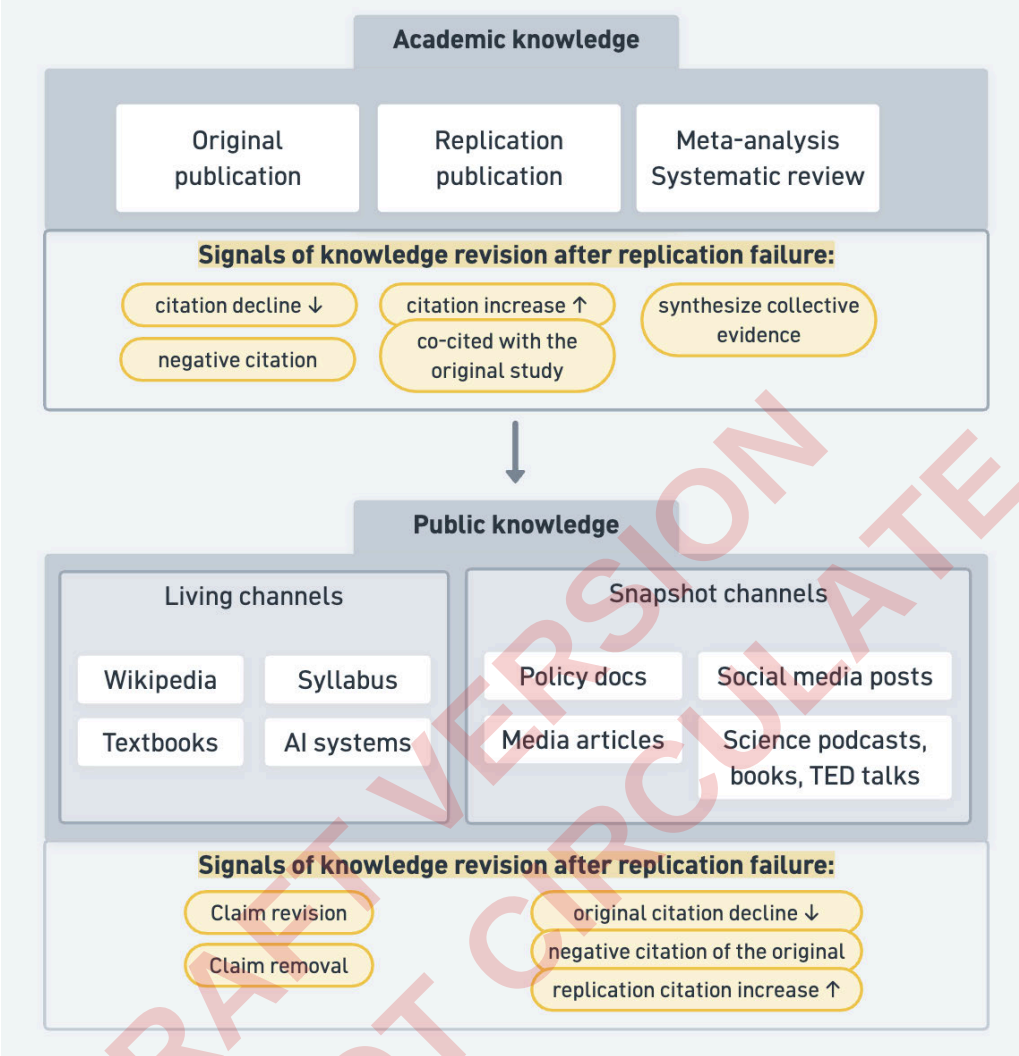
DRAFT VERSION
DO NOT CIRCULATE

Introduction

More than a decade ago, *Science* published a landmark study showing that fewer than half of 100 psychology findings replicated (OSC, 2015). The result sparked widespread attention to the replicability of science and inspired replication efforts across disciplines. Most recently, *Nature* published a series of papers examining the replicability and robustness of 3,900 studies across the social and behavioural sciences. They again found that roughly half of findings failed to replicate (Tyner et al., 2026).

Yet conducting replications is only a starting point of science's self-correction process. Replication findings must then ripple downstream, updating academic discourse and ultimately shaping public understanding. Figure 1 illustrates how knowledge revision should theoretically occur following a replication failure, and how the outcome can be empirically observed.

Figure 1. The knowledge revision pipeline: how replication failures can propagate through academic and public knowledge channels



DRAFT VERSION DO NOT CIRCULATE

Within academic d scourse, knowledge revision may be signalled by declining citations of the original paper, increased cita ons f the replication paper, or co-citation of both. Several studies have directly examined the citation impact of replication failures (e.g., Clark et al., 2023). The effects vary by analyt cal approach, ranging from modest citation declines following a replication failure to no detectable change.

Beyond academia, new replication evidence should also update what the public and students learn. We distinguish between two types of public knowledge channels: snapshot channels (e.g., media reports, policy documents, and social media posts) capture information at a single point in time and are rarely revised; living channels (e.g., Wikipedia, textbooks, and AI systems) can be revised as new evidence emerges.

To our knowledge, no study has examined whether replication failures propagate across public knowledge channels. We aim to address that gap. In our view, a robust replication failure does not automatically invalidate the original study, but it should be acknowledged and incorporated into accounts of the topic.

The current paper will focus on living channels like Wikipedia because they provide a clear, binary signal of knowledge revision. For each mention of the original study, we measure whether the replication failure has been acknowledged or not. We did not focus on snapshot channels such as media reports because of data and methodological limitations. Since individual records are rarely revised, knowledge revision must instead be inferred from citation patterns over time (e.g., Clark et al., 2023). This approach relies on sustained attention to the original paper, which is typical in academic literature but rare in public knowledge channels. Supplementary Text 1 provides a more detailed explanation.

Methods

Our sample of replication records came from the Framework for Open and Reproducible Research Training (FORRT, Hartmann et al., 2025). The FORRT Replication Databa e (FReD) is among the most comprehensive repositories of original papers and their replication outcomes in the social, cognitive, and behavioural sciences. Vo unteers continuously recorded references, phenomenon descriptions, and effect sizes for both original and replication papers. We drew additional replication records from a predecessor project “FORRT Replications and Reversals”.

We restricted our sample to 153 original papers with documented replication failures based on four criteria: (1) replications must have been published in a peer-reviewed journal; (2) the original study had not been retracted, ruling out retraction as an alternative explanation for knowledge revision; (3) replication studies had sample sizes at least 2.5 times that of the original study, a commonly accepted threshold for adequate statistical power (Simonsohn, 2015); (4) all available replication attempts were unsuccessful, as determined by the replication authors. These criteria were designed to isolate robust, unambiguous replication failures that clearly warrant revision of the original scientific record. Further methodological details can be found in Supplementary Text 2.

For each original paper, we tracked its presence and assessed knowledge revision across three traditional channels: Wikipedia, textbooks, and university course syllabi. We further queried AI systems as a fourth, emerging knowledge channel. Complete methodological details can be found in Supplementary Text 3.

Wikipedia has long been a primary gateway to science for non-experts. Altmetric Explorer was used to identify Wikipedia pages that mention any of the original papers in our sample. We then manually examined 124 Wikipedia mentions for evidence of knowledge revision (yes / no). Knowledge revision was defined as the removal of claims based on the original finding, or any mention of replication failure or controversy. We adopted this generous knowledge revision criterion because partial updates, such as a brief note of controversy, still represent meaningful engagement with the replication results. We excluded cases where original papers were not referenced in relation to the replicated finding, or appeared merely as bibliographic records such as on the author’s profile page.

DRAFT VERSION
DO NOT CIRCULATE

Textbooks translate scientific knowledge into educational canon. We examined 16 introductory social psychology textbooks for knowledge revision, drawn from a list recommended by the Social Psychology Network (2025) for teaching purposes. We focused on social psychology given its prominence in the replication crisis (Lakens, 2025) and our domain expertise in evaluating knowledge revision. For each textbook, the most current edition was compared with a previous edition to determine knowledge revision. Knowledge revision was again defined as the removal of claims based on the original paper, or any mention of replication failure or controversy. We evaluated 73 textbook mentions and excluded cases where original papers were not referenced in relation to the replicated finding (e.g., cited only for its methods or definitions).

Course syllabi shape the scientific diet of the next generation. S llabi data were retrieved from Open Syllabus. We identified 116 syllabi that assigned original papers that later f iled replication. For each syllabus, we tracked whether the assigned paper was subsequently removed following the replication failures. Because Open Syllabus data do not c ntain full syllabus content, we cannot assess the context of each paper assignment. A paper may be removed for reasons unrelated to replication failu e, or retained but taught critically. These two possibilities likely offset each other, resulting in a middle estimate

AI systems now mediate information consumption for billions. We therefore included AI as a fourth knowledge channel to assess knowledge revision, focusing on two popular platforms. ChatGPT is the most widely used conversational AI platform, with 1 billion active users in June 2026. Google AI Overview have 2 billion monthly users globally. We investigated Google AI Overview also because of its passive delivery model: AI-generated summaries appear automatically atop search results. Sixty percent of U.S. adults say they read AI search engine summaries (Gottfried et al., 2026).

We queried both AI systems about the 153 original papers and coded their responses. We used two types of prompts for AI. The first was a “lay prompt”: “Tell me about the research on [summary of the finding].” Summaries of the findings were provided by FORRT volunteers, typically written as a phenomenon name followed by a brief description (e.g., “Elderly priming; hearing about old age makes people walk slower.”). Where summaries were absent, we constructed them in the same format. The second was an “expert prompt”: “Tell me about this research: [the APA reference of the paper].” Knowledge revision was defined as any mention of replication failure or related controversy in the AI-generated response.

Our deliberately minimal prompting strategy was intended to approximate real-world use rather than optimise for detection of replication information. Critically, we did not explicitly prompt the AI to discuss replication, and each query was conducted in a fresh session without prior context.

DRAFT VERSION
DO NOT CIRCULATE

Results

Knowledge revision rate following replication failures

Figure 2 summarises the knowledge revision rates for the four knowledge channels. Among the traditional channels, Wikipedia displayed the highest knowledge revision rate (36%), followed by course syllabi (35%) and textbooks (27%). Supplementary Texts 3.1 to 3.3 provide further descriptive analyses, including whether revisions are concentrated among a few papers and whether revision occurred through claim removal or modification.

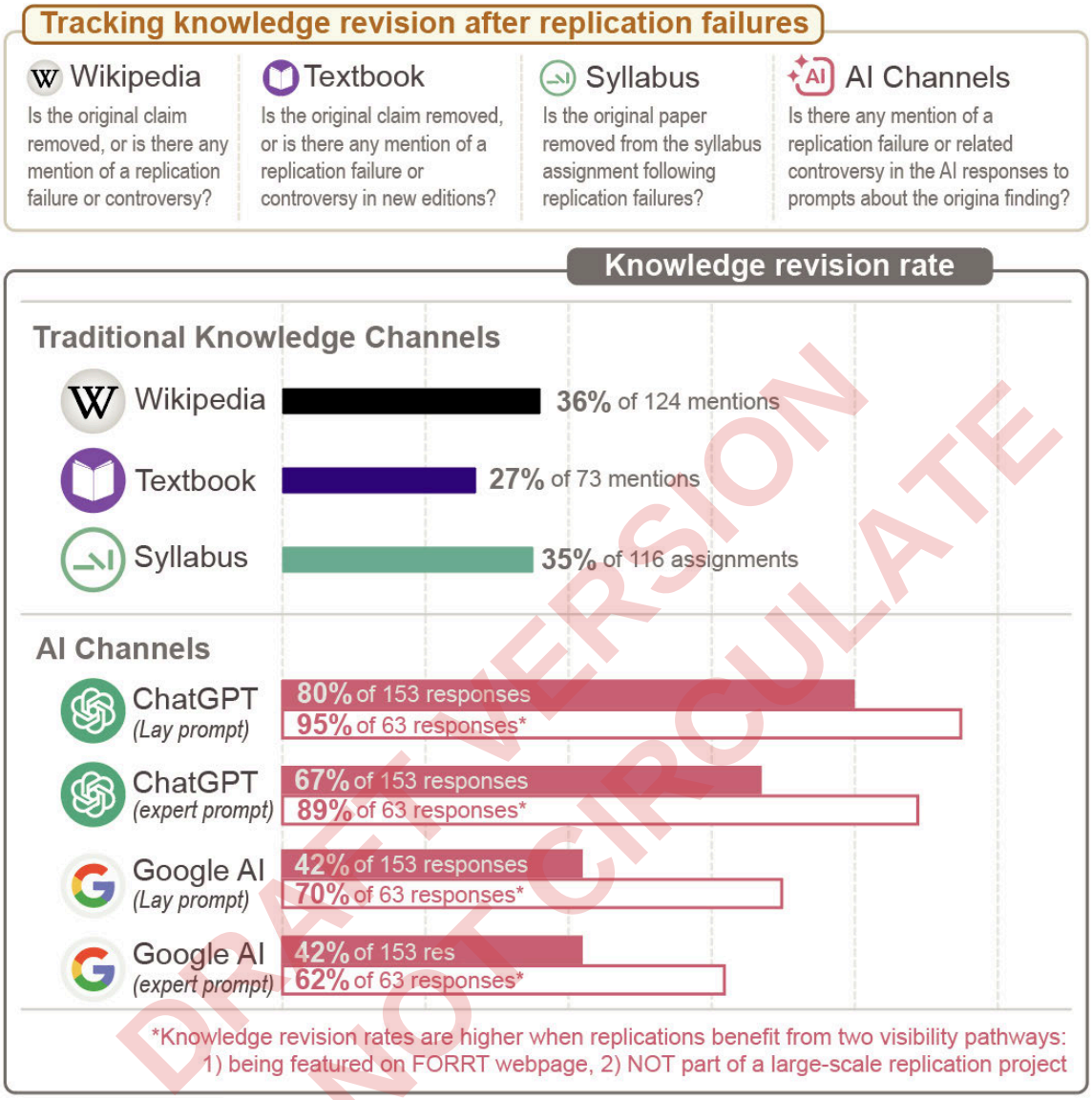
We found that these observed rates exceeded researchers' expecta ions. In a pre-registered survey, 24 attendees at the International Conference on the Science o Science and Innovation 2026 estimated knowledge revision rates for traditional channels before seeing our indings. Participants were pessimistic about knowledge revision following eplication failure (*M* = 21.0%, *Mdn* = 15.0%, *SD* = 18.2%). Supplementary Text 3.4 includes more details.

AI systems generally showed higher knowledge revision rates than traditional channels. ChatGPT outperformed Google AI Overview across both prompt ypes (lay prompt: 80% vs. 42%; expert prompt: 67% vs. 42%). We also found that ChatGPT responses are longer and more detailed than Google AI Overview. Supplementary Text 3.5 provides examples of AI responses and additional analyses including revision rate across phenomenon groups (e.g., priming effects).

The strong performance is unsurprising given AI's access to broad internet sources, including the scientific literature. When we repeated the ChatGPT queries with web search disabled, revision rates fell from 80% to 42% for lay prompts and from 67% to 39% for expert prompts. This suggests that access to external sources contributes substantially to knowledge revision.

Figure 2. How often knowledge channels revise information about original findings after replication failures.

DRAFT VERSION
DO NOT CIRCULATE



DRAFT VERSION
DO NOT CIRCULATE

Potential pathways for improving knowledge revision

We further exami ed wo pathways that may explain why some findings were more likely to attract knowledge revision than others. The first pathway is visibility via the FORRT Replications and Reversals webpage (https://forrt.org/reversals/). The website consolidates replication records on a single, openly accessible page, making them easy to find for both human readers and AI systems. According to FORRT organisers, this webpage attracts much more traffic than complex, database-format resources such as FReD. Findings featured on the webpage showed significantly higher revision rates across all AI systems (14–39 percentage points higher, see Table S2). A similar pattern appeared in Wikipedia. No significant effects emerged for textbooks or syllabi, possibly because too few of their findings (n = 11 and 12) were featured on the webpage to detect an effect. Supplementary Text 4.3 shows that the effect is unlikely to be

driven by the overrepresentation of particular phenomenon groups (e.g., priming effects) on FORRT.

A second pathway is reduced visibility when replications are embedded within large-scale replication projects (e.g., ManyLabs). These initiatives typically aim to estimate overall replication rates, with individual findings receiving little standalone attention. Consistent with this, findings from large-scale projects had significantly lower revision rates across all AI systems (34–47 percentage points lower, see Table S2). Similar patterns were observed for Wikipedia and syllabi. No significant effect emerged for textbooks, possibly because too few of the findings (n = 9) were from large-scale projects.

Combining both visibility pathways — salient web presence and individual replication attention — the 63 findings meeting both criteria achieved knowledge revision rates of 95% on ChatGPT (Figure 2). These results suggest that targeted dissemination strategies, combined with AI's ability to surface information, can potentially greatly expand the reach of replication evidence to the general public.

DRAFT VERSION
DO NOT CIRCULATE

Conclusions and policy implications

Do replication findings reach the general public? We found that knowledge revision following replication failure occurs roughly one third of the time across traditional knowledge channels: Wikipedia, textbooks, and university course syllabi. This is higher than what researchers predicted, and should encourage continued replication efforts. Yet there is still a long way to go. Our sample comprised the most robust, unambiguous replication failures available, and all of them deserve to be mentioned alongside the original finding in any knowledge channel.

AI systems represent a promising new pathway for replication evidence to reach the public. Replication records that are more visible and accessible online were associated with higher AI revision rates, indicating that targeted dissemination strategies may further close the gap.

Disseminating and consuming replication evidence

Our findings point to several concrete recommendations across different actors.

Meta-researchers should continue building infrastructure to systematically collect and curate replication records, as exemplified by FORRT. Accessibility of these resources for both human access and AI indexing is key to effective dissemination. We discovered that simple, openly accessible webpages listing individual replication records outperform complex database-format resources.

Researchers conducting replications should actively disseminate findings beyond journal publication — by contributing to replication databases, publishing accessible summaries on project websites, or sharing findings on social media. Broader web presence increases the likelihood of both knowledge curators and AI systems surfacing replication evidence.

Researchers from large-scale replication projects should ensure individual findings receive standalone attention rather than being subsumed within aggregate statistics.

Research funders should invest in dissemination as much as in replication itself, supporting both meta-scientific infrastructure such as FORRT and CurateScience (https://curatescience.org/), and individual researchers seeking to publicise their findings.

Knowledge curators, such as book authors, science journalists, Wikipedia editors, and educators, should make use of the available tools to periodically review and update their materials in light of replication evidence. One example is the FORRT Library of Reproduction and Replication Attempts (FLoRA) Annotator (https://forrt.org/annotator/), which allows users to upload manuscripts, reference lists, or reading materials to identi wh h cited studies have been replicated and what the outcomes were.

Knowledge consumers, such as readers, students, practi ioners, and policymakers, can use AI as a first point of access to replication information, w ile v rifying claims through dedicated resources such as those developed by FORRT. In our analysis, ChatGPT demonstrated better knowledge revision than Google AI Overview and is a more reliable starting point.

Finally, knowledge revision is not only a matter of noticing and adding replication information. Incorporating the new evidence into a coherent narrative is a challenge in itself. Knowledge curators such as textbook authors and Wikipedia editors must contextualise replication results alongside original findings. There is often a strong incentive to retain original narratives even when the underlying evidence is uncertain. Training and guidance for knowledge curators are needed to navigate this science communication challenge.

DRAFT VERSION
DO NOT CIRCULATE

References

Clark, C. J., Connor, P., & Isch, C. (2023). Failing to replicate predicts citation declines in psychology. *Proceedings of the National Academy of Sciences*, *120*(29), e2304862120.

Gottfried, J., Bishop, W., Anderson, M., Faverio, M., Park, E., & McClain, C. (2026). Americans and AI 2026: Chatbots, Smart Devices and Views on Impact. In *Pew Research Center Report*. https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/

Hartmann, H., Azevedo, F., Röseler, L., Wallrich, L., Aldoh, A., Elsherif, M. M., et al. (2025). Tracking and mainstreaming replications in the social, cognitive, and behavioral sciences. *MetaArXiv*.

Lakens, D. (2025). Concerns About Replicability Across Two Crises in Social Psychology. *International Review of Social Psychology, 38*(1).

Open Science Collaboration (OSC). (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569.

Social Psychology Network (2025). Social Psychology Textbooks. In Socialpsychology.org. https://www.socialpsychology.org/texts.htm

Tyner, A. H., Abatayo, A. L., Daley, M., ... Errington, T. M. (2026). Investigating the replicability of the social and behavioural sciences. *Nature*, *652*(8108), 143–150.

DRAFT VERSION
DO NOT CIRCULATE

242. Future of Cancer Care Today

I am responding: On behalf of an organization

Name: Bayli DiVita Dean

Name of Organization: Future of Cancer Care Today

Type of Organization: Research Participant/Patient Advocacy Organization/Patient Advocate/Caregiver

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: Future of Cancer Care Today (FCCT) is a non-profit that helps cancer families navigate the path beyond standard of care. Our work gives us a recurring view of a specific failure in the biomedical system: important science often exists, but patients cannot readily use it. Data and tissue may be difficult to obtain or move; relevant trials may be difficult to identify; promising technologies may encounter regulatory or manufacturing barriers; and knowledge that has been generated may never become accessible in clinical practice.

We therefore encourage NIH to measure scientific impact not only by what investigators produce, but by what their work makes possible: whether knowledge and scientific capabilities can be accessed, reused, validated, adopted, and ultimately translated into meaningful benefit for patients.

Families we work with consistently identify a recurring barrier: their records, imaging, genomic data, and tissue are difficult to move on urgent timelines, particularly when tissue preservation decisions may need to be made within days. The recurring failure is not missing science; it is that data and tissue which already exist are hard to obtain, move, combine, or reuse.

NIH should reward infrastructure and research practices that make underlying clinical, genomic, imaging, and tissue-derived data available to the individual and portable across the research and care ecosystem. Drawing on the barriers we watch families hit, this means rewarding: patient control of tumor tissue and comprehensive sequencing of it, preserved in forms that enable downstream testing rather than foreclosing it; access to the underlying data itself, including raw DNA/RNA sequencing files and exploratory diagnostics, not only a summary report; modernized privacy rules and a strong patient right of access under HIPAA; interoperability that frees data from EMRs, including action to ensure systems exchange data rather than trapping it; and patient-directed release that lets a person send data and samples to any recipient they choose, for care or for research.

The researchers, repositories, platforms, and institutions that make patient-directed sharing possible, and that honor it promptly, should be credited for it, and NIH public repositories, such as NCI's Childhood Cancer Data Initiative, should accept clinical, longitudinal data that patients can both submit and access.

Measurable indicators: whether patients can obtain usable copies of their clinical, imaging, and genomic data, including raw files; time to complete a release and completeness of the record; portability, meaning whether a patient can transmit data and tissue to another party without rebuilding the record from scratch; the share of specimens preserved in formats enabling downstream research; and documented downstream reuse in research, trial identification, or care.

Feasibility. NIH should fund shared infrastructure and standards rather than requiring each institution to build its own patient-facing system. Absent that, this indicator would reward well-resourced institutions and penalize the sites serving families least able to move their own data.

Safeguards. Patient-submitted data must be matched by rigorous consent, privacy protection, and data-quality safeguards, and incentives should reward returning value to contributors.

3. Training and Mentorship: N/A

4. Collaboration: Team-based and community-engaged science depends on labor that is invisible to author-order and citation metrics, such as sustaining patient partnerships, convening advisory boards, coordinating multi-site teams, and co-developing questions with the people affected, yet this is often what makes findings valid, adoptable, and trusted. NIH should extend a contributor taxonomy (e.g., CRediT) to name patient-advisory, recruitment, co-design, and dissemination roles; reward sustained engagement over one-time consultation; and make coordinating and integrating roles assessable so credit is not concentrated on first and last authors. Our substantive recommendations sit under Data Sharing and Translation.

5. Entrepreneurship and Translation: For patients with rare or advanced disease, the binding constraint is usually not invention but the path from a scientific rationale to responsible access. A definition of translational impact that stops at patents, licenses, and company formation misses the work that most determines whether a discovery reaches a person. We ask NIH to credit barrier-removal in the areas our families repeatedly encounter.

Faster, responsible paths to investigational therapies. NIH should credit work that improves early-phase pathways and access: allowing investigators to choose any federally compliant IRB; a notification pathway for Phase 1 safety studies where IRB approval is sufficient; Right to Try and conditional approval after Phase 1, so promising therapies are not lost; and appropriate weighting of the resulting evidence, so that data from expanded-access and Right to Try use must clear a high bar before being held against an experimental drug. Making real-time, adaptive trial infrastructure permanent, including the kind of real-time response monitoring that defines a first-principles approach, would shorten the path from question to enrollment.

Fit-for-purpose production for personalized medicine. Personalized therapeutics are frequently low-volume by nature. NIH should treat as translational impact the development of "GMP-light" manufacturing for low drug volumes, verifying the output rather than only the production process, and regulatory pathways that approve algorithms and processes, not only individual molecules, so a validated personalized method can reach the next patient without restarting approval from zero.

Reimbursement matched to rare disease. NIH should credit the development of context-driven evidence

and reimbursement models appropriate to populations too small for conventional development economics.

Policy translation as creditable work. Much of this is policy translation, generating the evidence for a change, drafting it, and building the endorsement to enact it, progressing through four stages: an evidence base; a concrete change such as regulatory or legislative language; support from affected professional, industry, and patient communities; and adoption or implementation. Investigators should receive credit for substantive contributions at any stage, weighted toward implementation and patient benefit.

Defining a "patient-first industry." NIH-supported commercialization should be judged not by whether a company formed, but by whether the resulting organization makes science materially more useful to individuals: returning usable data to patients, preserving patient control of tissue, enabling access beyond a single institution, reducing the time and cost of personalized manufacturing, supporting responsible early-access pathways, and designing evidence generation around small populations.

Applying NIH's three verbs. Incentivize documented contributions to translation-enabling change, not only commercialization. Assess them through verifiable artifacts, including published analyses, regulatory submissions, adopted language, formal endorsements, and implementation records, rather than self-report. Integrate them into the biosketch and progress report, with reviewers authorized to weigh them.

Safeguard. Company formation, patents, endorsements, and capital raised are not impact by themselves. The measure is what became possible for people, and these measures must reward responsible access, never premature access or weaker safety standards.

6. Foundational Scientific Exploration: N/A

7. Public Impact: Publication and citation counts measure influence within the research community; they do not measure whether knowledge reached, was trusted by, or benefited the people it was meant to serve. A discovery has limited public impact if patients cannot learn it exists, access the relevant trial or diagnostic, or reach the clinicians able to use it. We recommend NIH complement bibliometric measures with verifiable indicators of downstream impact: documented uptake into clinical guidelines (e.g., NCCN), care pathways, and validated patient-education resources; accessible, plain-language dissemination usable by patients and clinicians; and equity of reach into the populations most affected. NIH should capture these through a brief, structured "Public and Community Impact" statement in the biosketch and progress report, modeled on the Data Management and Sharing Plan, with explicit language authorizing reviewers to weigh it.

8. Other Areas: Organizing principle. The ultimate purpose of biomedical research is not to maximize publications, approvals, or institutional compliance; it is to improve the health and lives of individuals. In clinical settings, that means orienting the system to maximize patient survival and benefit rather than to minimize institutional or practitioner liability. NIH's impact framework should give greater weight to work as it moves closer to meaningful health benefit, while preserving appropriate measures for foundational research whose effects emerge over longer horizons.

Feasibility across career stage and institution. Any new indicator should be easy to report, independently verifiable, and interpreted in light of an investigator's resources and career stage, so that community-engaged and less-resourced investigators are not disadvantaged.

Structured evidence over automated scoring. Reward the substance of impact — knowledge that improves health and is trusted by those it serves — through structured evidence and brief narrative assessed by informed reviewers, not a new automated score.

Guard against gaming: count outcomes, not artifacts. NIH should not reward the number of datasets deposited, advisory boards convened, companies formed, whitepapers written, or endorsement letters collected. These matter only when they demonstrably contribute to reuse, adoption, implementation, improved access, or patient benefit.

We would welcome the opportunity to provide further detail, including anonymized patterns from our case portfolio.

243. Massachusetts Biotechnology Council

I am responding: On behalf of an organization

Name: Patrick Brown

Name of Organization: Massachusetts Biotechnology Council

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: An investigator whose work yields an issued patent, an executed license, a company, an IND, and a first-in-human trial has produced precisely the public value this RFI describes. Under current metrics, almost none of it registers. NIH already collects much of the data needed to change that. What is missing is the decision to treat these as measures of scientific impact rather than as compliance filings.
Recommendations
• Establish a defined set of translational indicators. We recommend: license and option agreements executed; companies formed; follow-on private capital raised by those companies; SBIR and STTR awards received; INDs and investigational device exemptions filed; clinical trials initiated; FDA designations granted, including Orphan Drug, Fast Track, Breakthrough Therapy, and Regenerative Medicine Advanced Therapy; and approved products traceable to the award. Making these indicators publicly available would itself change institutional behavior, at minimal marginal cost.
• Reduce the structural frictions facing investigator-founders. Certain current rules, such as founder equity and SBIR primary employment requirements, routinely necessitate management plans that distance the founder from the research. NIH should examine whether these rules match how academic discoveries actually reach companies.
• Strengthen SBIR and STTR as the translational bridge and measure the bridge. The most consequential failure point is between Phase II completion and the first institutional private financing.
• Recognize translational effort that does not succeed commercially. A program that fails for reasons of biology after a genuine translational attempt should not be scored below a program that never attempted translation at all.

6. Foundational Scientific Exploration: Private capital cannot fund foundational research. Venture funds operate on defined horizons and cannot underwrite ten to fifteen years of basic research work with an unknown endpoint. The reason our members can build companies at all is that NIH does the work that no investor will fund. Applying near-term translational indicators to that portfolio would be self-

defeating. It would shift NIH toward the incremental, fundable work the private sector is already positioned to do.
Recommendations
• Protect a defined share of the portfolio from any translational metric. NIH should identify basic and high-risk mechanisms as explicitly exempt, so that investigators and study sections can rely on it.
• Evaluate high-risk programs on the distribution of outcomes, not the average. A portfolio of transformative awards in which most projects fail and a few change a field is a successful portfolio.

7. Public Impact: Public impact is the hardest of these areas to measure and the most important to measure well, because it is the basis on which sustained federal investment in biomedical research is justified.
Recommendations
• Report health outcomes traceable to NIH funding: approved therapeutics, diagnostics, and devices with a documented line back to NIH-supported research; changes to clinical practice guidelines citing NIH-funded evidence; and, where data permit, patient populations reached.
• Report economic impact at the state and regional level. NIH funding supports employment, institutional research capacity, and regional cluster formation. Massachusetts is the clearest available case study of the full sequence, from NIH-funded academic discovery through spinout formation, private financing, clinical development, manufacturing, and employment. Estimates commonly place significant returns on economic activity per every dollar of NIH funding. NIH should adopt a consistent methodology, publish results by state, and update them annually.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/8.19.26-MassBio-NIH-RFI-Scientific-Impact.pdf

MASSBIO

700 Technology Square, 5th Floor
Cambridge, MA 02139
617.674.5100 | massbio.org

August 19, 2026

National Institutes of Health
Office of Science Policy
6705 Rockledge Drive, Suite 750
Bethesda, MD 20817

Re: Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

To Whom It May Concern:

The Massachusetts Biotechnology Council (MassBio) appreciates the opportunity to respond to the National Institutes of Health (NIH) Request for Information on Measuring and Rewarding Scientific Impact. MassBio represents more than 1,700 member organizations, including biotechnology and pharmaceutical companies, academic medical centers, research universities, disease foundations, and the service providers that support them. Our members include many institutions and companies that perform NIH-funded research, with Massachusetts a national leader, receiving approximately $3.5 billion in NIH funding annually, more per capita than any other state. The funding translates into new treatments for patients, high wage jobs and a growing workforce, and significant economic development across the State and the country.

We support the premise of this RFI. Publication counts and citation metrics capture only one dimension of the value NIH creates. They also render invisible much of the activity that actually moves a discovery toward a patient. Realigning NIH measures of success is a worthwhile undertaking that if implemented successfully will unlock new transformative innovations for patients across the country. We focus our comments on three of the areas identified in the notice: entrepreneurship and translation, foundational scientific exploration, and public impact.

Entrepreneurship and Translation

An investigator whose work yields an issued patent, an executed license, a company, an IND, and a first-in-human trial has produced precisely the public value this RFI describes. Under current metrics, almost none of it registers. NIH already collects much of the data needed to change that. What is missing is the decision to treat these as measures of scientific impact rather than as compliance filings.

Recommendations

- Establish a defined set of translational indicators. We recommend: license and option agreements executed; companies formed; follow-on private capital raised by those companies; SBIR and STTR awards received; INDs and investigational device exemptions filed; clinical trials initiated; FDA designations granted, including Orphan Drug, Fast Track, Breakthrough Therapy, and Regenerative Medicine Advanced Therapy; and

approved products traceable to the award. Making these indicators publicly available would itself change institutional behavior, at minimal marginal cost.

- Reduce the structural frictions facing investigator-founders. Certain current rules, such as founder equity and SBIR primary employment requirements, routinely necessitate management plans that distance the founder from the research. NIH should examine whether these rules match how academic discoveries actually reach companies.
- Strengthen SBIR and STTR as the translational bridge and measure the bridge. The most consequential failure point is between Phase II completion and the first institutional private financing.
- Recognize translational effort that does not succeed commercially. A program that fails for reasons of biology after a genuine translational attempt should not be scored below a program that never attempted translation at all.

Foundational Scientific Exploration

Private capital cannot fund foundational research. Venture funds operate on defined horizons and cannot underwrite ten to fifteen years of basic research work with an unknown endpoint. The reason our members can build companies at all is that NIH does the work that no investor will fund. Applying near-term translational indicators to that portfolio would be self-defeating. It would shift NIH toward the incremental, fundable work the private sector is already positioned to do.

Recommendations

- Protect a defined share of the portfolio from any translational metric. NIH should identify basic and high-risk mechanisms as explicitly exempt, so that investigators and study sections can rely on it.
- Evaluate high-risk programs on the distribution of outcomes, not the average. A portfolio of transformative awards in which most projects fail and a few change a field is a successful portfolio.

Public Impact

Public impact is the hardest of these areas to measure and the most important to measure well, because it is the basis on which sustained federal investment in biomedical research is justified.

Recommendations

- Report health outcomes traceable to NIH funding: approved therapeutics, diagnostics, and devices with a documented line back to NIH-supported research; changes to clinical practice guidelines citing NIH-funded evidence; and, where data permit, patient populations reached.
- Report economic impact at the state and regional level. NIH funding supports employment, institutional research capacity, and regional cluster formation. Massachusetts is the clearest available case study of the full sequence, from NIH-funded academic discovery through

spinout formation, private financing, clinical development, manufacturing, and employment. Estimates commonly place significant returns on economic activity per every dollar of NIH funding. NIH should adopt a consistent methodology, publish results by state, and update them annually.

Conclusion

MassBio appreciates NIH taking on this important effort. We would welcome the opportunity to discuss these recommendations further, and we are prepared to convene NIH staff with our member companies, academic medical centers, and investors to work through implementation and feasibility questions in more detail.

Thank you for your consideration.

Sincerely,



Kendalle Burlin O'Connell
Ceo & President
Massachusetts Biotechnology Council (MassBio)

244. Case Western Reserve University School of Medicine

I am responding: On behalf of an organization

Name: CTSC of Northern Ohio

Name of Organization: Case Western Reserve University School of Medicine

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: In the framework defined by NCATS, rigor and reproducibility determine whether research findings are dependable, transferable, and reusable. This alignment with the NIH gold-standard science document emphasizes that transparency is directly tied to the credibility and utility of biomedical research. To improve this, NIH should pivot from box-checking compliance to rewarding verifiable practices and transparent learning.
NIH should distinguish clearly between related concepts. Computational reproducibility refers to obtaining consistent results using the same data, code, and analytical procedures. Replicability refers to obtaining sufficiently consistent findings through new data collection or an independently conducted study. Generalizability concerns whether findings apply across populations, settings, models, or conditions beyond those originally examined. These distinctions should inform study-type-specific expectations rather than be collapsed into one generic rigor score. NIH should shift from documentation of intent toward evidence of practice.
We recommend that NIH adopt the Transparency and Openness Promotion (TOP) framework to distinguish between study registration, protocol availability, data and code transparency, and verification.
Key indicators of scientific rigor should include:
Time-stamped protocols, analysis plans, or preregistrations when appropriate to the research design.
Clear distinctions of confirmatory, exploratory, and post hoc analyses.
Transparent justification and reporting of sample size, randomization, blinding, inclusion and exclusion criteria, missing-data handling, and analytical decisions.
Authentication and provenance documentation for critical data, specimens, models, reagents, and software.
Availability of data, code, workflows, and computational environments sufficient to reproduce reported analyses, subject to ethical and legal constraints.
Timely reporting of null and negative results.
Independent computational reproduction, robustness testing, sensitivity analysis, or replication where scientifically warranted.
Evidence that replication findings or methodological audits led to improved protocols, standards, or subsequent research.
Concise definitions are essential for consistent review: reproducibility should refer to obtaining consistent computational results with the same data and code, while replicability refers to obtaining consistent results in new studies. NIH can operationalize this shift through study-type-specific review guidance and pilot rigor-and-transparency modules in progress reports, ensuring that the goal remains

improved practice rather than administrative burden. In addition, replication studies should be evaluated according to their methodological quality and informational value, not whether they reproduce the original finding.

2. Data, Software, and Model Sharing: NIH's 2023 Data Management and Sharing (DMS) Policy was a significant step toward universal sharing, but mandating behaviors is distinct from rewarding value. Compliance focuses on infrastructure; true incentive requires a framework that captures the downstream utility of shared resources—impact that is frequently delayed, distributed, and invisible to traditional metrics.

Data, software, and models generate scientific value only when they are interoperable, reusable, and accompanied by rich provenance. The decisive work occurs at the inception of a study, where standardized definitions, documented workflows, and validated capture procedures determine the future utility of research assets. To incentivize this, NIH must reward the quality of sharing rather than the mere presence of a repository link.

We recommend that NIH:

Incentivize Infrastructure: Officially recognize data stewardship, software engineering, and documentation as essential direct project costs.

Acknowledge Professionals: Ensure research librarians and data specialists receive named contributor credit for their roles in managing these resources.

Shift Measurement Criteria: Evaluate sharing success through indicators like persistent identifiers, versioned code, executable workflows, and clear metadata, rather than raw download counts.

Prioritize Responsible Usability: Assess sharing quality based on "responsible usability"—whether data is as open as ethically permissible with predictable access routes for qualified users—rather than penalizing necessary, clearly justified restrictions. Usability indicators may include installation and execution success, machine-readable metadata, interoperable formats, executable workflows, accessible documentation, user support, and reasonable time between a qualified request and access.

Evidence Value via Reuse: Value indicators may include independent reuse, adaptation in other settings, inclusion in external pipelines or platforms, use in replication studies, citation in research or guidance, incorporation into standards, and evidence that the resource reduced duplication, cost, or time.

Mandates alone are insufficient. NIH must cultivate an ecosystem where specialized labor, standardized infrastructure, and aligned rewards turn shared outputs into sustainable, collaborative scientific assets.

3. Training and Mentorship: The CTSA Program builds the workforce required for modern translational science. By prioritizing cross-disciplinary team science and boundary-crossing partnerships, the network aligns directly with the Translational Science Principles defined by the National Center for Advancing Translational Sciences (NCATS). These principles call for "translational scientists" who function as domain experts, process innovators, and team players—an identity that contrasts sharply with the traditional, individual-focused R01 research paradigm.

The CTSA network’s training mechanisms, such as the KL2 Mentored Career Development Award and the TL1 Linked Training Award, are specifically designed to cultivate this multi-disciplinary expertise. However, recent evidence suggests current efforts must evolve. Freedman et al. (2026) found that researchers affiliated with CTSA hubs report only slight-to-moderate confidence in translational science (TS) principles, particularly in areas like efficiency, productive failure, and risk-taking. Most participants lacked formal training in TS. Crucially, the authors conclude that training alone is insufficient without

systemic reward structures (e.g., tenure, honors) that recognize hours worked and risks taken, rather than solely counting research outputs.
To better evaluate these efforts, NIH should move beyond publication metrics and instead measure the development of core translational capabilities. The 14-item confidence scale developed by Freedman et al. (2026) offers a resource for assessing workforce development, but it must be paired with behavioral indicators of competence.
NIH should distinguish among mentoring inputs, mentoring processes, developmental outcomes, career outcomes, and ecosystem outcomes. Potential indicators include:
Individualized development plans, documented mentoring agreements, and regular review of progress.
Evidence of mentor preparation, including training in feedback, research integrity, sponsorship, and conflict management.
Demonstrated mastery of rigorous methods, team science, and data management.
Successful interdisciplinary and cross-sector collaboration.
Progress toward trainee-specific goals and career advancement.
Timely advancement, career stability, role alignment, and persistence in the biomedical or health-related workforce.
Quality of the mentoring relationship, psychological safety, clarity of expectations, and access to professional networks.
NIH should recognize instructional contributions from research professionals—including librarians, statisticians, and regulatory specialists—as valuable investments. Strong mentors and team-mentoring structures should receive grant-review credit and institutional protection. Finally, evaluation of these outcomes should be aggregated over time and interpreted contextually, ensuring that effective mentorship and workforce growth remain the priority over traditional prestige markers.

4. Collaboration: As a Clinical and Translational Science Collaborative (CTSC), we advocate for rewarding translational science principles that transcend disciplinary boundaries. While holistic health strategies—addressing the physical, mental, social, and environmental dimensions of well-being—are increasingly central to clinical care, research assessment remains largely disease-specific. The NCATS focus on cross-cutting translational barriers offers a vital alternative to the traditional NIH Institute and Center (IC) structure, yet researchers operating within this framework frequently struggle with existing grant mechanisms, particularly the R01. Study sections often favor disciplinary silos, frequently penalizing the novelty and innovation inherent in team-based, cross-disciplinary proposals.
Collaboration is the operational foundation of CTSA hubs. To better recognize this work, NIH must move beyond assessment models that obscure the contributions of essential team members. We recommend requiring structured contributor statements—utilizing the CRediT taxonomy—to acknowledge diverse roles including software engineering, data curation, statistical analysis, and project management, rather than focusing solely on PI status.

Furthermore, NIH should pilot cross-IC referral and review pathways for proposals that address shared translational barriers. Review panels for such proposals should integrate diverse scientific, operational, and community expertise. Finally, NIH should provide explicit support for research software engineers, statisticians, and project managers, while ensuring new assessment criteria do not impose undue administrative burdens on early-career investigators.
Potential collaboration indicators include:
Clarity and completeness of contributor attribution.

Evidence that the team's composition matches the scientific and translational problem.
Integration across disciplines, institutions, professions, sectors, or communities.
Shared decision-making and governance.
Inclusion of relevant methodological, operational, clinical, and lived-experience expertise at appropriate stages.
Interoperable processes, shared protocols, and effective data or resource exchange.
Resolution of scientific, operational, or regulatory barriers through collaboration.
Durability of productive partnerships when continuity is scientifically valuable.
Development of new shared infrastructure, methods, standards, or networks.
Evidence that contributions were distributed fairly and that junior or less institutionally powerful members were not rendered invisible.
Social network analysis may offer useful program-level evidence concerning the development of multidisciplinary and cross-institutional ecosystems. Measures such as connectivity, brokerage, disciplinary diversity, cross-sector reach, and participation over time can help characterize collaborative structures. However, network measures should be interpreted with caution under the guidance of network scientists.

5. Entrepreneurship and Translation: The National Center for Advancing Translational Sciences (NCATS) provides a highly comprehensive and systematically organized commercialization infrastructure within the federal biomedical research enterprise. By integrating national-scale training initiatives, such as I-Corps at NCATS, funding mechanisms including SBIR/STTR, physical bio-incubators, and the expert networks of Clinical and Translational Science Awards (CTSA) hubs, NCATS has established a powerful engine for translating discoveries into real-world health solutions. This structural strength is derived not from isolated programs, but from a coordinated national ecosystem governed by the 2025–2030 Strategic Plan.
The CTSA network offers a distinct value proposition that differentiates it from elite university Technology Transfer Offices (TTOs). Rather than replicating the high-scale venture capital integration characteristic of individual institutions like Stanford or MIT, the CTSA network provides a standardized and nationally coordinated commercialization infrastructure. These systems demonstrate maximum effectiveness when operating collaboratively, as evidenced by successful models at Northwestern, Minnesota, Virginia, and Washington.
While translation and entrepreneurship represent vital pathways to enhanced health outcomes, patents, licenses, and company formation constitute intermediate milestones rather than definitive evidence of public benefit. CTSA hubs assist investigators in defining unmet needs, testing feasibility, navigating regulatory pathways, and engaging end-users to determine whether an innovation should advance. The NIH should reward this responsible progress through evidence-based indicators, such as the validation of unmet needs; meaningful patient and stakeholder involvement; reproducible prototypes; independent technical validation; pre-specified go/no-go milestones; regulatory and clinical benchmarks; manufacturing readiness; and demonstrable improvements in accessibility, affordability, or quality. Assessment measures must align precisely with the specific stage of innovation, avoiding the assumption that every discovery warrants commercialization. Furthermore, the NIH should recognize the specialized expertise of translational professionals in regulatory science, implementation, manufacturing, and partnership development. We recommend that the NIH enhance the visibility of this work through milestone-based supplements, proof-of-concept awards, and increased credit in biosketches and grant renewals. Finally, transparency in 'negative' de-risking results—which prevent investment in unsafe or

ineffective products—should be celebrated as valuable scientific learning. Assessment frameworks must prioritize evidence of validity, utility, and responsible deployment over mere revenue or valuation, while ensuring rigorous conflict-of-interest management and the protection of research participants.

6. Foundational Scientific Exploration: The CTSA network provides vital cross-cutting infrastructure through entities such as the Trial Innovation Network and the SMART IRB platform to coordinate multi-institutional challenges and support research transcending individual Institute and Center missions. Formalizing this framework at the NIH level would provide the resources necessary to institutionalize these critical translational activities.
NIH should evaluate foundational research through expert narrative review emphasizing:
The importance of the scientific question.
The quality and transparency of the underlying reasoning.
Methodological rigor and the appropriateness of the experimental or theoretical approach.
The degree to which the work opens a new research direction or challenges a consequential assumption.
Creation of new methods, models, resources, or conceptual frameworks.
Informative negative or contradictory findings.
Evidence that the work changed subsequent scientific questions, methods, or investment.
Long-term influence on fields, training, or scientific infrastructure.

Productive failure should be recognized when a well-designed study tests a consequential proposition, generates credible evidence, and changes future decisions. NIH should avoid retrospectively labeling every unsuccessful project as productive; the value must be demonstrated through the quality of the work and the learning produced.
Longer assessment horizons are essential. Some foundational contributions will not demonstrate influence within a typical three- to five-year grant cycle. NIH could permit delayed impact updates, retrospective portfolio analyses, and renewal narratives describing how previous findings altered research trajectories. Such evidence should inform organizational learning but should not create indefinite reporting requirements for individual investigators.

7. Public Impact: Public impact represents a core component of translational science, yet it is seldom attributable to an individual investigator or a single publication. Consequently, the NIH should evaluate impact through a transparent contribution pathway that links research activities to ultimate public benefit, while appropriately accounting for temporal lags, external influences, community contexts, and shared attribution. In accordance with the CTSC's emphasis on community participation, broad applicability, implementation, and enhanced health outcomes within Northern Ohio and beyond, this approach should incorporate proximal indicators. These indicators include sustained patient and community involvement in priority-setting and research design, accessible return of results, adoption of validated tools, incorporation of findings into clinical guidelines, standards, public policy, education, or research infrastructure, and prompt rectification of inaccurate claims.
Intermediate indicators may include reach, adoption, implementation fidelity, sustainability, cost, accessibility, acceptability, appropriateness, feasibility, service quality, changes in professional or organizational behavior, and equitable implementation across settings. Long-term indicators may include changes in clinical outcomes, population health, quality of life, productivity, economic activity, system performance, or public trust. Such outcomes should be attributed to a specific body of research only when the contribution claim is supported by credible evidence and competing influences are

acknowledged. Similarly, public trust should be evaluated based on transparency regarding uncertainty and conflicts, public comprehension, responsiveness to affected communities, accessibility of results, and the quality and durability of partnerships, rather than popularity, favorable media coverage, or social media activity. Furthermore, the NIH should recognize community-engagement expertise, partnership maintenance, implementation support, public communication, and accessibility efforts as budgeted and credited contributions. Community organizations and patient partners should receive appropriate compensation and formal acknowledgment in contributor statements, and reviewers must evaluate whether their engagement substantively influenced the research rather than merely verifying the existence of an advisory group.

• Other Areas.

It is recommended that the National Institutes of Health implement a structured contribution portfolio framework rather than substituting one simplistic metric for another. NIH should implement a structured contribution portfolio rather than replace one simplistic metric with another. Ethical conduct, scientific rigor, complete reporting, and applicable data sharing should serve as baseline expectations, after which broader impact could be assessed through expert review of a limited number of selected contributions, consistent with responsible research-assessment principles (Declaration on Research Assessment, 2013; Moher et al., 2020). Applicants could identify no more than five significant contributions, explain their specific roles and those of other contributors, provide persistent identifiers or other verifiable evidence where available, and present concise evidence of rigor, reuse, influence, mentorship, collaboration, translation, foundational learning, implementation, or public benefit. Eligible contributions could include peer-reviewed publications, datasets, software, methods, standards, replication studies, training programs, team leadership, research infrastructure, community partnerships, and responsible translational milestones; however, grant funding itself should not be treated as evidence of impact. Evaluation criteria should be published in advance, tailored to the type of research, and interpreted in the context of career stage, discipline, institutional mission, available resources, and the time required for impact to emerge. To reduce administrative burden, NIH should draw evidence from repositories, ORCID records, and existing progress reports wherever feasible, while providing reviewers with appropriate training, evaluation anchors, and calibration examples. No applicant should be expected to demonstrate excellence in every category. Before widespread adoption, NIH should pilot the framework across multiple Institutes and Centers, funding mechanisms, career stages, and institution types. The CTSA network could provide a useful pilot environment because it integrates disciplines, institutions, infrastructure, workforce development, community engagement, and translation. Evaluation of the pilot should examine reviewer agreement, administrative burden, differential effects across groups and institutions, and whether the framework improves research practices and funding decisions. NIH should maintain transparency by publishing its rubrics, pilot protocols, aggregate findings, audits, and revisions while monitoring for unintended consequences, including inflated contributor claims, disadvantages to fields with fewer repositories or longer impact horizons, uncompensated documentation work, and the conversion of qualitative judgment into a hidden numerical score. Indicators that do not improve decisions or scientific behavior should be revised or retired. Ultimately, NIH should reward the practices, people, and infrastructure that make scientific findings trustworthy and useful without creating another prestige hierarchy or burdensome compliance exercise.

8. Other Areas: On behalf of the Clinical and Translational Science Collaborative (CTSC) of Northern Ohio at Case Western Reserve University (CWRU), we respectfully submit this formal response to the National Institutes of Health (NIH) Request for Information regarding Measuring and Rewarding Scientific Impact (NOT-OD-26-087). As a designated Clinical and Translational Science Award (CTSA) hub with nearly 19 years of experience fostering clinical and translational research across a seven-institution consortium, we appreciate this essential dialogue and provide our perspective as an organization deeply dedicated to the core values the NIH intends to enhance and incentivize.

The recent inauguration of the agency-wide Replication and Reproducibility Initiative by the NIH marks a significant and necessary paradigm shift in the definition and measurement of success within the biomedical research enterprise. The mission and objectives of CTSA hubs correspond directly with the thematic interests outlined in this Request for Information. For CTSA institutions, this reflects the fundamental, daily operations of translational science: establishing robust systems and cross-sector partnerships that enhance research rigor, improve operational efficiency, expand scientific inclusivity, and maximize public health outcomes.

The CTSA Program was strategically implemented to mitigate the structural and operational impediments that decelerate the progression of scientific discoveries along the translational continuum—from initial basic observations to clinical implementation and subsequent population-level health improvements. Utilizing our extensive background as a translational science hub, our institutional dedication to advancing health outcomes, and our perspective that prospective NIH metrics must accurately evaluate both individual and collective achievements, scholarly and operational contributions, and institutional and public-facing initiatives that collectively define meaningful scientific advancement. We incorporate insights from peer-reviewed literature across the translational science spectrum and align with established standards to facilitate the institutionalization and promotion of replication and reproducibility practices throughout the biomedical research enterprise.
The proposed approach should rest on five principles.
First, scientific impact is multidimensional. No single indicator can adequately represent scientific quality, usefulness, influence, workforce development, translation, or public benefit.
Second, indicators must be aligned with the purpose, stage, and modality of the work. Appropriate evidence for a basic-science discovery differs from evidence for a dataset, clinical intervention, implementation strategy, training program, community partnership, or regulatory tool.
Third, quantitative indicators should inform rather than replace expert judgment. Responsible research-assessment frameworks consistently emphasize that metrics must be transparent, field-sensitive, contextualized, and subordinate to qualitative assessment.
Fourth, NIH should reward contributions and behaviors that improve the research ecosystem, including rigorous methods, transparent reporting, stewardship, replication, mentorship, collaboration, infrastructure development, and responsible decisions not to advance ineffective or unsafe innovations.
Fifth, NIH must evaluate its own assessment system. Any new framework should be piloted, audited for differential effects and gaming, and revised or retired when indicators create burden without improving decisions or scientific behavior.

245. American Association of Physicists in Medicine (AAPM)

I am responding: On behalf of an organization

Name: Emily Townley

Name of Organization: American Association of Physicists in Medicine (AAPM)

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Framing

Medical physics is the discipline of applying physics to medicine including but not limited to applying physical principles to medical imaging and cancer treatment. As such, medical physicists are engaged in clinical practice, research and education in these fields. They are trained to support all aspects of patient safety and efficacy in the usage of highly sophisticated technologies (MRIs, CT scanners, Radiation Therapy treatment modalities) in advancing the detection, diagnosis, assessment and treatment of disease.

As part of their training, clinical practice and research practice, medical physicists perform the actions that make biomedical measurement trustworthy, shareable, and translatable. Making measurements reproducible - through standardized test objects, standardized methods, and shared reference data and tools - is among the field's highest-value contributions, and it is precisely the kind of work that conventional impact metrics systematically undercount.

Areas ripe for replication and reproducibility efforts

Some specific relevant examples are that medical physicists commonly perform measurements on equipment using standardized test objects, frequently referred to as phantoms, using a standardized set of test procedures. These tests are used to assess equipment performance that ranges from meeting a manufacturer's specification during acceptance testing, to performing tests to characterize the performance under a wide range of clinical conditions (for example, for pediatric vs. adult patients).

Thus, medical physicists are well versed in developing and using standardized test objects and standardized testing methods, both of which are strong prerequisites for replication and reproducibility efforts. In addition, many medical physicists have been instrumental in other efforts which are key to replication and reproducibility efforts including, but not limited to: (a) developing innovative test objects (physical phantoms), (b) digital reference objects which serve as test objects that can be used to evaluate algorithm performance in specific ways; (c) customized analysis tools (e.g. software tools) that they have released as open source tools to colleagues and to the public; (d) publicly available datasets including, but not limited to: (i) the Lung Imaging Database Consortium (LIDC) dataset hosted on NCI's TCIA database; (ii) more recently the Medical Imaging and Data Resource Center (MIDRC)

(https://data.midrc.org) and many others hosted on TCIA or public websites and repositories such as Zenodo; and (e) references and/or open source analysis tools to facilitate the replication and reproducibility evaluations (an example of this is the MIDRC MetricTree https://www.midrc.org/performance-metrics-decision-tree, which the FDA has adopted as one of its regulatory science tools https://cdrh-rst.fda.gov/mic-met-tree-decision-tree-medical-imaging-aiml-classification-metrics).

New strategies for the AI era

AI and machine-learning models used in biomedical science raise a distinct reproducibility challenge. Many now operate in settings where a clinician must act on a model's output quickly, such as AI-assisted treatment planning or contouring, leaving little room for after-the-fact validation. New strategies are therefore needed to establish reproducibility before and during deployment, not only at publication. These could include routine pre-deployment testing against standardized reference cases, periodic re-testing on a defined schedule, such as at initial implementation and then again at set intervals or per patient before treatment, and layered quality checks that continue to monitor a live model's performance rather than validating it only once. Sustaining this kind of ongoing verification will require dedicated infrastructure and reporting standards, including appropriate documentation of model performance, transparency, limitations, and, where applicable, explainability. NIH should also consider what dissemination and reporting incentives would make this practical for investigators to sustain, recognizing that clinical reimbursement mechanisms such as CPT codes fall outside NIH's own purview.

Incentives and infrastructure

Medical physicists have been able to find ways to carry out reproduction and replication activities in spite of a lack of incentives, but they face serious headwinds such as lack of funding, perception of a lack of innovation (typically required for publication and certainly for grant funding) and difficulty in sharing patient data (sharing phantom data and digital reference objects are much easier, but there are still some obstacles). There are examples of studies in our field where reproducibility efforts were carried out, but many of these were funded by grant mechanisms such as U01 (NCI's Quantitative Imaging Network or QIN and the aforementioned LIDC). Our journals, Medical Physics and the Journal of Applied Clinical Medical Physics (JACMP) both support a Medical Physics Dataset Article (MPDA), but these have been lightly used.

Furthermore, the recent effort of research practices to enhance reproducibility and replicability has brought an emphasis on transparency and explainability, which has fostered new attention to these issues. Practices that have been used elsewhere in medicine may be useful to consider here as well, including pre-registration of a study, transparency of data sharing, AI-models, and a persistent researcher identifier (such as ORCID), which could facilitate the use of advanced computational techniques as well as the globalization of research capabilities and collaboration.

To sustain incentives and infrastructure that would support the replication and reproduction of existing findings, it would be helpful to be able to: (a) fund research projects that would be specifically intended to support multicenter collaboration to replicate and reproduce results in a specific context and in such a

way as to support teams at various institutions to carry out the reproducibility exercise (funding both the data collection and analysis); (b) provide support for efforts to publish open source tools that would enhance these replication and reproducibility efforts, including supporting the additional expenses in making open source tools available (many journals add costs to make publications Gold OA).

Defining Causal Research

NIH's request for comment on defining causal research is also relevant to medical physics, particularly as AI and computational tools become embedded in clinical decision-making. Medical physicists routinely use physical phantoms, digital reference objects, and other known reference conditions to evaluate technologies under controlled experimental conditions. These approaches can isolate the effect of a specific technical intervention, such as a reconstruction algorithm, calibration method, or AI model, while controlling sources of variability that may confound evaluation in clinical data. Such controlled experiments can provide important evidence about cause and effects relationships for technical and computational interventions and can complement, rather than replace, causal inference from clinical and population-level data.

2. Data, Software, and Model Sharing: Framing

Medical physics is the discipline that makes biomedical measurement trustworthy, shareable, and translatable. Much of that work - validating that a result reproduces, building the tools that let others reuse a dataset, maintaining the infrastructure a whole subfield depends on - is exactly the kind of contribution that conventional impact metrics (citations, first/last authorship, grant dollars) systematically undercount. Software and data infrastructure should be recognized and funded as first-class scientific output, not treated as a byproduct of research grants aimed primarily at publications (see Whelan BM, Brock KK, Li Z, "Software from publicly funded research should be free and open source for research," Med Phys 2024; https://doi.org/10.1002/mp.17107).

Measurable Indicators NIH Could Use

• Software/tool releases as a distinct, credited output, tracked separately from publications - e.g., a persistent identifier (RRID, or DOI via Zenodo/Software Heritage) logged in RPPR/progress reports, not just mentioned in a methods section. Also, linking of DOI of peer-reviewed publication with DOI of the dataset, as on MIDRC.
• Reuse and dependency metrics, not just downloads: the number of independent labs or downstream products building on a tool. Widely adopted community platforms in imaging and radiotherapy physics (e.g., ITK, VTK, 3D Slicer, SciPy, matRad, CERR) demonstrate that reuse, not citation count, is the real signal of scientific value.
• Reproduction/replication attempts enabled: whether another lab can independently verify a published result using the released code and data. This is measurable - e.g., linked reproduction reports, containerized pipelines with recorded successful re-runs - in a way citation counts are not.
• Maintenance activity, not just initial release: issue response time, commit frequency, or a documented succession plan when a PI moves on. Release without upkeep tends to produce unusable, poorly documented, task-specific code, so any indicator scheme should reward sustained stewardship over a one-time drop.

• Documented dissemination plans and their fulfillment, since NIH grant review already asks applicants to state a dissemination plan; NIH could track completion rates against those stated plans as an accountability metric.

Implementation Approaches

• Tiered expectations rather than a uniform mandate. A strong presumption of open release fits standard research grants well, but a single rule applied uniformly across standard research grants, SBIR/STTR small-business awards, and Academic-Industry Partnership grants risks working against NIH's own commercialization goals. A "free for research use, protected for commercial use" model — similar to research-only licensing already common in biomedical technology transfer — is a better fit for SBIR/STTR and industry-partnered awards.

• Budget for it explicitly. Dissemination-quality software (documentation, packaging, a stable repository, periodic updates) is a real cost that standard research grant budgets do not currently anticipate. NIH could add an allowable, reviewable budget line for software maintenance/dissemination (complementing the open-source-tool support recommended in our Rigor and Reproducibility response section), analogous to data-management-plan costs under the 2023 Data Management and Sharing Policy, with continuation funding tied to a minimal maintenance track record, not just initial release.

• Credit infrastructure builders directly. Long-lived, well-maintained platforms require continuous, dedicated funding, not incidental grant output. A supplement or dedicated mechanism for maintaining widely reused community tools - distinct from the award that originally produced them - would recognize infrastructure maintenance as its own category of scientific labor rather than a byproduct of an unrelated project grant.

• Use existing reporting infrastructure. Since NIH already asks applicants to state a dissemination plan and factors it into review, NIH could formalize follow-through reporting (e.g., in RPPRs) rather than creating a new compliance layer.

Two Clear Examples

Example 1: the Operational Ontology for Oncology (O3)

Radiation oncology offers a working model of what professional society-led, multistakeholder data infrastructure can look like in practice. The Operational Ontology for Oncology (O3), developed under the American Association of Physicists in Medicine's Big Data Subcommittee (https://aapmbdsc.azurewebsites.net/) beginning in 2019, standardizes the clinical elements, attributes, value sets, and relationships needed to make electronic health record and treatment-planning data usable for aggregation, quality improvement, and AI-ready research (Mayo CS et al., "Operational Ontology for Oncology (O3)," Int J Radiat Oncol Biol Phys 2023; https://doi.org/10.1016/j.ijrobp.2023.05.033; PMCID PMC10741247). The current version defines 42 key elements, 359 attributes, 144 value sets, and 155 relationships, prioritized by clinical significance, likelihood of being captured in existing systems, and feasibility of workflow change.

Several features of how O3 was built and disseminated map directly onto the recommendations above:

• Multi-professional-society governance rather than a single-lab effort. O3 was developed jointly with the American Society of Radiation Oncology, Canadian Association of Radiation Oncology, Canadian Organization of Medical Physicists, European Society of Therapeutic Radiation Oncology, and NRG Oncology, with additional input from government agencies (NIH, NCI, VA) and electronic-record vendors.

This distributes both the domain expertise and the long-term stewardship burden across institutions rather than concentrating it in one grant-funded team.
• A public, iterative comment process. An open-access website was used for two public comment cycles (2021 and 2022), alongside targeted review by specialty groups, vendors, and standards bodies — a low-cost, replicable model for incorporating broad stakeholder input into a data standard without requiring every contributor to be a co-investigator.
• Interoperability with existing infrastructure, not a parallel standard. O3 was deliberately mapped to SNOMED CT, NCI Thesaurus, and HL7 FHIR, and several of its concept definitions were incorporated directly into HL7 FHIR and IHE-RO interoperability testing. This is a concrete example of the "extend and interoperate" principle that keeps a domain-specific standard from becoming an isolated silo.
• Machine-usable, freely available outputs. Beyond the ontology itself, the project released a JSON file and SQL database-creation scripts into the public domain, directly enabling other institutions to implement the standard rather than merely read about it.
• Staged rollout tied to feasibility, not an all-at-once mandate. Recommendations to clinics explicitly phase implementation — full capture of first-priority elements within one to two years for most clinics and three years for the majority — which reflects the kind of tiered, capacity-aware timeline recommended above rather than a uniform compliance deadline.
O3 also illustrates a limitation NIH should weigh: this multi-year, 30-plus-member effort was sustained largely through volunteer committee time within a professional society rather than a dedicated funding line. That model produced a high-quality standard, but it is not obviously scalable to every subfield of biomedicine, and it underscores why dedicated funding for infrastructure-building consortia — not just for the investigators who happen to originate a standard — is worth NIH's direct support.

Example 2: the Medical Imaging and Data Resource Center (MIDRC)
A complementary model comes from medical imaging data infrastructure and software tools to help understand ones' data, rather than terminology standards. The Medical Imaging and Data Resource Center (https://www.midrc.org/), initiated in 2020 and funded by the National Institute of Biomedical Imaging and Bioengineering, is co-led by the American College of Radiology, the Radiological Society of North America, and the American Association of Physicists in Medicine, and hosted at the University of Chicago on the Gen3 platform. It was created to accelerate machine-learning research by making large volumes of imaging data — originally COVID-19 chest imaging, now expanding to cancer, diabetes, chronic liver disease, coronary artery disease, and COPD — available through a common access point rather than scattered across individual institutional repositories.

Several structural choices here are directly relevant to how NIH could support data and model sharing at scale:
• A dedicated funding vehicle for infrastructure, not a research grant repurposed for it. The center is funded through an NIH contract mechanism rather than an investigator-initiated grant, with data intake, curation, de-identification, and quality assessment treated as a funded/established core function rather than an unfunded expectation placed on individual labs. This is a working example of the "distinct infrastructure funding category" recommended above, rather than a hypothetical.
• Centralized open data commons with intake with distributed, harmonized access. Imaging and meta data flow in through dedicated intake portals, is anonymized at the contributing site, is standardized centrally, and is then made available through an open data commons with a public access point for

building cohorts, sparing individual research teams from having to solve data harmonization independently for every new dataset.
• A reserved, non-public sequestered data commons. A portion of incoming data is deliberately held out from the public commons specifically to support independent validation, benchmarking, and data-science challenges. This is a concrete, exportable version of the "measurable indicator" recommended above — a held-out test set enabling objective benchmarking of models, which were trained on the shared data, rather than relying on self-reported performance; expediting translation through regulatory and to clinical care.
• A distributed biomedical imaging hub (BIH) for indexing other existing repositories to enable more comprehensive searches and discoveries for cohort building, especially useful for rare diseases where only a few cases exist at multiple institutions.
• A suite of user-friendly web-based and downloadable tools to assist AI researchers in tasks such as cohort building through stratified sampling, choosing appropriate metrics and statistical analyses for robust evaluations, determining and comparing data representativeness across datasets, evaluating both performance and presence of bias when comparing AI algorithms, deidentification tools for imaging data and radiology reports, and harmonization tools – all tools that investigators can use to expedite their research.
• A stated bridge to commercialization and regulatory translation. The held-out validation resource is explicitly framed as supporting industry translation and regulatory approval, showing that a centrally funded public data commons and a commercialization pathway are not mutually exclusive — a direct answer to the concern that data/software sharing mandates might undercut translation incentives – and can serve as a mechanism of sustainability.
• Governance that scales the scope over time. What began as a narrowly scoped, urgent-need resource (COVID-19 imaging) has since broadened its disease coverage and, through a recent philanthropic grant, its international reach, rather than being built to a fixed, one-time specification. Multiple professional-society-run subcommittees and working groups govern this expansion, echoing the multistakeholder governance structure described above.
Together, the two examples cover different but complementary layers of the same problem: one standardizes what the data means, and the other standardizes how the data is collected, held, and accessed. NIH's ability to point to concrete infrastructure funded this way is a strong signal that dedicated infrastructure investment, held out from standard research-grant budgets, is a proven vehicle rather than a speculative one.

Benefits

• Reduced duplicated effort across labs solving the same measurement, registration, or reconstruction problems.
• Stronger reproducibility posture, directly responsive to well-documented replication concerns in biomedical research more broadly.
• Increased translational reach for tools that would otherwise sit unused in a single lab, particularly benefiting smaller institutions and low-resource settings that cannot license commercial equivalents.

Unintended Consequences to Guard Against

Penalizing small businesses and industry partnerships. A uniform open-source mandate applied to SBIR/STTR or Academic-Industry Partnership grants could undercut NIH's own directive to promote commercialization, a pathway that is at least as impactful as open release for translating research into usable clinical tools.
• Repositories full of low-value code. Mandating release without funding the work needed to make code usable by others risks producing repositories filled with undocumented, task-specific scripts that add compliance burden without adding scientific value. Any incentive structure should reward usable release, not mere release.
• Slower research pace. If investigators must build to open-source, deployable standards throughout a project rather than at its end, exploratory and early-stage work may slow down. Decoupling "release standard" from "in-progress code" — e.g., requiring open availability at publication/award-closeout rather than continuously — would mitigate this.

Feasibility across Career Stages, Disciplines, and Institution Types

Early-career investigators and small/under-resourced labs face the steepest relative cost of packaging and maintaining software; funding mechanisms and review credit should explicitly account for this rather than assuming uniform capacity to comply. Standard research grants do not currently fund this work, and not every investigator has a software-engineering skill set — though this is an increasingly expected and learnable skill, not a permanent excuse for opacity.
• Small businesses and startups operating under SBIR/STTR need protection for competitively sensitive code even while participating in a "research-use" sharing norm. Enforcement of "research use only" licensing is an unresolved operational question NIH guidance should address directly, since the practical burden of proving misuse would otherwise fall on small companies least equipped to pursue it.
• Large platform-maintaining groups need a fundamentally different, continuous funding model than a typical project grant. Policy should recognize infrastructure maintenance as a distinct award category rather

3. Training and Mentorship: Framing

The mission of the AAPM is to “advance medicine through excellence in the science, education and professional practice of medical physics.” Accordingly, AAPM is committed to the education, training, and mentorship of aspiring and early-career medical physics professionals and researchers. While a number of NIH funding opportunities exist specifically to promote the growth and stability of the biomedical research workforce through training and mentorship (such as the Institutional National Research Service Award (T32) and the Ruth L. Kirschstein National Research Service Award (NRSA) Individual Predoctoral Fellowship (F31)), training and mentorship should be expected to pervade all scientific research endeavors. NIH-funded research should include an expectation of training and mentorship activities as a component of impact and success measures to signify the value of the next generation of biomedical scientists being trained alongside those researchers who are making cutting-edge advances in research today.

The typical training trajectory for medical physicists begins with a strong undergraduate physics background, followed by an accredited graduate program in medical physics. After graduate training, medical physicists contribute in several ways to biomedical science and healthcare. Some pursue

research-focused roles in academia or participate in industry innovation; some go on to a clinical physics residency, after which they will focus on clinical work (possibly combined with research activities) in a hospital setting. Consequently, medical physicists, whether employed in academia, industry, or the clinic, have been trained by their advisors and mentors in rigorous research practices, which are integral for both the advancement of science and for state-of-the-art clinical care. Imparting rigorous research skills to future generations of medical physicists is a hallmark of AAPM's mission that has resulted in a tradition of excellence in scientific research training and mentorship and, importantly, measures of success that can be translated to the biomedical research workforce at large.

Measurable indicators NIH could use

AAPM places a high value on intentional programming to foster rigor in research training for medical physicists. It has two aspects of measurable indicators that the NIH could translate for the assessment of scientific impact broadly, both for biomedical fields of study and for individual investigators providing mentorship and training.

• Undergraduate mentorship. AAPM currently sponsors two undergraduate research fellowship programs: the Summer Undergraduate Fellowship Program (SUFP) and the Driving Recruitment through Education and Mentoring Program (DREAM). AAPM conducts an annual evaluation of immediate program outcomes. It also recently began surveying past participants to measure a range of long-term outcomes, including research deliverables from the fellowship and continuation to graduate school and other training programs. Specific measures of training and mentorship success that could be translated broadly to biomedical research include the number of summer undergraduate research fellows the research team mentors through an NIH-funded research project, research deliverables within such fellowships, the number of mentored undergraduates who later enter graduate programs, and impact on pursuing research.
• Graduate and residency mentorship. The accreditation of medical physics graduate and residency programs by the Commission on Accreditation of Medical Physics Education Programs (CAMPEP) provides a structure that tracks trainees across clinical, research, and industry career destinations. This framework could be translated to biomedical science at large. For example, annual CAMPEP reporting includes number of applicants, number of trainees admitted, numbers of graduates, and subsequent positions. There are additional research training features in the CAMPEP standards: required graduate research, research-integrated residencies, responsible conduct of research training, and research ethics instruction. These metrics could all be used to quantify the impact of graduate and resident involvement in NIH-funded research.

Implementation approaches

Several implementation paths could turn the measures already collected by the field of medical physics into indicators usable across biomedical science. AAPM could publish a white paper documenting the indicators it collects, how it collects them, and the reasoning behind each, then jointly author companion pieces with stakeholders in other biomedical fields to identify which indicators transfer directly and which require adaptation. This would give the broader biomedical community a proven starting point rather than requiring each discipline to design workforce indicators from scratch.

Undergraduate research programs could collectively implement moderately frequent survey cycles (e.g., every five years) and adopt use of a persistent researcher identifier (such as Open Researcher and Contributor ID (ORCID) now commonly used to track scientific publications of individual researchers) captured at the time of the fellowship so participants can be followed across institutions and career stages without requiring formal reporting by them.

Graduate and other advanced training programs could initiate or extend existing annual reporting with a small set of standardized measures of research outcomes, such as entry into research training, research outputs produced during training, and sustained engagement in research. Other areas of biomedical science could adopt the same measures, giving NIH indicators that are consistent, standardized, and comparable across disciplines.

Finally, individual trainers and mentors in medical physics could adopt these simple measures on their own, accessing and aggregating them (such as ORCID-tracked outputs) or collecting them, providing brief reports of research training impact in research proposal applications and reports.

Benefits

Medical physics has trained researchers for scientific impact for decades and already has solid indicators that are being collected, enhancing the potential for translation to other fields. Aggregation at both the field level (in the case of undergraduate research programs and graduate/advanced training programs) and the investigator level can help inform the public about the impact that training and mentorship has in science.

Unintended consequences to guard against

It is important to note that reporting could be burdensome for groups and individuals without sufficient staffing and resources. Thus, the chosen measures should be both simple and informative to maximize the collection and reporting of the information.

Counting research deliverables could penalize programs whose main purpose is exposure to the field. For example, in short undergraduate experiences, research output is not the central goal and the time window (often 10 weeks or less) limits what a trainee can potentially produce. Therefore, measures of the impact of training and mentorship should be considered holistically. Further, trainees who move out of a given field should not be counted as attrition. Their research activities can impact other areas of biomedical science. It is also beneficial for individuals trained in rigorous research to go on to impact other sectors, such as finance, education, and non-profit organizations.

4. Collaboration: Framing

Team-based science is an area in which medical physicists make substantial contributions that are often inadequately recognized by conventional measures of scientific impact. As biomedical research becomes more convergent, the medical physicist's role is increasingly a connective one - designing the measurements, harmonizing data across sites and scanners, building and validating the shared

resources, and defining the metrics that let a large, multi-institutional effort function as a single scientific enterprise. This work is essential to whether a collaboration's results can be trusted, yet it rarely produces first- or last-author publications and is therefore largely invisible to evaluation built around authorship position, citation counts, and individual grant dollars.

Medical physics also offers concrete, working models of large-scale team science. Illustrating the value of team science, the Operational Ontology for Oncology (O3) and the Medical Imaging and Data Resource Center (MIDRC), both described in our responses to other sections of this RFI, are each sustained not by a single laboratory but by multi-society, multi-institution collaboration.

How Team Science is Actually Organized: MIDRC

MIDRC (https://www.midrc.org/) is a multi-institutional initiative co-led by three professional societies — the American College of Radiology, the Radiological Society of North America, and the American Association of Physicists in Medicine — funded by NIBIB and hosted at the University of Chicago. It coordinates imaging data from academic medical centers, community hospitals, and other contributors into a single harmonized resource, and it is governed through a structure of subcommittees, working groups, Technology Development Projects, and Collaborative Research Projects, recently expanding internationally through MIDRC-Global.
What makes MIDRC work is not principally any one discovery; it is sustained coordination across disciplines. Radiologists, medical physicists, data scientists, informaticists, statisticians, and clinicians must be aligned on how data are curated, de-identified, annotated, harmonized, and evaluated. Within that structure, medical physicists occupy a characteristically connective role, establishing harmonization and quality-assessment methods, designing the validation resources and community benchmarking (e.g., Grand Challenges) that let algorithms be compared objectively, and defining the performance and bias metrics the whole community relies on. This is high-value scientific labor that determines whether the shared resource is trustworthy, yet it maps poorly onto first/last-author metrics.

Measurable Indicators NIH could Use

• Role-based contribution taxonomies. NIH could incorporate structured contribution taxonomies, such as CRediT, into its evaluation framework to identify and credit the specific roles that convergent science depends on – including data curation and harmonization, annotation, validation-set design, standards and ontology development, quality assurance, and consortium governance - rather than relying primarily on authorship position.
• Recognition of consortium and working-group leadership as a creditable scientific contribution in biosketches and progress reports, so that stewardship of a shared standard or resource (e.g., leading a MIDRC working group or an O3 development effort) is visible and valued.
• Outputs enabled by collaboration. NIH could recognize measurable scientific products that depend on coordinated contributions across disciplines or institutions, such as harmonized multicenter datasets, shared validation resources, consensus measurement protocols, interoperable standards, or independent multisite validation. These outputs provide stronger evidence of meaningful team science than simply counting the number of participating investigators, institutions, or organizations.
• Contribution distinct from sole attribution. Indicators should let evaluators credit an essential link in a

longer collaborative chain without requiring that the contributor be able to claim the whole result - the connective, enabling contribution should be assessable on its own terms.

Implementation Approaches

• Use existing reporting infrastructure. A structured contribution taxonomy could be captured in RPPRs and biosketches rather than through a new compliance layer, letting team roles be recorded where NIH already collects investigator information.
• Credit and fund the coordinating core. Sustaining team science at scale requires supporting the coordination functions themselves — intake, harmonization, annotation, validation, and governance — as dedicated infrastructure, complementing the infrastructure-funding recommendations in our Data, Software, and Model Sharing response rather than assuming these functions emerge from participating investigators' individual project grants.
• Recognize professional societies as collaboration infrastructure. O3 and MIDRC both show professional societies serving as durable, neutral conveners that distribute domain expertise and long-term stewardship across institutions; NIH could formally recognize this convening role when assessing the feasibility and sustainability of team-based efforts.

Benefits

Recognizing team-based and connective contributions would encourage researchers, including medical physicists, to take on the essential coordinating work that convergent science depends on, rather than being disincentivized by metrics that ignore it. Such recognition would also give NIH a more accurate picture of how modern biomedical results are actually produced — in which measurement, harmonization, and validation contributions are often the difference between trustworthy and untrustworthy findings. And it would make credit more equitable for contributors whose work does not surface as first/last authorship, including early-career researchers and staff scientists who frequently perform this connective labor.

Unintended Consequences to Guard Against

• Box-checking and gaming. A contribution taxonomy could devolve into rote role-tagging; indicators should reward substantive, verifiable contribution rather than mere listing.
• Over-granular credit that dilutes meaning. Assigning credit to many small roles risks making all contributions look equivalent; the aim is to make essential connective work visible, not to flatten distinctions between central and peripheral contributions.
• Penalizing the invisible contributor twice. If team-science credit is added but evaluation still defaults to authorship position and citation counts, the contributors this reform is meant to recognize will remain disadvantaged; the indicators must carry real weight in review, not sit alongside unchanged traditional metrics.
• Equating team size with scientific value. Metrics intended to recognize collaboration should not inadvertently disadvantage rigorous research conducted by individual investigators or small teams. Collaboration should be recognized when it adds scientific value or enables work that could not otherwise be accomplished, rather than rewarded based on the number of investigators, disciplines, or institutions involved.

5. Entrepreneurship and Translation: Framing

AAPM recommends that NIH define translation broadly and evaluate it according to the pathway by which a discovery produces practical benefit. Patents, licenses, company formation, and investment can be relevant indicators, but they are neither necessary nor sufficient evidence of scientific or public impact. In medical physics, translation is not limited to starting a company, licensing intellectual property, or obtaining a patent; consequential translation often occurs through independent technical validation, clinical commissioning, regulatory science, integration into clinical workflows, incorporation into professional standards, open-source dissemination, and adoption across healthcare institutions. Translation may consist of moving a physical principle, algorithm, device, imaging method, dosimetric technique, or decision-support tool through technical validation, independent replication, regulatory evaluation, clinical commissioning, workflow integration, professional guidance or standardization, multisite adoption, and demonstrated improvement in safety, quality, efficiency, access, or outcomes.

A concrete example of translational impact that would be largely missed by conventional commercialization metrics is AAPM's TG-51 protocol for clinical reference dosimetry. Published in 1999 and maintained through subsequent addenda (2014 for photon beams, 2024 for electron beams) by a standing AAPM working group, it moved a measurement-science advance into standardized clinical practice with no patent or commercial product - and, according to IROC Houston, has been implemented by 99.8% of North American institutions with compliant machines. It exemplifies high-impact translation achieved through professional standardization and multisite adoption rather than commercialization.

Measurable Indicators NIH Could Use

Evaluation should be appropriate to the innovation's maturity. Early-stage work should not be judged by sales or clinical outcomes, while mature work should not be judged solely by publications. Stage-appropriate indicators could include:

Early-stage indicators:
- Reproducible technical performance.
- Benchmarking against accepted reference methods.
- Documentation of uncertainty and failure modes.
- Assessment of clinical or user needs.
- Engagement of patients, clinicians, and intended users.

Translational-development indicators:
- Independent or multisite validation.
- Human-factors and workflow testing.
- Regulatory milestones.
- Interoperability with clinical systems.
- Performance across populations, equipment, and institutional settings.
- Consideration of safety, cybersecurity, maintainability, and lifecycle costs.

Adoption and deployment indicators:

• Clinical commissioning or adoption.
• Inclusion in clinical trials, standards, or professional guidance.
• Improvements in quality, safety, access, cost, or efficiency.
• Uptake outside the originating institution.
• Availability of training resources and post-deployment monitoring.
• Sustainable maintenance and technical support.

This framework should recognize multiple dissemination models. An openly available algorithm, reference dataset, quality-assurance method, or consensus standard may, in some circumstances, create greater public value than a patented technology that achieves limited adoption. Some high-impact innovations are intentionally disseminated through open-source software, public datasets, standards, or nonexclusive licensing, and these should not be treated as less valuable than proprietary commercialization.

Implementation Approaches

Implementation mechanisms could include a structured translational-impact narrative in progress reports and biosketches, milestone plans appropriate to the technology's maturity, supplements supporting independent validation and clinical implementation, and access to regulatory and commercialization expertise for institutions without extensive technology-transfer infrastructure. NIH should also support protected time, entrepreneurial education, regulatory expertise, implementation science, and partnerships among academia, healthcare organizations, government, and industry. NIH should support transparent management of financial conflicts while recognizing that responsible collaboration among academia, healthcare institutions, government, and industry is often necessary for successful translation.
NIH should recognize the contributions of the full translational team, including investigators responsible for measurement science, software, validation, quality assurance, regulatory strategy, clinical workflow integration, training, and long-term technical stewardship. Metrics should recognize each individual's contribution rather than attributing all success to the patent holder, founder, or laboratory director (consistent with our Collaboration response).

Benefits

Defining translation broadly, and recognizing open dissemination alongside commercialization, increases the reach of innovations that would otherwise sit unused, particularly for smaller institutions and low-resource settings. Failed or discontinued translation attempts can also produce valuable information about safety, generalizability, usability, cost, or clinical feasibility: a rigorously conducted translational effort that identifies a safety limitation, lack of generalizability, or poor clinical fit can prevent ineffective or harmful implementation and should be regarded as a meaningful contribution. NIH should reward rigorous learning, not only successful products.

Unintended Consequences to Guard Against

NIH should avoid reliance on raw counts of patents, licenses, start-ups, or investment. These indicators

can be readily inflated and do not establish clinical usefulness, accessibility, safety, or benefit to the public; NIH should avoid rewarding patent accumulation, as patents vary substantially in quality and public value and may never lead to a usable intervention. Translational impact should instead be evaluated through converging evidence that an innovation is valid, usable, responsibly disseminated, and capable of improving health or healthcare delivery. Conflict-of-interest transparency is essential, but industry collaboration should not itself be regarded as evidence of bias or diminished scientific value.

Feasibility Across Career Stages, Disciplines, and Institution Types

Smaller institutions and investigators without extensive technology-transfer offices need equitable access to translational assistance. Stage-appropriate evaluation is itself a feasibility safeguard: judging early-stage or foundational work by sales or clinical outcomes would disadvantage exactly the investigators and institutions least able to carry a technology to market alone.

6. Foundational Scientific Exploration: Framing

Traditional metrics of scientific evaluation — citation rates, high-impact journal counts, and short-term deliverables — systematically undervalue the capital-intensive infrastructure, methodological validation, and multi-center tool generation required for high-risk, foundational biomedical innovations. Using FLASH radiotherapy (ultra-high-dose-rate irradiation) and advanced radiation dosimetry (for FLASH-RT and radiopharmaceutical therapies) as core exemplars, this response details specific non-citation indicators, implementation mechanisms, risk mitigations, and career-stage safeguards to realign NIH funding and evaluation with modern foundational scientific research. These recommendations are intended to apply to high-risk foundational research generally.

FLASH Radiotherapy as a Paradigm for Foundational Scientific Exploration

FLASH radiotherapy (FLASH-RT) delivers radiation at ultra-high dose rates, greater than 40 Gy/s, and is under active investigation for its potential to widen the therapeutic window by sparing healthy tissue while preserving tumor control. The underlying biological mechanisms, including oxygen depletion, free-radical kinetics, and mitochondrial effects, and the physical requirements for delivering and monitoring ultra-fast beams, remain incompletely understood and are the subject of ongoing preclinical and early clinical study. Progress depends on sustained investment in beamline engineering, physics infrastructure, and standardized assay protocols. This work is necessary to determine whether and how FLASH-RT can be safely translated to patients, yet it does not register on short-term citation counts. FLASH-RT illustrates why conventional metrics undervalue foundational work: its scientific value lies as much in rigorously mapping mechanism and uncertainty as in confirming a positive result.

Measurable indicators.

Specific, measurable indicators of foundational impact include
• Community technology adoption - the number of independent research and clinical centers adopting an open-source beam-delivery architecture, target chamber, or pulse-control framework.
• Standardized protocol adherence - community-wide implementation of standardized radiobiological protocols, oxygen-control standards, and beam-calibration procedures.

• Shared open-data repositories - public deposition of raw, multi-center biological and physics datasets, such as the NCI Cancer Data Service, to support independent mechanistic modeling.
• Interdisciplinary team capabilities - validated institutional infrastructure bringing together accelerator physicists, radiobiologists, computational scientists, and clinical oncologists.

Incentives and funding mechanisms

Proposed NIH incentives and funding mechanisms include milestone-driven infrastructure grants (i.e., U01 and U24 cooperative agreements and RM1 Research Opportunity Awards to support long-term physical beam-line construction, standardization, and inter-institutional calibration networks) and rewarding rigorous non-replications, establishing funding lines explicitly devoted to independent replication of the FLASH effect across variable dose rates, tissue types, and oxygenation levels, rewarding biological boundary-mapping over positive-outcome bias.

Unintended consequences

One unintended consequence should be managed directly: over-incentivizing 'high-risk' novel modalities without sufficient mechanistic rigor could lead to premature clinical translation or unvalidated biological claims. This can be mitigated by tying translational funding milestones to mandatory, independent multi-center validation of physical beam characteristics and biological endpoints prior to Phase I/II clinical-trial initiation.

Career-stage and institutional considerations

Early-career investigators who lead team-science infrastructure work, such as accelerator design or dosimetry modeling, face a particular risk under current evaluation norms, since this work is essential to a project's success but rarely generates first- or last-author publications. Structured contribution credit along the lines recommended in our Collaboration response, such as recognizing methodological leadership in biosketches and progress reports through a taxonomy like CRediT, would let these investigators document their role without relying on authorship position alone to demonstrate their contribution.

Advanced Radiation Dosimetry as an Enabling Foundational Science

Advanced radiation dosimetry provides the essential quantitative foundation for safe, effective, and reproducible research in novel modalities, including FLASH radiotherapy, MR-guided adaptive therapy, and targeted radiopharmaceutical therapeutics (RPTs) such as Lutetium-177 (177Lu) and Actinium-225 (225Ac). Because dosimetry outputs take the form of open-source software packages, Monte Carlo codes, physical phantoms, and metrology standards, rather than high-volume primary-outcome papers, traditional metrics dismiss dosimetry as a routine support service rather than a primary scientific driver.

Measurable indicators

Specific, measurable indicators of foundational impact include

• Regulatory and guideline integration - formal adoption of novel dosimetric frameworks into FDA regulatory filings, SNMMI/AAPM consensus protocols, or NCI National Clinical Trials Network (NCTN) trial designs
• Computational reach and software adoption - unique downloads, active code forks, and platform integrations of open-access dosimetry software such as MIRDcalc and voxel-level Monte Carlo tools
• Standardized calibration distribution - deployment of validated physical or digital reference phantoms across academic and community clinical-trial sites.
Proposed NIH incentives and funding mechanisms

Two mechanisms are proposed. First, dedicated methodology-infrastructure awards — funding mechanisms such as R24 or U24 — focused on open-source dosimetry-tool creation, computational-phantom expansion, and multi-center metrological traceability. Second, mandatory dosimetric rigor standards that require patient-specific or voxel-level dosimetry across all NIH-funded clinical trials investigating novel radiopharmaceuticals or ultra-high-dose-rate modalities.

Unintended consequences

The principal unintended consequence to manage is incentivizing pure software or hardware development detached from biological or clinical utility; this can be mitigated by requiring all funded dosimetry projects to demonstrate explicit integration with biological-response models, toxicity-reduction endpoints, or clinical-trial harmonization.

Career-stage and institutional considerations

Finally, with respect to institutional equity and access, open-source dosimetry suites and shared reference datasets reduce financial barriers for non-R01-intensive institutions and community medical centers, broadening diverse patient participation in advanced radiopharmaceutical clinical trials.

Conclusion

NIH should explicitly recognize technology creation, open infrastructure, consensus standards, and multi-institutional collaboration as primary indicators of scientific impact within the Foundational Scientific Exploration category. By expanding evaluation metrics beyond traditional citation counts, NIH can better sustain long-term support for high-risk foundational research of this kind.

7. Public Impact: Framing

AAPM supports NIH's effort to recognize the broader public impact of biomedical research. Public impact should encompass not only new drugs and devices, but also improvements in the safety, quality, accessibility, reliability, and efficiency of healthcare. Medical physics illustrates the importance of this broader approach, and provides an example of public impact that is important but often largely invisible to patients. Much of the field's public benefit arises from the measurement, validation, dosimetry, quality-assurance, and technical infrastructure that enables imaging and radiation therapy to be delivered safely and accurately, including safer imaging, more accurate radiation therapy, reduced unnecessary exposure, consistent quantitative measurements, earlier detection, fewer errors, better

workflows, and wider access to high-quality care. The Image Gently and Image Wisely campaigns — multidisciplinary national radiation-safety initiatives in pediatric and adult imaging - provide visible examples of how evidence generated through imaging science, radiation protection, protocol optimization, and quality improvement can ultimately be translated through professional initiatives into broad public benefit. Their impact is distributed across patients and institutions and is realized through changes in practice rather than through a single product or discovery.

In safety-focused disciplines, public impact may also consist of adverse events that do not occur: an unnecessary imaging examination avoided or optimized, a treatment error prevented, or an unsafe or ineffective technology identified before widespread clinical implementation. Such preventive impact is inherently difficult to capture through conventional outcome metrics, but it can represent substantial value to patients and the healthcare system.

Measurable Indicators NIH Could Use

Public impact could be evaluated at several levels, spanning patient, clinical, health-system, population, and societal outcomes:
• Patient - safety, outcomes, experience, and treatment burden.
• Clinical - guideline adoption, diagnostic quality, treatment accuracy, and workflow reliability.
• Health system - reduced errors, avoided procedures, efficiency, interoperability, and cost.
• Population - access, geographic reach, disparities, and scalability.
• Societal - public trust, scientific literacy, workforce development, informed public policy, and economic activity.

Concrete examples across these levels include improvements in diagnostic performance or treatment accuracy; reductions in unnecessary radiation exposure, errors, repeat procedures, or treatment burden; adoption of validated practices or standards; improved interoperability and workflow reliability; dissemination to community, rural, or other resource-constrained settings; reductions in disparities in access or performance; and evidence that research findings have informed clinical guidance, regulation, or public policy.

Outcomes should be interpreted in relation to the investigator's contribution. Process and intermediate outcomes are legitimate indicators when they are credibly connected to public benefit, and metrics should account for the potentially long interval between foundational research and widespread clinical impact.

Implementation Approaches

Assessment should combine appropriately selected indicators with a concise narrative describing the investigator's contribution, the affected population, the strength of the evidence, and the expected time horizon. NIH should permit intermediate and preventive outcomes when investigators can describe a credible pathway to public benefit. For example, a validated quality-assurance method may reduce the probability of future errors even when the errors prevented cannot be directly counted.

Several practical measures would support fair and reliable evaluation of public impact:
• Use a short menu of indicators rather than one universal score, and allow investigators to select the indicators relevant to their work.
• Require evidence or links supporting major impact claims.
• Give reviewers guidance on contribution versus attribution and on differing translational timelines.
• Keep reporting concise and reuse information already collected in biosketches and progress reports.
• Recognize maintenance, replication, negative findings, and prevention of harm.
• Avoid composite rankings or formulas that imply false precision, and periodically assess whether particular indicators are encouraging gaming or distorting research priorities.

Evaluators should distinguish contribution from sole attribution: individual researchers often provide an essential link in a longer chain of discovery, validation, implementation, and clinical care, but cannot independently determine final health outcomes. (This principle recurs across our responses – see our Collaboration and Entrepreneurship and Translation responses - and applies with equal force to public-impact assessment.)

Public trust should be supported through transparent communication of benefits, limitations, uncertainty, conflicts of interest, and the performance of emerging technologies such as artificial intelligence. Accessibility should mean more than publishing a paper; it can include usable documentation, interoperability, affordability, training, disability access, language access where relevant, and feasibility in community and rural settings. Patient and community engagement should be recognized when it substantively influences research questions, study design, communication, dissemination, implementation, or access — not simply by counting the number of meetings held.

Benefits

A broad, multi-level view of public impact captures how biomedical science actually delivers durable value: not only through new products, but through safer, more accurate, more accessible, and more reliable care. Population-level safety initiatives (such as Image Gently and Image Wisely) exemplify this durable value, the benefit accrues broadly across many patients and institutions rather than through a single marketable product. Recognizing preventive and negative findings - for example, research demonstrating that a technology is unsafe, biased, non-generalizable, or unsuitable for clinical use - protects patients and conserves public resources, and represents genuine public value even when it does not lead to product adoption. A framework that credits enabling and infrastructure contributions ensures that the largely invisible work behind safe, effective imaging and radiation therapy is neither overlooked nor undervalued.

Unintended Consequences to Guard Against

NIH should avoid requiring investigators to claim direct changes in mortality, economic return, or other distal outcomes that cannot be reliably attributed to one project. Over-reliance on simplistic "lives saved" or economic-return calculations invites over-claiming and disadvantages foundational, enabling, safety-focused, and infrastructure-oriented research. A portfolio of contextualized evidence would better capture how biomedical science delivers durable value to the public. Because a new framework could

itself introduce distortions, NIH could pilot it in selected programs before applying it broadly, and evaluate it for bias by discipline, career stage, institution type, and access to commercialization infrastructure.

8. Other Areas: The American Association of Physicists in Medicine (AAPM) is the professional society of medical physicists — scientists who apply physics to medicine to advance the safe, accurate, and effective use of imaging and radiation therapy in patient care and research. On behalf of our members, we thank NIH for its leadership in examining how scientific impact is measured and rewarded, and we welcome the opportunity to respond to NOT-OD-26-087. Medical physics sits at the intersection of measurement science, technology development, and clinical translation, and our members contribute to biomedical research in ways - reproducibility infrastructure, data and software sharing, standardization, training, and team-based enabling work - that are essential to rigorous science yet often undervalued by conventional impact metrics. We offer the following comments in that spirit and stand ready to serve as a continuing resource to NIH as this important effort moves forward.

Description: The American Association of Physicists in Medicine (AAPM) is the professional society of medical physicists — scientists who apply physics to medicine to advance the safe, accurate, and effective use of imaging and radiation therapy in patient care and research. On behalf of our members, we thank NIH for its leadership in examining how scientific impact is measured and rewarded, and we welcome the opportunity to respond to NOT-OD-26-087. Medical physics sits at the intersection of measurement science, technology development, and clinical translation, and our members contribute to biomedical research in ways - reproducibility infrastructure, data and software sharing, standardization, training, and team-based enabling work - that are essential to rigorous science yet often undervalued by conventional impact metrics. We offer the following comments in that spirit and stand ready to serve as a continuing resource to NIH as this important effort moves forward.

246. Inter-university Consortium for Political and Social Research (ICPSR)

I am responding: On behalf of an organization

Name: Elizabeth Moss, with Maggie Levenstein, and Trent Alexander

Name of Organization: Inter-university Consortium for Political and Social Research (ICPSR)

Type of Organization: Other

Type of Organization - Other: Research Data Repository

Role: Other

Role - Other: Librarian (Moss); Investigator/Researcher (Levenstein and Alexander)

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: The Inter-university Consortium for Political and Social Research (ICPSR) appreciates the opportunity to respond to NIH's Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087). ICPSR is one of the world's largest digital repositories of social and behavioral science research data. As a replication-ready repository, ICPSR preserves data, documentation, and provenance metadata, with the technical and legal infrastructure to support sharing, including for data that are large, complex, or sensitive. Certified by CoreTrustSeal (ICPSR, 2024), ICPSR is an archive dedicated to providing good data stewardship that maximizes the value of federal investments, ensures that scientists are credited for creating data products, supports replication of original research, and creates new opportunities for research. ICPSR currently preserves and disseminates approximately 27,000 data studies containing hundreds of thousands of individual data files, with over 850,000 downloads each year. As a strong supporter of replication in biomedical, behavioral, and social science research, ICPSR hosts more than 7,000 research replication packages in journal repositories on its self-publishing platforms. Additionally, for 63 years, successive generations of early-career scholars have received training in quantitative research methodology through courses in the ICPSR Summer Program, where they are exposed to best practices in research. Further, ICPSR promotes transparency and replicability by publishing extensive guidelines for data depositors on how to prepare data files and documentation to ensure their reuse (ICPSR, 2020). ICPSR strongly supports NIH's interest in identifying the "critical research and infrastructure needs necessary to advance rigorous and reliable scientific findings" (NIH, 2026). ICPSR further supports NIH's efforts to expand measures of scientific success through acknowledging, elevating, and rewarding scientists who share data. In tandem with NIH, ICPSR also recognizes the value in rewarding scholars who reuse shared data and then make adequate information and tools available to replicate their results.

In addition to serving as data stewards and supporting study replication, ICPSR has a long history of linking publication-associated replication materials to the curated data it distributes. Data distributed by ICPSR form the foundation for tens of thousands of research articles, reports, dissertations, books, and other works that advance science, and are collected in the ICPSR Bibliography of Data-related Literature (ICPSR, 2026), a one-of-a-kind searchable database that captures, organizes, and makes discoverable research findings stemming from ICPSR datasets. Each citation record in the database is enhanced with

metadata linking the publication to the underlying data used by its author. The Bibliography plays a crucial role in demonstrating the long-term impact of NIH-funded datasets. However, a further step could be taken to recognize authors who reuse data and make it easier to test their methods and reproduce their findings.

Data repositories are well positioned to play a larger role in recognizing investigators who reuse data and provide replication packages when they publish their findings, since the platforms on which shared data are made available form a natural home for source data, published findings based on those data, and the tools essential to replicating study findings. Linking such resources together makes each meaningful in context, multiplies the usefulness and impact of NIH investments, and helps recognize investigators' rigorous efforts to make results from publicly funded research reproducible.

Measurable indicators of science that should be recognized and incentivized:
--Researchers publish in journals or on platforms that assign scholarly works with persistent identifiers, e.g., DOIs, to ensure long-term accessibility.
--Researchers reuse archived data and include links to their replication packages that contain their analytic code, software, and other tools essential for replication.
--Data repositories flag or highlight publications in which authors include links to their replication packages.
--Data repositories encourage authors to deposit their replication packages, which are linked to the original dataset they reused.
--Data repositories that house replication packages requiring restricted access offer a streamlined process for granting access for testing replicability.
--Data repositories systematically collect and link their distributed data to the citing literature, and they also link findings, replication packages, and original archived data to disseminate them together, thereby simplifying and promoting the reproduction of results.

Implementation approaches:
--Trusted repositories should provide a data citation and an accompanying persistent data identifier for all deposited datasets, and require data reusers to cite the data in publications in which the data were analyzed.
--Publishers should require the explicit citation of source data and replication packages in scholarly works. Without explicit citation and persistent identifiers, automatic linking among data, literature, and replication packages is more difficult.
--NIH should support archival system infrastructure that links research data to the primary and secondary findings generated from those data.
--NIH should support efforts by archives and repositories to single out publications that provide access to replication packages and use their platforms to promote best practices.
--Investigators who have received recognition from trusted data repositories can include such accomplishments in their NIH grant applications and progress reports.
--NIH should encourage archives not only to collect links to data-related publications but also to find and link to, or even archive, the replication packages associated with those publications.
--NIH should encourage archives and repositories to reach out to authors whose replication materials are not available online but are available by request, and encourage authors to preserve them.

--NIH should support efforts by archives and repositories to disseminate data with the citing papers and the replication packages associated with them.

Potential benefits or unintended consequences:
--Authors may more frequently make use of provided citations and unique identifiers to formally cite data.
--It becomes easier to identify which shared datasets support which publications, analyses, and downstream discoveries.
--It becomes easier for researchers to build upon existing research, since the citing literature is identified and available for replication or extension.
--Transparency and replicability are prioritized by making replication materials available alongside the data.
--More authors will provide links to replication packages when publishing results.
--By collecting publications that reuse data and including replication packages, archives and repositories will be able to identify investigators who enable replication.
--Greater public trust in, and a stronger evidence base for, biomedical and behavioral science will result.
--A potential unintended consequence is that creating viable replication packages when restricted data were used can be difficult due to privacy concerns, indirectly penalizing researchers who reuse those data. Archives could mitigate this by providing secure options for replication testing. Also, by highlighting publications whose authors provided replication packages, repositories risk endorsing the content of their analysis rather than the practice of sharing the necessary tools to replicate the experiment or study using the same methods and conditions to achieve the same result.

References
Inter-university Consortium for Political and Social Research (ICPSR). (2026). Data-related publications. Retrieved August 17, 2026, from https://www.icpsr.umich.edu/sites/icpsr/find-data/citations

ICPSR. (2024). "2027-08-22 - Inter-university Consortium for Political and Social Research (ICPSR) - CoreTrustSeal Requirements 2023-2025. Retrieved August 8, 2026 from https://doi.org/10.34894/WCWZU0

ICPSR. (2020). Guide to social science data preparation and archiving best practice throughout the data life cycle, 6th Edition. https://wpvip.icpsr.umich.edu/icpsr/wp-content/uploads/sites/11/dataprep-wcag.pdf

National Institutes of Health (NIH). (2026). Enhancing scientific rigor, transparency and replicability. Retrieved August 17, 2026 from https://grants.nih.gov/funding/find-a-fit-for-your-research/highlighted-topics/66

2. Data, Software, and Model Sharing: Over the past 64 years, ICPSR has built infrastructure for researchers to share data with the academic research community, receiving significant funding from NIH and other federal agencies (e.g., the Office of Justice Programs), as well as private foundations and member institutions. Data sharing has become a core expectation for researchers conducting federally funded scientific research. Today, nearly every federal agency providing research funding requires a data management and sharing plan as a condition of receiving an award (e.g., DOE, 2023; NASA, 2022; NEH,

2025; NIH, 2023; NSF, 2024). Both funding agencies and scholarly journals recognize the value of publishing corroborating data, and they require authors to make their data publicly available in trusted repositories to ensure robust metadata adequately describe and give context to the data, provide systems that protect confidential and sensitive information, register data with persistent, machine-readable identifiers, make data available in formats that will be useable over time, and provide online catalogs and discovery tools to ensure the data can be found and reused. ICPSR supports NIH's efforts to realign its measures of success, in order to recognize scholarly contributions beyond traditional measures of productivity. Research data as a scholarly information object has measurable impact that informs scholarship in ways publication metrics, alone, cannot. ICPSR supports NIH in its efforts to give credit to investigators who do the hard work of preparing, documenting, and sharing data to extend its value for the future.

Tracking the use and reuse of these shared data collections is a practical way for data repositories to help data creators demonstrate the scientific value they have generated. Like many repositories and archives, ICPSR measures the number of accounts created, dataset page views, downloads, data-use agreements, and access requests. However, these metrics do not necessarily indicate that data were cited, analyzed in publications, or used to replicate or extend prior findings. A researcher may download a dataset but never use it, or multiple members of a research team may access data that results in one publication. Conversely, a dataset may be used extensively over time in ways that are not fully captured by download counts alone. To address this limitation, ICPSR supports creating linkages between shared datasets and publications based on those datasets. By linking datasets to downstream publications, ICPSR provides an infrastructure that gives investigators the credit they deserve for preparing, documenting, and sharing data. This architecture is the database known as the ICPSR Bibliography of Data-related Literature (Moss et al., 2015; ICPSR, 2026). A publicly available tool for discovering data via the citing literature, it serves as an example of how repositories can demonstrate the scientific value generated by shared data resources.

Continually expanding, the Bibliography's database currently contains nearly 130,000 citations to publications that meaningfully use data distributed at ICPSR. The collection represents over 60 years of scholarship, including in public health, behavioral health, and biomedical research funded by NIH. Often, output based on these data is published decades after the data were originally collected and shared, proving their continued usefulness. The Bibliography was built with support from the National Science Foundation (SES 9977984) under a continuing grant from 1999 to 2003. Thereafter, it has been funded by ICPSR's 800 member institutions, as well as by foundations and federal agencies. In the past decade, thanks in part to increased support from several NIH-funded data repositories at ICPSR, the Bibliography has grown at an accelerated pace, adding 5,000 to 7,000 citations per year. To populate the Bibliography, a small team of librarians and information professionals set up manual or API queries within bibliographic databases to find evidence of data use in the scholarly literature. They collect citations to publications that use the data and enhance each citation record with links to the data in ICPSR's collection. All publications associated with a given dataset then appear displayed with the data on ICPSR's public websites, like any other important metadata element. Careful human curation is required in order to make the collection meaningful, and not diluted with publications that merely mention datasets but do not actually conduct analysis using them.

The Bibliography can be searched, returning relevant results that point to publications linked to data. In this way, the Bibliography serves as both a data discovery tool and a compilation of scholarly works based on each dataset. Funders, data creators, researchers, and students can see not only the available data in the Bibliography but also how those data were used by others. Each publication's citation links a site user to the full text if they have access to it, and to the underlying data held at ICPSR. Thus, site users can discover analyses of data to replicate, avoid duplicating analyses that have already been done, identify cross-disciplinary implications and uses of the data, study how data resources are used, conduct citation analyses, better understand the data life cycle, and learn of methodological issues associated with the data, some of which are covered solely in the published literature. Crucially, the Bibliography acts as a de facto reward mechanism for data creators. A list of citing works, permanently linked to their shared data, helps demonstrate its value and justify future support from funders.

As the Bibliography has grown, bibliometricians and other scholars have found the full collection of citations, with its highly curated linkages between data and the citing literature, to be a unique dataset (ICPSR, 2026) for conducting research. They take advantage of its rich citation metadata, often combining the Bibliography with other sources to better understand the range of scholarship enabled by shared data. Some researchers have used co-citation analysis to measure the diversity and novelty of research data reuse (Lafia et al., 2022b). Others take advantage of the fact that, by definition, the authors in the Bibliography have analyzed data. By accessing corresponding author information, researchers can contact authors who reuse data and survey them about their experiences working with secondary data (e.g., Million et al., 2025). Hence, having convenient access to a sample of this group of published authors can inform researchers about successful data reuse. Still other researchers have used the data-literature links in the Bibliography to study the downstream impacts of poorly reused or problematic data noted in Retraction Watch, drawing links between retracted papers and archived datasets (Narayan et al., 2026). With the increasing number of retractions in the biomedical literature (Hwang et al., 2023), it is more important than ever to have tools like the Bibliography to use in conjunction with other resources to help make flaws in data more visible to potential reusers.

Due to several factors, it can be difficult and costly to find the universe of published and unpublished works resulting from analyses of data held at ICPSR. For instance, librarians and information professionals who populate the Bibliography must make do with tools not designed for data discovery and use. Most platforms that host bibliographic databases focus on scientific article retrieval and do not support field searching that can isolate search matches where they appear in the methods and data sections of publications. Ironically, due to the inconsistency in how data are described in works and tagged or linked where those works are hosted, the best source for accessing metrics of data use is considered to be the data repository (Apartis, et al., 2024). Poor citing practices also can be a barrier for repositories trying to track data reuse (Vannan, 2020). In ICPSR's case, each citation is evaluated to ensure it is linked accurately to a specific dataset(s), since authors can be imprecise, opaque, or negligent in how they cite research data (Moss & Lyle, 2018). Inevitably, some reuse goes undetected; therefore, the data creator never receives credit for it. More computational efforts are being developed to determine how often shared data are cited. To help in this endeavor, ICPSR supplies data-literature linkage metadata from the Bibliography to the Data Citation Corpus (DataCite & Make Data Count, 2024), which in turn enables linkages to be made by entities like OpenAlex (Priem et al.), a free, open-source catalog and database of the global research ecosystem. With data citation information supplied by

repositories and publishers, their services enable tracking publications and data that are registered with Digital Object Identifiers (DOIs). But these services still depend heavily on repositories to provide vetted data citations. For these reasons and more, the cost of collecting and linking shared data to the literature using it can be too high. As a result, ICPSR is one of only a few large-scale domain repositories tracking data use with human review to ensure the accuracy and relevance of the linkages.

In the future, researchers might leverage machine learning and generative AI (GenAI) to reduce the costs and frustrations associated with manually curating a collection like the Bibliography. In one experiment using the Bibliography, machine learning models were trained to recognize data use in publications (Lafia et al., 2022a), in an effort to reduce the amount of human judgement required to identify actual data use, and to do so at scale. It may be possible to use GenAI to build a domain-agnostic prototype of a tool that other repositories could use. A dynamic bibliographic metadata service could identify candidate related publications from external open source corpora, evaluate likely dataset-publication relationships, and route recommendations for staff review. Other tools may be created using GenAI to mine collections like the Bibliography to reveal and highlight existing interdisciplinary collaborations and other unforeseen benefits of sharing and reusing data. GenAI may also help reveal unexpected aspects of the secondary research being conducted across the social and biomedical sciences, surfacing more ways in which those who share data or share replication packages deserve recognition for the scholarship they enable.

Measurable indicators of science that should be recognized and incentivized:
--Data creators prepare, document, and share data for reuse in trusted data repositories.
--Data repositories link shared data to works that reused the shared data, to the extent possible.
--Repositories and archives help demonstrate the scientific value generated by shared data.
--Funders and data creators can verify that their data are reused, by which researchers, and in what disciplines.
--Data repositories, data creators/agencies, data funders, and information scientists can determine whether reuse extends beyond the original research team, whether the data support new research questions, whether the data contribute to interdisciplinary science, whether the data are used in training, replication, or policy-relevant work, and whether investments in curation and metadata increase downstream reuse.

Implementation approaches:
--Trusted repositories should provide a data citation and an accompanying persistent data identifier for all deposited datasets, requiring data reusers to cite the data in publications in which the data were analyzed.
--NIH should support trusted repositories in their efforts to discover and make linkages between shared data and the literature that uses it.
--NIH should invest in the infrastructure of data repositories and build on existing tools to create additional metrics of downstream reuse, connecting datasets to researchers, institutions, grants, and publications.
--NIH should support repository efforts to enhance publication citations beyond standard citation elements by leveraging robust metadata and persistent identifiers available in open sources.
--NIH should support data repositories seeking to develop GenAI mechanisms to increase efficiency in

finding data for reuse.
--NIH should invest in research to uncover networks of interdisciplinary data reuse hidden in the citing literature.

Potential benefits or unintended consequences:
--Data seekers may gain an understanding of the data by reading works that use it.
--Funders, investigators, and others have access to a metric of shared data reuse.
--High-value data may be recognized by reuse in publications, not simply by download metrics.
--Data reusers may gain a better understanding of the uses and limitations of datasets, avoid duplicating effort, ask new questions, apply new analytical methods, and extend findings from the original dataset, thereby increasing its long-term value.
--Data repositories may be able to measure the impact of NIH-funded data after grants are completed.
--Potential unintended consequences could include placing additional burdens on small data repositories. By providing links between data studies and the citing literature, data repositories risk prioritizing data reuse quantity over the quality of scholarship.

References

Apartis, S., Catalano, G., Consiglio, G., Costas, R. Delugas, E. Dulong de Rosnay, M., Grypari, I., Kárász, I., Klebel, T., Kormann, E., Manola, N., Papageorgiou, H., Seminaroti, E., Stavropoulos, P., Stoy, L., Traag, V. A., van Leeuwen, T. N., Venturini, T., Vignetti, S., … Willemse, T. (2024). Open science impact indicator handbook. Zenodo. 2024. https://doi.org/10.5281/zenodo.14538442.

DataCite, & Make Data Count. (2024). Data Citation Corpus data file. DataCite. https://doi.org/10.5281/zenodo.11216814.

Hwang, S. Y., Yon, D. K., Lee, S. W., Kim, M. S., Kim, J. Y., Smith, L., Koyanagi, A., Solmi, M., Carvalho, A. F., Kim, E., Shin, J. I., & Ioannidis, J. P. A. (2023). Causes for retraction in the biomedical literature: A systematic review of studies of retraction notices. Journal of Korean Medical Science, 38(41), e333. https://doi.org/10.3346/jkms.2023.38.e333

Inter-university Consortium for Political and Social Research (ICPSR). (2026). Data-related publications. Retrieved August 17, 2026, from https://www.icpsr.umich.edu/sites/icpsr/find-data/citations

Inter-university Consortium for Political and Social Research (ICPSR). (2026). ICPSR Bibliography of Data-related Literature: Data-linked Publication Citation Files, Full Download. [Data set]. https://doi.org/10.3886/E218241V3

Lafia, S., Fan, L., & Hemphill, L. (2022a). A natural language processing pipeline for detecting informal data references in academic literature. Proceedings of the Association for Information Science and Technology (ASIS&T). https://doi.org/10.1002/pra2.614

Lafia, S., Fan, L., Thomer, A., & Hemphill, L. (2022b). Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network. Quantitative Science Studies (QSS). https://doi.org/10.48550/arXiv.2205.08395

Million, A. J., York, J., Lafia, S., & Hemphill, L. (2025). Data, not documents: Moving beyond theories of information-seeking behavior to advance data discovery. Journal of the Association for Information Science and Technology, 76(4), 649-664. https://doi.org/10.1002/asi.24962

Moss, E., Cave, C., & Lyle, J. (2015). Chapter 4: Sharing and citing research data: A repository's perspective. In H.K.P. Jayasuriya (Ed.), Big data, big challenges in evidence-based policy making (pp. 47-65). West Academic Publishing. https://hdl.handle.net/2027.42/115490

Moss, E., & Lyle J. (2018, February 20). Opaque data citation: Actual citation practice and its implication for tracking data use. [Poster presentation]. 13th International Digital Curation Conference, Barcelona, Spain. https://deepblue.lib.umich.edu/handle/2027.42/142393

National Aeronautics and Space Administration, Science Mission Directorate (NASA). (2022). Scientific information policy for the Science Mission Directorate (SPD-41a). Retrieved August 8, 2026 from https://www.neh.gov/publicaccess

National Endowment for the Humanities (NEH). (2025). Public access at NEH. Retrieved August 8, 2026 from https://www.neh.gov/publicaccess

National Institutes of Health (NIH). (2023). NIH Policy for Data Management and Sharing. NOT-OD-21-013. Retrieved August 8, 2026 from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-013.html

National Science Foundation (NSF). (2024). Proposal and award policies and procedures guide (NSF 24-1), Section XI.D.4: Dissemination and sharing of research results. Retrieved August 8, 2026 from https://www.nsf.gov/policies/pappg/24-1/ch-11-other-post-award-requirements#ch11D4

Narayan, A., Million, A. J., & Kubale, J. T. (2026). Examining the prevalence and persistence of data implicated in article retractions [Preprint]. OSF. https://osf.io/preprints/metaarxiv/fqezp_v1/

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. [Preprint]. ArXiv. https://arxiv.org/abs/2205.01833

U.S. Department of Energy (DOE). (2023). DOE public access plan. Retrieved August 8, 2026 from https://www.energy.gov/sites/default/files/2023-06/doe-public-access-plan-2023.pdf

Vannan, S., Downs, R. R., Meier, W., Wilson, B. E., & Gerasimov, Irina V. (2020). Data sets are foundational to research. Why don't we cite them?, Eos, 101. https://doi.org/10.1029/2020EO151665.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

247. Engineering Biology Research Consortium

I am responding: On behalf of an organization

Name:

Name of Organization: Engineering Biology Research Consortium

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: 1. Rigor and Reproducibility Engineering biology contributes to biomedical discovery through the development and validation of new biological systems, technologies, and methods. Rigor and reproducibility are therefore important not only for individual hypothesis-driven findings but also for the tools and capabilities that enable subsequent research. Development of an engineered cell system, organoid model, synthetic gene circuit, genome-editing component, or other biological technology may require extensive characterization and benchmarking across biological systems, laboratories, protocols, and measurement environments before reliably supporting other research. NIH should therefore consider evidence such as: -Independent replication, validation, or benchmarking of methods, platforms, biological systems, or findings, documented through publications, reports, datasets, or other appropriate research outputs -Characterization and documentation of limitations, failure modes, and appropriate uses through methods publications, technical reports, protocols, or other research outputs -Development and validation of standardized protocols, reference materials, or measurement approaches, including evidence of independent testing or adoption -Reproducibility of computational models, software, and experimental workflows, supported by accessible code, data, documentation, or replication studies -Evidence that methods, platforms, or biological systems perform consistently across laboratories, biological systems, protocols, measurement environments, or novel application contexts Where possible, NIH should recognize these contributions through existing or emerging research outputs, such as validated protocols, benchmarking datasets, technical reports, methods publications, repository records, or independent validation studies, rather than requiring entirely new reporting mechanisms. These indicators should support expert assessment rather than function as automatic scores or thresholds. NIH should also recognize well-documented negative results, limitations, unsuccessful approaches, and replication failures when they generate information that prevents duplication of ineffective approaches, clarifies the conditions under which findings hold, or improves subsequent research.

2. Data, Software, and Model Sharing: 2. Data, Software, and Model Sharing Engineering biology can contribute to biomedical discovery through shared datasets, computational tools, biological resources, experimental models, and research infrastructure that enable research beyond the projects that created them. A validated genetic-component library, engineered cell line, organoid system, biological dataset, or computational design tool may enable other researchers to pursue new questions, validate findings, or develop new applications. These contributions can advance biomedical discovery even when they are not fully reflected in publication or citation metrics. NIH should not only assess whether resources are

shared, but also their quality, usability, interoperability, and downstream scientific value. Relevant evidence could include: -Quality and completeness of documentation and metadata -Interoperability and adherence to relevant standards -Independent validation or benchmarking -Repository reuse metrics and attribution tracking (e.g., software downloads, dataset citations, biological repository requests) alongside narrative descriptions of downstream adoption -Integration into other research workflows, models, platforms, or novel application contexts -Contributions to reproducibility or new research questions The absence of immediate reuse should not automatically indicate a lack of value, particularly for new resources whose applications may emerge over longer time horizons. NIH should also recognize the development of standards, reference materials, protocols, and other community infrastructure that enables research across institutions

3. Training and Mentorship: 3. Training and Mentorship NIH should recognize training and mentorship that build scientific capabilities across the biomedical research enterprise rather than simply counting trainees supervised or graduated. Engineering biology often requires expertise spanning molecular biology, engineering, computation, automation, and quantitative methods; developing this interdisciplinary capacity can enable new research approaches across an entire field. NIH could consider evidence such as: -Development of interdisciplinary training programs and curricula -Creation of shared training resources and educational materials -Development of new technical competencies -Contributions to interdisciplinary research communities -Evidence of trainees subsequently contributing to research, technology development, public-sector activities, industry, education, or other parts of the biomedical ecosystem Measures should account for differences across career stages and research environments. NIH should recognize both direct mentorship and contributions to broader training infrastructure, including shared curricula, workshops, and community resources.

4. Collaboration: 4. Collaboration As biomedical research becomes increasingly interdisciplinary and convergent, NIH should develop approaches for assessing and recognizing substantive contributions to team-based science without equating contribution with authorship position or individual publication output. Engineering biology is inherently collaborative, often bringing together biologists, engineers, computational scientists, clinicians, and technology developers. Contributions such as enabling technologies, biological platforms, verified datasets, computational methods, and experimental workflows may be essential to a team's scientific output without being fully reflected in authorship position or publication counts. NIH should therefore consider: -Structured contribution statements detailing specific technical, methodological, or platform contributions to team science beyond authorship position -Development of enabling technologies and shared experimental capabilities -Creation of computational methods, biological platforms, or experimental workflows -Contributions to team-based infrastructure -Leadership or coordination of interdisciplinary research efforts -Evidence that a contribution enabled new research, novel applications, or capabilities Evaluation should recognize meaningful scientific contribution within collaborative research rather than simply rewarding the number of collaborations or co-authorships. Structured contribution statements, building on existing NIH approaches to describing contributions to science, could provide useful evidence for reviewers.

5. Entrepreneurship and Translation: 5. Entrepreneurship and Translation Engineering biology spans

6. Foundational Scientific Exploration: 6. Foundational Scientific Exploration Engineering biology also illustrates the importance of recognizing contributions whose significance may emerge only over time. The development of a new biological capability and the generation of large-scale datasets can precede

recognition of the questions, phenomena, or applications they ultimately make possible, creating opportunities for discovery and innovation that are difficult to anticipate in advance. NIH should therefore: -Avoid systematically favoring short-term outcomes over exploratory or foundational research -Recognize enabling technologies and platforms whose applications may be difficult to predict -Account for well-documented negative results, limitations, and iterative development -Allow impact to emerge over longer time horizons -Ensure support for a broad portfolio of biomedical discovery pathways, including hypothesis-driven science as well as the creation of datasets, methodologies, technologies, platforms, and other enabling resources The technologies underlying modern genome editing, such as CRISPR-based techniques, emerged from decades of foundational research whose eventual applications could not necessarily have been predicted. Similarly, large-scale datasets such as those generated through the Human Genome Project have continued to enable discoveries and applications decades after their initial development. More broadly, developing a biological platform or computational capability may itself be a significant scientific contribution because it allows researchers to investigate questions that were previously inaccessible.

7. Public Impact: 7. Public Impact Engineering biology can contribute to public impact through multiple pathways, including advances in health, scientific infrastructure, workforce development, biotechnology and biomanufacturing innovation, and public-sector capabilities. These contributions may be direct or may occur through technologies and platforms that enable subsequent research and applications. NIH should therefore consider evidence such as: -Advances in health and biomedical research -Improvements in clinical practice or patient outcomes -Scientific and technological infrastructure -Standards and reference systems -Workforce development -Public-sector research and technological capabilities -Biotechnology and biomanufacturing innovation Public impact should not be defined solely by immediate clinical outcomes, commercialization, or economic activity. Doing so could systematically disadvantage foundational and enabling research whose ultimate public benefits emerge through subsequent scientific and technological advances or use of technologies in novel settings or application areas. The appropriate pathway and timeframe for demonstrating public impact will vary across research areas, career stages, institution types, and stages of technology development.

8. Other Areas: Other Comments: Implementation, Administrative Burden, and Unintended Consequences NIH should implement any new approach without creating substantial reporting burdens or replacing one rigid metrics system with another. Wherever possible, NIH should use information already available through grant applications, publications, repositories, and other existing systems. Proposed indicators should also be evaluated for unintended consequences, including metric gaming, pressure toward short-term clinical or commercial outcomes, and disadvantages for early-career researchers, interdisciplinary fields, or less-resourced institutions. NIH should consider piloting approaches that give reviewers structured evidence of contribution and impact without creating fixed quantitative thresholds. For example, NIH could allow researchers to identify a limited number of significant contributions and provide evidence such as: -Documented use of a dataset, software tool, biological resource, platform, protocol, or technology by researchers or organizations outside the originating group -Independent replication, external benchmarking, or validation of a method, model, platform, or biological system -Technology transfer, licensing, industry or public-sector adoption, standards development, or incorporation into clinical, research, or manufacturing workflows, where appropriate to the maturity of the work -Structured contribution statements describing substantive scientific or technical contributions that may not be reflected by authorship position -Letters or

statements documenting contributions from relevant stakeholders beyond academic collaborators, including industry, clinical, public-sector, or other research partners, where those stakeholders can provide direct evidence of the contribution's significance, utility, or application Evaluation frameworks should also remain adaptable as scientific practices, technologies, and forms of research output evolve.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/EBRC-Response-to-Request-for-Information-on-Measuring-and-Rewarding-Scientific-Impact_NOT-OD-26-087.pdf

Description: introduction and conclusion to our response.

EBRC Engineering Biology Research Consortium

Response to NIH's Request for Information on Measuring and Rewarding Scientific Impact

The Engineering Biology Research Consortium (EBRC) is a nonprofit, public-private partnership that brings together scientists, engineers, and industry leaders to advance the field of engineering biology to address national and global needs. EBRC's members include experts from over 100 universities and research institutes, alongside leaders from more than 30 companies, philanthropies, and other organizations. Working closely with partners across the engineering biology ecosystem, EBRC focuses on four key areas: Research Roadmapping, Policy & International Engagement, Education, and Security.

EBRC appreciates the opportunity to respond to the National Institutes of Health (NIH) Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087). EBRC supports NIH's efforts to broaden how scientific contribution and impact are recognized across the biomedical and biological research enterprise. Engineering biology exemplifies the interdisciplinary, collaborative, iterative, and translational approaches NIH seeks to better recognize. Traditional publication- and citation-based measures may underrepresent important engineering biology contributions to biomedical research, including shared biological platforms, computational tools, standards, datasets, enabling technologies, research infrastructure, and intellectual property that enable subsequent research, discoveries, and breakthroughs across fields. As NIH considers updating its evaluation criteria, capturing these contributions requires a system designed for flexible, multifaceted research models. EBRC therefore encourages NIH to develop a flexible, multidimensional, principles-based approach that recognizes diverse forms of scientific contribution without replacing one rigid metrics-based system with another.

Cross-Cutting Principles for Measuring and Rewarding Scientific Impact

Across the seven areas identified in the RFI, EBRC recommends that NIH use indicators as complementary evidence for expert assessment, rather than as proxies for scientific value or as automatic scores or thresholds. Evaluation should account for differences in research context, career stage, discipline, institution type, and the time horizon over which impact may emerge. NIH should also avoid indicators that encourage metric gaming, prioritize short-term outcomes, or reward quantity over meaningful contribution.

1. Rigor and Reproducibility

Engineering biology contributes to biomedical discovery through the development and validation of new biological systems, technologies, and methods. Rigor and reproducibility are therefore important not only for individual hypothesis-driven findings but also for the tools and capabilities that enable subsequent research. Development of an engineered cell system,

organoid model, synthetic gene circuit, genome-editing component, or other biological technology may require extensive characterization and benchmarking across biological systems, laboratories, protocols, and measurement environments before reliably supporting other research.

NIH should therefore consider evidence such as:

- Independent replication, validation, or benchmarking of methods, platforms, biological systems, or findings, documented through publications, reports, datasets, or other appropriate research outputs
- Characterization and documentation of limitations, failure modes, and appropriate uses through methods publications, technical reports, protocols, or other research outputs
- Development and validation of standardized protocols, reference materials, or measurement approaches, including evidence of independent testing or adoption
- Reproducibility of computational models, software, and experimental workflows, supported by accessible code, data, documentation, or replication studies
- Evidence that methods, platforms, or biological systems perform consistently across laboratories, biological systems, protocols, measurement environments, or novel application contexts

Where possible, NIH should recognize these contributions through existing or emerging research outputs, such as validated protocols, benchmarking datasets, technical reports, methods publications, repository records, or independent validation studies, rather than requiring entirely new reporting mechanisms.

These indicators should support expert assessment rather than function as automatic scores or thresholds. NIH should also recognize well-documented negative results, limitations, unsuccessful approaches, and replication failures when they generate information that prevents duplication of ineffective approaches, clarifies the conditions under which findings hold, or improves subsequent research.

2. Data, Software, and Model Sharing

Engineering biology can contribute to biomedical discovery through shared datasets, computational tools, biological resources, experimental models, and research infrastructure that enable research beyond the projects that created them. A validated genetic-component library, engineered cell line, organoid system, biological dataset, or computational design tool may enable other researchers to pursue new questions, validate findings, or develop new applications. These contributions can advance biomedical discovery even when they are not fully reflected in publication or citation metrics.

NIH should not only assess whether resources are shared, but also their quality, usability, interoperability, and downstream scientific value. Relevant evidence could include:

- Quality and completeness of documentation and metadata
- Interoperability and adherence to relevant standards
- Independent validation or benchmarking
- Repository reuse metrics and attribution tracking (e.g., software downloads, dataset citations, biological repository requests) alongside narrative descriptions of downstream adoption
- Integration into other research workflows, models, platforms, or novel application contexts
- Contributions to reproducibility or new research questions

The absence of immediate reuse should not automatically indicate a lack of value, particularly for new resources whose applications may emerge over longer time horizons. NIH should also recognize the development of standards, reference materials, protocols, and other community infrastructure that enables research across institutions.

3. Training and Mentorship

NIH should recognize training and mentorship that build scientific capabilities across the biomedical research enterprise rather than simply counting trainees supervised or graduated. Engineering biology often requires expertise spanning molecular biology, engineering, computation, automation, and quantitative methods; developing this interdisciplinary capacity can enable new research approaches across an entire field.

NIH could consider evidence such as:

- Development of interdisciplinary training programs and curricula
- Creation of shared training resources and educational materials
- Development of new technical competencies
- Contributions to interdisciplinary research communities
- Evidence of trainees subsequently contributing to research, technology development, public-sector activities, industry, education, or other parts of the biomedical ecosystem

Measures should account for differences across career stages and research environments. NIH should recognize both direct mentorship and contributions to broader training infrastructure, including shared curricula, workshops, and community resources.

4. Collaboration

As biomedical research becomes increasingly interdisciplinary and convergent, NIH should develop approaches for assessing and recognizing substantive contributions to team-based science without equating contribution with authorship position or individual publication output.

Engineering biology is inherently collaborative, often bringing together biologists, engineers, computational scientists, clinicians, and technology developers. Contributions such as enabling technologies, biological platforms, verified datasets, computational methods, and experimental workflows may be essential to a team's scientific output without being fully reflected in authorship position or publication counts.

NIH should therefore consider:

- Structured contribution statements detailing specific technical, methodological, or platform contributions to team science beyond authorship position
- Development of enabling technologies and shared experimental capabilities
- Creation of computational methods, biological platforms, or experimental workflows
- Contributions to team-based infrastructure
- Leadership or coordination of interdisciplinary research efforts
- Evidence that a contribution enabled new research, novel applications, or capabilities

Evaluation should recognize meaningful scientific contribution within collaborative research rather than simply rewarding the number of collaborations or co-authorships. Structured contribution statements, building on existing NIH approaches to describing contributions to science, could provide useful evidence for reviewers.

5. Entrepreneurship and Translation

Engineering biology spans fundamental discovery, engineering, technology development, validation, scale-up, and application, with meaningful contributions occurring throughout this continuum. A novel genetic circuit, genome-editing technology, computational design tool, engineered cell system, or organoid model may demonstrate translational value through technical validation, characterization, standardization, and de-risking well before commercialization.

NIH should consider technology maturity, including where appropriate frameworks such as Technology Readiness Levels (TRLs). To assess translational progress without creating rigid commercialization targets, reviewers should consider contextual evidence such as:

- Documented progress from proof-of-concept toward testing in relevant application environments
- External characterization, multi-laboratory validation, or testing by industry, clinical, or public-sector partners
- Executed MTAs, licensing agreements, or other technology-transfer activities
- Contributions to consensus standards or regulatory readiness activities
- Adoption of design tools, genetic parts, or protocols into third-party research or DBTL workflows

- Application or implementation of emerging technologies in new domains or translational contexts

These indicators should serve as qualitative evidence for expert assessment rather than requirements for successful research. NIH should avoid creating pressure to commercialize research that is better suited to foundational or enabling work. Approaches to data and resource sharing should similarly balance openness with legitimate considerations related to intellectual property, privacy, biosecurity, and responsible access.

6. Foundational Scientific Exploration

Engineering biology also illustrates the importance of recognizing contributions whose significance may emerge only over time. The development of a new biological capability and the generation of large-scale datasets can precede recognition of the questions, phenomena, or applications they ultimately make possible, creating opportunities for discovery and innovation that are difficult to anticipate in advance.

NIH should therefore:

- Avoid systematically favoring short-term outcomes over exploratory or foundational research
- Recognize enabling technologies and platforms whose applications may be difficult to predict
- Account for well-documented negative results, limitations, and iterative development
- Allow impact to emerge over longer time horizons
- Ensure support for a broad portfolio of biomedical discovery pathways, including hypothesis-driven science as well as the creation of datasets, methodologies, technologies, platforms, and other enabling resources

The technologies underlying modern genome editing, such as CRISPR-based techniques, emerged from decades of foundational research whose eventual applications could not necessarily have been predicted. Similarly, large-scale datasets such as those generated through the Human Genome Project have continued to enable discoveries and applications decades after their initial development. More broadly, developing a biological platform or computational capability may itself be a significant scientific contribution because it allows researchers to investigate questions that were previously inaccessible.

7. Public Impact

Engineering biology can contribute to public impact through multiple pathways, including advances in health, scientific infrastructure, workforce development, biotechnology and biomanufacturing innovation, and public-sector capabilities. These contributions may be direct

or may occur through technologies and platforms that enable subsequent research and applications.

NIH should therefore consider evidence such as:

- Advances in health and biomedical research
- Improvements in clinical practice or patient outcomes
- Scientific and technological infrastructure
- Standards and reference systems
- Workforce development
- Public-sector research and technological capabilities
- Biotechnology and biomanufacturing innovation

Public impact should not be defined solely by immediate clinical outcomes, commercialization, or economic activity. Doing so could systematically disadvantage foundational and enabling research whose ultimate public benefits emerge through subsequent scientific and technological advances or use of technologies in novel settings or application areas. The appropriate pathway and timeframe for demonstrating public impact will vary across research areas, career stages, institution types, and stages of technology development.

Other Comments: Implementation, Administrative Burden, and Unintended Consequences

NIH should implement any new approach without creating substantial reporting burdens or replacing one rigid metrics system with another. Wherever possible, NIH should use information already available through grant applications, publications, repositories, and other existing systems. Proposed indicators should also be evaluated for unintended consequences, including metric gaming, pressure toward short-term clinical or commercial outcomes, and disadvantages for early-career researchers, interdisciplinary fields, or less-resourced institutions.

NIH should consider piloting approaches that give reviewers structured evidence of contribution and impact without creating fixed quantitative thresholds. For example, NIH could allow researchers to identify a limited number of significant contributions and provide evidence such as:

- Documented use of a dataset, software tool, biological resource, platform, protocol, or technology by researchers or organizations outside the originating group
- Independent replication, external benchmarking, or validation of a method, model, platform, or biological system
- Technology transfer, licensing, industry or public-sector adoption, standards development, or incorporation into clinical, research, or manufacturing workflows, where appropriate to the maturity of the work

- Structured contribution statements describing substantive scientific or technical contributions that may not be reflected by authorship position
- Letters or statements documenting contributions from relevant stakeholders beyond academic collaborators, including industry, clinical, public-sector, or other research partners, where those stakeholders can provide direct evidence of the contribution's significance, utility, or application

Evaluation frameworks should also remain adaptable as scientific practices, technologies, and forms of research output evolve.

Conclusion

EBRC supports NIH's effort both to broaden and to more accurately define how scientific contribution and impact are recognized and rewarded across the biomedical research enterprise. Engineering biology demonstrates both the need for this evolution and the challenges of developing measures that can fairly capture modern, interdisciplinary, technology-driven research.

A flexible, multidimensional, principles-based approach can better recognize diverse contributions by:

- Recognizing indicators of scientific impact and contributions beyond traditional publications, citations, journal impact factor, and authorship position
- Maintaining expert judgment in evaluation
- Accounting for differences across disciplines, career stages, and research models
- Recognizing foundational and long-term contributions
- Recognizing enabling technologies, platforms, data, software, standards, and infrastructure as critical to biomedical discovery
- Minimizing administrative burden and unintended incentives
- Allowing evaluation frameworks to evolve alongside science and technology

Such an approach can strengthen incentives for rigorous, collaborative, innovative, and translational research while preserving support for foundational science and the diverse pathways through which biomedical impact emerges.

248. Michael E. Bales, Elizabeth A. Wood, Inmaculada de Melo-Martin, Matt Mikaelian, Kelly McConnell, Julianne Imperato-McGinley, Principal Investigator

I am responding: On behalf of myself

Name: Michael E. Bales, Elizabeth A. Wood, Inmaculada de Melo-Martin, Matt Mikaelian, Kelly McConnell, Julianne Imperato-McGinley, Principal Investigator

Name of Organization: Clinical and Translational Science Center, Weill Cornell Medicine

Type of Organization: Academic Institution

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/RFI-NOT-OD-26-087-response-WCM-CTSC.pdf

Description: RFI response from Weill Cornell Medicine CTSC

Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Memorandum / RFI Response
To: National Institutes of Health (NIH)
From: Michael E. Bales, Elizabeth A. Wood, Inmaculada de Melo-Martin, Matt Mikaelian, Kelly McConnell, Julianne Imperato-McGinley, Principal Investigator, Clinical and Translational Science Center, Weill Cornell Medicine
Date: August 19th, 2026
Re: Response to RFI on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Introduction

We appreciate the opportunity to provide feedback supporting the NIH's efforts to refine how scientific impact is measured and rewarded. It is widely recognized that traditional metrics of scientific output, such as publication counts, citation impact, and funded grants, cannot fully characterize how scientific work impacts clinical care and population health. As such, there are clear opportunities to refine measurement of scientific impacts and rewards which recognize the complex and non-linear nature of bench to bedside translational science.

A Dual Framework for Impact: Discovery vs. Community

We propose a framework that recognizes two distinct, but complementary, forms of scientific impact. First, *discovery impact* is the direct impact of the scientific work. It may encompass success in achieving a grant's specific aims, the advancement of knowledge, or the validation of a new theory. Discovery impact, which is foundational to advances in translational science, should itself be recognized and rewarded. Second, *community impact* reflects the downstream impact of scientific work on the communities it is meant to serve. Community impact is increasingly valued as an ultimate goal of translational research, as it indicates successful translation of scientific work to improved clinical and public health outcomes. While both types of impact are important, there has generally been a stronger emphasis on discovery impact. Academic reward structures have favored innovative or theoretically significant findings, often with a lack of emphasis on how these findings are leveraged to improve clinical care and population health.

Because there is nearly always a significant delay between a scientific discovery and its translation into practice, academic reward structures should be broadened to further incentivize downstream community impact – for example, changes in clinical practice

or healthcare policy – without discounting the importance of the basic discoveries that made these improvements possible.

Limitations of Current Frameworks

While tools such as the Translational Science Benefits Model (TSBM) and other approaches like logic models are valuable in understanding and communicating impact, there is a tendency to employ them linearly – for example, by identifying clear and direct connections between research findings and subsequent downstream impacts. However, the journey from bench to bedside is not always linear; it can be an indirect, long-term journey involving multiple pathways.

There are several important challenges in substantiating community impact. For one, improvements in clinical and public health practice often arise from cumulative knowledge and systemic changes, as noted above, rather than from a single discovery. Additionally, the data and information required to demonstrate clinical impact is often stored in healthcare delivery systems or community health records rather than in academic databases.

Bridging the Gap: Implementation Science and Community Engagement

There is increasing recognition among CTSA awardees of the importance of applying the principles of dissemination and implementation (D&I) science to translational science. Despite a substantial and well-researched D&I literature base, many academic health systems lack robust mechanisms to ensure that research is not just translated, but implemented and sustained.

To address this, we propose several strategic shifts in how impact is measured and rewarded. First, we must invest in training CTSC-supported researchers – especially basic scientists – in D&I science. It is essential to develop effective methods for teaching D&I that are relevant to researchers, especially those who may not traditionally view their work through an implementation lens. Second, community-partnered metrics should be incorporated into the research and implementation process. Rather than simply obtaining community validation of a pre-determined result, communities should have meaningful influence over research priorities, outcomes, and implementation. When this is accomplished, it is more straightforward to demonstrate that research projects were tailored to be responsive to community needs.

Given the complexity of team science, we believe that social network modeling approaches can reveal how knowledge diffuses through professional networks into

clinical practice, capturing the often less visible work of mentorship and knowledge transfer.

Recommendations for Rewarding Impact

To support collaborative research that leads to community impact, we recommend the NIH consider:

- Establish incentives that reward both high-value discovery impact (e.g., basic discovery or advancement of knowledge) as well as community impact (e.g., the creation of guidelines, changes to policy, or the successful adaptation of a treatment across diverse clinical settings).
- To account for longer-term implementation in clinical medicine and public health, expand the timeframe for measuring impact beyond the standard grant cycle.
- Given the essential role of implementation partners and community stakeholders in achieving downstream impact, ensure that their contributions are recognized.

A Note on Unintended Consequences

Finally, it bears mentioning that once measures are tied to funding and professional rewards, researchers may optimize their work to perform well on specific metrics. It may therefore be important to monitor unintended consequences, such as efforts to optimize performance on the measures, overstating causal contributions, or neglecting valuable work whose impact is difficult to quantify according to the measures used. We recommend that any new reward systems include mechanisms to monitor these unintended consequences. This will help ensure that the pursuit of measurable impact does not interfere with the pursuit of rigorous, responsible science.

249. Inter-university Consortium for Political and Social Research (ICPSR)

I am responding: On behalf of an organization

Name: Elizabeth Moss, with Maggie Levenstein, and Trent Alexander

Name of Organization: Inter-university Consortium for Political and Social Research (ICPSR)

Type of Organization: Other

Type of Organization - Other: Academic Research Data Repository

Role: Other

Role - Other: Librarian (Moss); Investigator/Researcher (Levenstein and Alexander)

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: The Inter-university Consortium for Political and Social Research (ICPSR) appreciates the opportunity to respond to NIH's Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087). ICPSR is one of the world's largest digital repositories of social and behavioral science research data. As a replication-ready repository, ICPSR preserves data, documentation, and provenance metadata, with the technical and legal infrastructure to support sharing, including for data that are large, complex, or sensitive. Certified by CoreTrustSeal (ICPSR, 2024), ICPSR is an archive dedicated to providing good data stewardship that maximizes the value of federal investments, ensures that scientists are credited for creating data products, supports replication of original research, and creates new opportunities for research. ICPSR currently preserves and disseminates approximately 27,000 data studies containing hundreds of thousands of individual data files, with over 850,000 downloads each year. As a strong supporter of replication in biomedical, behavioral, and social science research, ICPSR hosts more than 7,000 research replication packages in journal repositories on its self-publishing platforms. Additionally, for 63 years, successive generations of early-career scholars have received training in quantitative research methodology through courses in the ICPSR Summer Program, where they are exposed to best practices in research. Further, ICPSR promotes transparency and replicability by publishing extensive guidelines for data depositors on how to prepare data files and documentation to ensure their reuse (ICPSR, 2020). ICPSR strongly supports NIH's interest in identifying the "critical research and infrastructure needs necessary to advance rigorous and reliable scientific findings" (NIH, 2026). ICPSR further supports NIH's efforts to expand measures of scientific success through acknowledging, elevating, and rewarding scientists who share data. In tandem with NIH, ICPSR also recognizes the value in rewarding scholars who reuse shared data and then make adequate information and tools available to replicate their results.

In addition to serving as data stewards and supporting study replication, ICPSR has a long history of linking publication-associated replication materials to the curated data it distributes. Data distributed by ICPSR form the foundation for tens of thousands of research articles, reports, dissertations, books, and other works that advance science, and are collected in the ICPSR Bibliography of Data-related Literature (ICPSR, 2026), a one-of-a-kind searchable database that captures, organizes, and makes discoverable research findings stemming from ICPSR datasets. Each citation record in the database is enhanced with

metadata linking the publication to the underlying data used by its author. The Bibliography plays a crucial role in demonstrating the long-term impact of NIH-funded datasets. However, a further step could be taken to recognize authors who reuse data and make it easier to test their methods and reproduce their findings.

Data repositories are well positioned to play a larger role in recognizing investigators who reuse data and provide replication packages when they publish their findings, since the platforms on which shared data are made available form a natural home for source data, published findings based on those data, and the tools essential to replicating study findings. Linking such resources together makes each meaningful in context, multiplies the usefulness and impact of NIH investments, and helps recognize investigators' rigorous efforts to make results from publicly funded research reproducible.

Measurable indicators of science that should be recognized and incentivized:
--Researchers publish in journals or on platforms that assign scholarly works with persistent identifiers, e.g., DOIs, to ensure long-term accessibility.
--Researchers reuse archived data and include links to their replication packages that contain their analytic code, software, and other tools essential for replication.
--Data repositories flag or highlight publications in which authors include links to their replication packages.
--Data repositories encourage authors to deposit their replication packages, which are linked to the original dataset they reused.
--Data repositories that house replication packages requiring restricted access offer a streamlined process for granting access for testing replicability.
--Data repositories systematically collect and link their distributed data to the citing literature, and they also link findings, replication packages, and original archived data to disseminate them together, thereby simplifying and promoting the reproduction of results.

Implementation approaches:
--Trusted repositories should provide a data citation and an accompanying persistent data identifier for all deposited datasets, and require data reusers to cite the data in publications in which the data were analyzed.
--Publishers should require the explicit citation of source data and replication packages in scholarly works. Without explicit citation and persistent identifiers, automatic linking among data, literature, and replication packages is more difficult.
--NIH should support archival system infrastructure that links research data to the primary and secondary findings generated from those data.
--NIH should support efforts by archives and repositories to single out publications that provide access to replication packages and use their platforms to promote best practices.
--Investigators who have received recognition from trusted data repositories can include such accomplishments in their NIH grant applications and progress reports.
NIH should encourage archives not only to collect links to data-related publications but also to find and link to, or even archive, the replication packages associated with those publications.
--NIH should encourage archives and repositories to reach out to authors whose replication materials are not available online but are available by request, and encourage authors to preserve them.

--NIH should support efforts by archives and repositories to disseminate data with the citing papers and the replication packages associated with them.

Potential benefits or unintended consequences:
--Authors may more frequently make use of provided citations and unique identifiers to formally cite data.
--It becomes easier to identify which shared datasets support which publications, analyses, and downstream discoveries.
--It becomes easier for researchers to build upon existing research, since the citing literature is identified and available for replication or extension.
--Transparency and replicability are prioritized by making replication materials available alongside the data.
--More authors will provide links to replication packages when publishing results.
--By collecting publications that reuse data and including replication packages, archives and repositories will be able to identify investigators who enable replication.
--Greater public trust in, and a stronger evidence base for, biomedical and behavioral science will result.
--A potential unintended consequence is that creating viable replication packages when restricted data were used can be difficult due to privacy concerns, indirectly penalizing researchers who reuse those data. Archives could mitigate this by providing secure options for replication testing. Also, by highlighting publications whose authors provided replication packages, repositories risk endorsing the content of their analysis rather than the practice of sharing the necessary tools to replicate the experiment or study using the same methods and conditions to achieve the same result.

References
Inter-university Consortium for Political and Social Research (ICPSR). (2026). Data-related publications. Retrieved August 17, 2026, from https://www.icpsr.umich.edu/sites/icpsr/find-data/citations

ICPSR. (2024). "2027-08-22 - Inter-university Consortium for Political and Social Research (ICPSR) - CoreTrustSeal Requirements 2023-2025. Retrieved August 8, 2026 from https://doi.org/10.34894/WCWZU0

ICPSR. (2020). Guide to social science data preparation and archiving best practice throughout the data life cycle, 6th Edition. https://wpvip.icpsr.umich.edu/icpsr/wp-content/uploads/sites/11/dataprep-wcag.pdf

National Institutes of Health (NIH). (2026). Enhancing scientific rigor, transparency and replicability. Retrieved August 17, 2026 from https://grants.nih.gov/funding/find-a-fit-for-your-research/highlighted-topics/66

2. Data, Software, and Model Sharing: Over the past 64 years, ICPSR has built infrastructure for researchers to share data with the academic research community, receiving significant funding from NIH and other federal agencies (e.g., the Office of Justice Programs), as well as private foundations and member institutions. Data sharing has become a core expectation for researchers conducting federally funded scientific research. Today, nearly every federal agency providing research funding requires a data management and sharing plan as a condition of receiving an award (e.g., DOE, 2023; NASA, 2022; NEH,

2025; NIH, 2023; NSF, 2024). Both funding agencies and scholarly journals recognize the value of publishing corroborating data, and they require authors to make their data publicly available in trusted repositories to ensure robust metadata adequately describe and give context to the data, provide systems that protect confidential and sensitive information, register data with persistent, machine-readable identifiers, make data available in formats that will be useable over time, and provide online catalogs and discovery tools to ensure the data can be found and reused. ICPSR supports NIH's efforts to realign its measures of success, in order to recognize scholarly contributions beyond traditional measures of productivity. Research data as a scholarly information object has measurable impact that informs scholarship in ways publication metrics, alone, cannot. ICPSR supports NIH in its efforts to give credit to investigators who do the hard work of preparing, documenting, and sharing data to extend its value for the future.

Tracking the use and reuse of these shared data collections is a practical way for data repositories to help data creators demonstrate the scientific value they have generated. Like many repositories and archives, ICPSR measures the number of accounts created, dataset page views, downloads, data-use agreements, and access requests. However, these metrics do not necessarily indicate that data were cited, analyzed in publications, or used to replicate or extend prior findings. A researcher may download a dataset but never use it, or multiple members of a research team may access data that results in one publication. Conversely, a dataset may be used extensively over time in ways that are not fully captured by download counts alone. To address this limitation, ICPSR supports creating linkages between shared datasets and publications based on those datasets. By linking datasets to downstream publications, ICPSR provides an infrastructure that gives investigators the credit they deserve for preparing, documenting, and sharing data. This architecture is the database known as the ICPSR Bibliography of Data-related Literature (Moss et al., 2015; ICPSR, 2026). A publicly available tool for discovering data via the citing literature, it serves as an example of how repositories can demonstrate the scientific value generated by shared data resources.

Continually expanding, the Bibliography's database currently contains nearly 130,000 citations to publications that meaningfully use data distributed at ICPSR. The collection represents over 60 years of scholarship, including in public health, behavioral health, and biomedical research funded by NIH. Often, output based on these data is published decades after the data were originally collected and shared, proving their continued usefulness. The Bibliography was built with support from the National Science Foundation (SES 9977984) under a continuing grant from 1999 to 2003. Thereafter, it has been funded by ICPSR's 800 member institutions, as well as by foundations and federal agencies. In the past decade, thanks in part to increased support from several NIH-funded data repositories at ICPSR, the Bibliography has grown at an accelerated pace, adding 5,000 to 7,000 citations per year. To populate the Bibliography, a small team of librarians and information professionals set up manual or API queries within bibliographic databases to find evidence of data use in the scholarly literature. They collect citations to publications that use the data and enhance each citation record with links to the data in ICPSR's collection. All publications associated with a given dataset then appear displayed with the data on ICPSR's public websites, like any other important metadata element. Careful human curation is required in order to make the collection meaningful, and not diluted with publications that merely mention datasets but do not actually conduct analysis using them.

The Bibliography can be searched, returning relevant results that point to publications linked to data. In this way, the Bibliography serves as both a data discovery tool and a compilation of scholarly works based on each dataset. Funders, data creators, researchers, and students can see not only the available data in the Bibliography but also how those data were used by others. Each publication's citation links a site user to the full text if they have access to it, and to the underlying data held at ICPSR. Thus, site users can discover analyses of data to replicate, avoid duplicating analyses that have already been done, identify cross-disciplinary implications and uses of the data, study how data resources are used, conduct citation analyses, better understand the data life cycle, and learn of methodological issues associated with the data, some of which are covered solely in the published literature. Crucially, the Bibliography acts as a de facto reward mechanism for data creators. A list of citing works, permanently linked to their shared data, helps demonstrate its value and justify future support from funders.

As the Bibliography has grown, bibliometricians and other scholars have found the full collection of citations, with its highly curated linkages between data and the citing literature, to be a unique dataset (ICPSR, 2026) for conducting research. They take advantage of its rich citation metadata, often combining the Bibliography with other sources to better understand the range of scholarship enabled by shared data. Some researchers have used co-citation analysis to measure the diversity and novelty of research data reuse (Lafia et al., 2022b). Others take advantage of the fact that, by definition, the authors in the Bibliography have analyzed data. By accessing corresponding author information, researchers can contact authors who reuse data and survey them about their experiences working with secondary data (e.g., Million et al., 2025). Hence, having convenient access to a sample of this group of published authors can inform researchers about successful data reuse. Still other researchers have used the data-literature links in the Bibliography to study the downstream impacts of poorly reused or problematic data noted in Retraction Watch, drawing links between retracted papers and archived datasets (Narayan et al., 2026). With the increasing number of retractions in the biomedical literature (Hwang et al., 2023), it is more important than ever to have tools like the Bibliography to use in conjunction with other resources to help make flaws in data more visible to potential reusers.

Due to several factors, it can be difficult and costly to find the universe of published and unpublished works resulting from analyses of data held at ICPSR. For instance, librarians and information professionals who populate the Bibliography must make do with tools not designed for data discovery and use. Most platforms that host bibliographic databases focus on scientific article retrieval and do not support field searching that can isolate search matches where they appear in the methods and data sections of publications. Ironically, due to the inconsistency in how data are described in works and tagged or linked where those works are hosted, the best source for accessing metrics of data use is considered to be the data repository (Apartis, et al., 2024). Poor citing practices also can be a barrier for repositories trying to track data reuse (Vannan, 2020). In ICPSR's case, each citation is evaluated to ensure it is linked accurately to a specific dataset(s), since authors can be imprecise, opaque, or negligent in how they cite research data (Moss & Lyle, 2018). Inevitably, some reuse goes undetected; therefore, the data creator never receives credit for it. More computational efforts are being developed to determine how often shared data are cited. To help in this endeavor, ICPSR supplies data-literature linkage metadata from the Bibliography to the Data Citation Corpus (DataCite & Make Data Count, 2024), which in turn enables linkages to be made by entities like OpenAlex (Priem et al.), a free, open-source catalog and database of the global research ecosystem. With data citation information supplied by

repositories and publishers, their services enable tracking publications and data that are registered with Digital Object Identifiers (DOIs). But these services still depend heavily on repositories to provide vetted data citations. For these reasons and more, the cost of collecting and linking shared data to the literature using it can be too high. As a result, ICPSR is one of only a few large-scale domain repositories tracking data use with human review to ensure the accuracy and relevance of the linkages.

In the future, researchers might leverage machine learning and generative AI (GenAI) to reduce the costs and frustrations associated with manually curating a collection like the Bibliography. In one experiment using the Bibliography, machine learning models were trained to recognize data use in publications (Lafia et al., 2022a), in an effort to reduce the amount of human judgement required to identify actual data use, and to do so at scale. It may be possible to use GenAI to build a domain-agnostic prototype of a tool that other repositories could use. A dynamic bibliographic metadata service could identify candidate related publications from external open source corpora, evaluate likely dataset-publication relationships, and route recommendations for staff review. Other tools may be created using GenAI to mine collections like the Bibliography to reveal and highlight existing interdisciplinary collaborations and other unforeseen benefits of sharing and reusing data. GenAI may also help reveal unexpected aspects of the secondary research being conducted across the social and biomedical sciences, surfacing more ways in which those who share data or share replication packages deserve recognition for the scholarship they enable.

Measurable indicators of science that should be recognized and incentivized:
--Data creators prepare, document, and share data for reuse in trusted data repositories.
--Data repositories link shared data to works that reused the shared data, to the extent possible.
--Repositories and archives help demonstrate the scientific value generated by shared data.
--Funders and data creators can verify that their data are reused, by which researchers, and in what disciplines.
--Data repositories, data creators/agencies, data funders, and information scientists can determine whether reuse extends beyond the original research team, whether the data support new research questions, whether the data contribute to interdisciplinary science, whether the data are used in training, replication, or policy-relevant work, and whether investments in curation and metadata increase downstream reuse.

Implementation approaches:
--Trusted repositories should provide a data citation and an accompanying persistent data identifier for all deposited datasets, requiring data reusers to cite the data in publications in which the data were analyzed.
--NIH should support trusted repositories in their efforts to discover and make linkages between shared data and the literature that uses it.
--NIH should invest in the infrastructure of data repositories and build on existing tools to create additional metrics of downstream reuse, connecting datasets to researchers, institutions, grants, and publications.
--NIH should support repository efforts to enhance publication citations beyond standard citation elements by leveraging robust metadata and persistent identifiers available in open sources.
--NIH should support data repositories seeking to develop GenAI mechanisms to increase efficiency in

finding data for reuse.
--NIH should invest in research to uncover networks of interdisciplinary data reuse hidden in the citing literature.

Potential benefits or unintended consequences:
--Data seekers may gain an understanding of the data by reading works that use it.
--Funders, investigators, and others have access to a metric of shared data reuse.
--High-value data may be recognized by reuse in publications, not simply by download metrics.
--Data reusers may gain a better understanding of the uses and limitations of datasets, avoid duplicating effort, ask new questions, apply new analytical methods, and extend findings from the original dataset, thereby increasing its long-term value.
--Data repositories may be able to measure the impact of NIH-funded data after grants are completed.
--Potential unintended consequences could include placing additional burdens on small data repositories. By providing links between data studies and the citing literature, data repositories risk prioritizing data reuse quantity over the quality of scholarship.

References

Apartis, S., Catalano, G., Consiglio, G., Costas, R. Delugas, E. Dulong de Rosnay, M., Grypari, I., Kárász, I., Klebel, T., Kormann, E., Manola, N., Papageorgiou, H., Seminaroti, E., Stavropoulos, P., Stoy, L., Traag, V. A., van Leeuwen, T. N., Venturini, T., Vignetti, S., … Willemse, T. (2024). Open science impact indicator handbook. Zenodo. 2024. https://doi.org/10.5281/zenodo.14538442.

DataCite, & Make Data Count. (2024). Data Citation Corpus data file. DataCite. https://doi.org/10.5281/zenodo.11216814.

Hwang, S. Y., Yon, D. K., Lee, S. W., Kim, M. S., Kim, J. Y., Smith, L., Koyanagi, A., Solmi, M., Carvalho, A. F., Kim, E., Shin, J. I., & Ioannidis, J. P. A. (2023). Causes for retraction in the biomedical literature: A systematic review of studies of retraction notices. Journal of Korean Medical Science, 38(41), e333. https://doi.org/10.3346/jkms.2023.38.e333

Inter-university Consortium for Political and Social Research (ICPSR). (2026). Data-related publications. Retrieved August 17, 2026, from https://www.icpsr.umich.edu/sites/icpsr/find-data/citations

Inter-university Consortium for Political and Social Research (ICPSR). (2026). ICPSR Bibliography of Data-related Literature: Data-linked Publication Citation Files, Full Download. [Data set]. https://doi.org/10.3886/E218241V3

Lafia, S., Fan, L., & Hemphill, L. (2022a). A natural language processing pipeline for detecting informal data references in academic literature. Proceedings of the Association for Information Science and Technology (ASIS&T). https://doi.org/10.1002/pra2.614

Lafia, S., Fan, L., Thomer, A., & Hemphill, L. (2022b). Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network. Quantitative Science Studies (QSS). https://doi.org/10.48550/arXiv.2205.08395

Million, A. J., York, J., Lafia, S., & Hemphill, L. (2025). Data, not documents: Moving beyond theories of information-seeking behavior to advance data discovery. Journal of the Association for Information Science and Technology, 76(4), 649-664. https://doi.org/10.1002/asi.24962

Moss, E., Cave, C., & Lyle, J. (2015). Chapter 4: Sharing and citing research data: A repository's perspective. In H.K.P. Jayasuriya (Ed.), Big data, big challenges in evidence-based policy making (pp. 47-65). West Academic Publishing. https://hdl.handle.net/2027.42/115490

Moss, E., & Lyle J. (2018, February 20). Opaque data citation: Actual citation practice and its implication for tracking data use. [Poster presentation]. 13th International Digital Curation Conference, Barcelona, Spain. https://deepblue.lib.umich.edu/handle/2027.42/142393
National Aeronautics and Space Administration, Science Mission Directorate (NASA). (2022). Scientific information policy for the Science Mission Directorate (SPD-41a). Retrieved August 8, 2026 from https://www.neh.gov/publicaccess

National Endowment for the Humanities (NEH). (2025). Public access at NEH. Retrieved August 8, 2026 from https://www.neh.gov/publicaccess

National Institutes of Health (NIH). (2023). NIH Policy for Data Management and Sharing. NOT-OD-21-013. Retrieved August 8, 2026 from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-013.html

National Science Foundation (NSF). (2024). Proposal and award policies and procedures guide (NSF 24-1), Section XI.D.4: Dissemination and sharing of research results. Retrieved August 8, 2026 from https://www.nsf.gov/policies/pappg/24-1/ch-11-other-post-award-requirements#ch11D4

Narayan, A., Million, A. J., & Kubale, J. T. (2026). Examining the prevalence and persistence of data implicated in article retractions [Preprint]. OSF. https://osf.io/preprints/metaarxiv/fqezp_v1/

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. [Preprint]. ArXiv. https://arxiv.org/abs/2205.01833

U.S. Department of Energy (DOE). (2023). DOE public access plan. Retrieved August 8, 2026 from https://www.energy.gov/sites/default/files/2023-06/doe-public-access-plan-2023.pdf

Vannan, S., Downs, R. R., Meier, W., Wilson, B. E., & Gerasimov, Irina V. (2020). Data sets are foundational to research. Why don't we cite them?, Eos, 101. https://doi.org/10.1029/2020EO151665.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/ICPSR-response-to-NIH-RFI-Measuring-and-Rewarding-Scientific-Impact-NOT-OD-26-087.pdf

Description: Note this submission is a duplicate of one sent already today, but the copy of that submission emailed to us did not retain all of the text in some places. This is a second attempt to submit our response using the form. In case a similar issue occurs, we have uploaded a PDF file containing the same information being sent via the form. Thank you.

Rigor and Reproducibility

The Inter-university Consortium for Political and Social Research (ICPSR) appreciates the opportunity to respond to NIH's Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087). ICPSR is one of the world's largest digital repositories of social and behavioral science research data. As a replication-ready repository, ICPSR preserves data, documentation, and provenance metadata, with the technical and legal infrastructure to support sharing, including for data that are large, complex, or sensitive. Certified by CoreTrustSeal (ICPSR, 2024), ICPSR is an archive dedicated to providing good data stewardship that maximizes the value of federal investments, ensures that scientists are credited for creating data products, supports replication of original research, and creates new opportunities for research. ICPSR currently preserves and disseminates approximately 27,000 data studies containing hundreds of thousands of individual data files, with over 850,000 downloads each year. As a strong supporter of replication in biomedical, behavioral, and social science research, ICPSR hosts more than 7,000 research replication packages in journal repositories on its self-publishing platforms. Additionally, for 63 years, successive generations of early-career scholars have received training in quantitative research methodology through courses in the ICPSR Summer Program, where they are exposed to best practices in research. Further, ICPSR promotes transparency and replicability by publishing extensive guidelines for data depositors on how to prepare data files and documentation to ensure their reuse (ICPSR, 2020). ICPSR strongly supports NIH's interest in identifying the "critical research and infrastructure needs necessary to advance rigorous and reliable scientific findings" (NIH, 2026). ICPSR further supports NIH's efforts to expand measures of scientific success through acknowledging, elevating, and rewarding scientists who share data. In tandem with NIH, ICPSR also recognizes the value in rewarding scholars who reuse shared data and then make adequate information and tools available to replicate their results.

In addition to serving as data stewards and supporting study replication, ICPSR has a long history of linking publication-associated replication materials to the curated data it distributes. Data distributed by ICPSR form the foundation for tens of thousands of research articles, reports, dissertations, books, and other works that advance science, and are collected in the ICPSR Bibliography of Data-related Literature (ICPSR, 2026), a one-of-a-kind searchable database that captures, organizes, and makes discoverable research findings stemming from ICPSR datasets. Each citation record in the database is enhanced with metadata linking the publication to the underlying data used by its author. The Bibliography plays a crucial role in demonstrating the long-term impact of NIH-funded datasets. However, a further step could be taken to recognize authors who reuse data and make it easier to test their methods and reproduce their findings.

Data repositories are well positioned to play a larger role in recognizing investigators who reuse data and provide replication packages when they publish their findings, since the platforms on which shared data are made available form a natural home for source data, published findings based on those data, and the tools essential to replicating study findings. Linking such

resources together makes each meaningful in context, multiplies the usefulness and impact of NIH investments, and helps recognize investigators' rigorous efforts to make results from publicly funded research reproducible.

Measurable indicators of science that should be recognized and incentivized:

- Researchers publish in journals or on platforms that assign scholarly works with persistent identifiers, e.g., DOIs, to ensure long-term accessibility.
- Researchers reuse archived data and include links to their replication packages that contain their analytic code, software, and other tools essential for replication.
- Data repositories flag or highlight publications in which authors include links to their replication packages.
- Data repositories encourage authors to deposit their replication packages, which are linked to the original dataset they reused.
- Data repositories that house replication packages requiring restricted access offer a streamlined process for granting access for testing replicability.
- Data repositories systematically collect and link their distributed data to the citing literature, and they also link findings, replication packages, and original archived data to disseminate them together, thereby simplifying and promoting the reproduction of results.

Implementation approaches:

- Trusted repositories should provide a data citation and an accompanying persistent data identifier for all deposited datasets, and require data reusers to cite the data in publications in which the data were analyzed.
- Publishers should require the explicit citation of source data and replication packages in scholarly works. Without explicit citation and persistent identifiers, automatic linking among data, literature, and replication packages is more difficult.
- NIH should support archival system infrastructure that links research data to the primary and secondary findings generated from those data.
- NIH should support efforts by archives and repositories to single out publications that provide access to replication packages and use their platforms to promote best practices.
- Investigators who have received recognition from trusted data repositories can include such accomplishments in their NIH grant applications and progress reports.
- NIH should encourage archives not only to collect links to data-related publications but also to find and link to, or even archive, the replication packages associated with those publications.
- NIH should encourage archives and repositories to reach out to authors whose replication materials are not available online but are available by request, and encourage authors to preserve them.
- NIH should support efforts by archives and repositories to disseminate data with the citing papers and the replication packages associated with them.

Potential benefits or unintended consequences:

- Authors may more frequently make use of provided citations and unique identifiers to formally cite data.
- It becomes easier to identify which shared datasets support which publications, analyses, and downstream discoveries.
- It becomes easier for researchers to build upon existing research, since the citing literature is identified and available for replication or extension.
- Transparency and replicability are prioritized by making replication materials available alongside the data.
- More authors will provide links to replication packages when publishing results.
- By collecting publications that reuse data and including replication packages, archives and repositories will be able to identify investigators who enable replication.
- Greater public trust in, and a stronger evidence base for, biomedical and behavioral science will result.
- A potential unintended consequence is that creating viable replication packages when restricted data were used can be difficult due to privacy concerns, indirectly penalizing researchers who reuse those data. Archives could mitigate this by providing secure options for replication testing. Also, by highlighting publications whose authors provided replication packages, repositories risk endorsing the content of their analysis rather than the practice of sharing the necessary tools to replicate the experiment or study using the same methods and conditions to achieve the same result.

References

Inter-university Consortium for Political and Social Research (ICPSR). (2026). *Data-related publications*. Retrieved August 17, 2026, from https://www.icpsr.umich.edu/sites/icpsr/find-data/citations

ICPSR. (2024). *"2027-08-22 - Inter-university Consortium for Political and Social Research (ICPSR) - CoreTrustSeal Requirements 2023-2025*. Retrieved August 8, 2026 from https://doi.org/10.34894/WCWZU0

ICPSR. (2020). *Guide to social science data preparation and archiving best practice throughout the data life cycle, 6th Edition*. https://wpvip.icpsr.umich.edu/icpsr/wp-content/uploads/sites/11/dataprep-wcag.pdf

National Institutes of Health (NIH). (2026). *Enhancing scientific rigor, transparency and replicability.* Retrieved August 17, 2026 from https://grants.nih.gov/funding/find-a-fit-for-your-research/highlighted-topics/66

Data, Software, and Model Sharing

Over the past 64 years, ICPSR has built infrastructure for researchers to share data with the academic research community, receiving significant funding from NIH and other federal agencies (e.g., the Office of Justice Programs), as well as private foundations and member institutions. Data sharing has become a core expectation for researchers conducting federally funded scientific research. Today, nearly every federal agency providing research funding requires a data management and sharing plan as a condition of receiving an award (e.g., DOE, 2023; NASA, 2022; NEH, 2025; NIH, 2023; NSF, 2024). Both funding agencies and scholarly journals recognize the value of publishing corroborating data, and they require authors to make their data publicly available in trusted repositories to ensure robust metadata adequately describe and give context to the data, provide systems that protect confidential and sensitive information, register data with persistent, machine-readable identifiers, make data available in formats that will be useable over time, and provide online catalogs and discovery tools to ensure the data can be found and reused. ICPSR supports NIH's efforts to realign its measures of success, in order to recognize scholarly contributions beyond traditional measures of productivity. Research data as a scholarly information object has measurable impact that informs scholarship in ways publication metrics, alone, cannot. ICPSR supports NIH in its efforts to give credit to investigators who do the hard work of preparing, documenting, and sharing data to extend its value for the future.

Tracking the use and reuse of these shared data collections is a practical way for data repositories to help data creators demonstrate the scientific value they have generated. Like many repositories and archives, ICPSR measures the number of accounts created, dataset page views, downloads, data-use agreements, and access requests. However, these metrics do not necessarily indicate that data were cited, analyzed in publications, or used to replicate or extend prior findings. A researcher may download a dataset but never use it, or multiple members of a research team may access data that results in one publication. Conversely, a dataset may be used extensively over time in ways that are not fully captured by download counts alone. To address this limitation, ICPSR supports creating linkages between shared datasets and publications based on those datasets. By linking datasets to downstream publications, ICPSR provides an infrastructure that gives investigators the credit they deserve for preparing, documenting, and sharing data. This architecture is the database known as the ICPSR Bibliography of Data-related Literature (Moss et al., 2015; ICPSR, 2026). A publicly available tool for discovering data via the citing literature, it serves as an example of how repositories can demonstrate the scientific value generated by shared data resources.

Continually expanding, the Bibliography's database currently contains nearly 130,000 citations to publications that meaningfully use data distributed at ICPSR. The collection represents over 60 years of scholarship, including in public health, behavioral health, and biomedical research funded by NIH. Often, output based on these data is published decades after the data were originally collected and shared, proving their continued usefulness. The Bibliography was built with support from the National Science Foundation (SES 9977984) under a continuing grant from 1999 to 2003. Thereafter, it has been funded by ICPSR's 800 member institutions, as well

as by foundations and federal agencies. In the past decade, thanks in part to increased support from several NIH-funded data repositories at ICPSR, the Bibliography has grown at an accelerated pace, adding 5,000 to 7,000 citations per year. To populate the Bibliography, a small team of librarians and information professionals set up manual or API queries within bibliographic databases to find evidence of data use in the scholarly literature. They collect citations to publications that use the data and enhance each citation record with links to the data in ICPSR's collection. All publications associated with a given dataset then appear displayed with the data on ICPSR's public websites, like any other important metadata element. Careful human curation is required in order to make the collection meaningful, and not diluted with publications that merely mention datasets but do not actually conduct analysis using them.

The Bibliography can be searched, returning relevant results that point to publications linked to data. In this way, the Bibliography serves as both a data discovery tool and a compilation of scholarly works based on each dataset. Funders, data creators, researchers, and students can see not only the available data in the Bibliography but also how those data were used by others. Each publication's citation links a site user to the full text if they have access to it, and to the underlying data held at ICPSR. Thus, site users can discover analyses of data to replicate, avoid duplicating analyses that have already been done, identify cross-disciplinary implications and uses of the data, study how data resources are used, conduct citation analyses, better understand the data life cycle, and learn of methodological issues associated with the data, some of which are covered solely in the published literature. Crucially, the Bibliography acts as a de facto reward mechanism for data creators. A list of citing works, permanently linked to their shared data, helps demonstrate its value and justify future support from funders.

As the Bibliography has grown, bibliometricians and other scholars have found the full collection of citations, with its highly curated linkages between data and the citing literature, to be a unique dataset (ICPSR, 2026) for conducting research. They take advantage of its rich citation metadata, often combining the Bibliography with other sources to better understand the range of scholarship enabled by shared data. Some researchers have used co-citation analysis to measure the diversity and novelty of research data reuse (Lafia et al., 2022b). Others take advantage of the fact that, by definition, the authors in the Bibliography have analyzed data. By accessing corresponding author information, researchers can contact authors who reuse data and survey them about their experiences working with secondary data (e.g., Million et al., 2025). Hence, having convenient access to a sample of this group of published authors can inform researchers about successful data reuse. Still other researchers have used the data-literature links in the Bibliography to study the downstream impacts of poorly reused or problematic data noted in Retraction Watch, drawing links between retracted papers and archived datasets (Narayan et al., 2026). With the increasing number of retractions in the biomedical literature (Hwang et al., 2023), it is more important than ever to have tools like the Bibliography to use in conjunction with other resources to help make flaws in data more visible to potential reusers.

Due to several factors, it can be difficult and costly to find the universe of published and unpublished works resulting from analyses of data held at ICPSR. For instance, librarians and information professionals who populate the Bibliography must make do with tools not designed

for data discovery and use. Most platforms that host bibliographic databases focus on scientific article retrieval and do not support field searching that can isolate search matches where they appear in the methods and data sections of publications. Ironically, due to the inconsistency in how data are described in works and tagged or linked where those works are hosted, the best source for accessing metrics of data use is considered to be the data repository (Apartis, et al., 2024). Poor citing practices also can be a barrier for repositories trying to track data reuse (Vannan, 2020). In ICPSR's case, each citation is evaluated to ensure it is linked accurately to a specific dataset(s), since authors can be imprecise, opaque, or negligent in how they cite research data (Moss & Lyle, 2018). Inevitably, some reuse goes undetected; therefore, the data creator never receives credit for it. More computational efforts are being developed to determine how often shared data are cited. To help in this endeavor, ICPSR supplies data-literature linkage metadata from the Bibliography to the Data Citation Corpus (DataCite & Make Data Count, 2024), which in turn enables linkages to be made by entities like OpenAlex (Priem et al.), a free, open-source catalog and database of the global research ecosystem. With data citation information supplied by repositories and publishers, their services enable tracking publications and data that are registered with Digital Object Identifiers (DOIs). But these services still depend heavily on repositories to provide vetted data citations. For these reasons and more, the cost of collecting and linking shared data to the literature using it can be too high. As a result, ICPSR is one of only a few large-scale domain repositories tracking data use with human review to ensure the accuracy and relevance of the linkages.

In the future, researchers might leverage machine learning and generative AI (GenAI) to reduce the costs and frustrations associated with manually curating a collection like the Bibliography. In one experiment using the Bibliography, machine learning models were trained to recognize data use in publications (Lafia et al., 2022a), in an effort to reduce the amount of human judgement required to identify actual data use, and to do so at scale. It may be possible to use GenAI to build a domain-agnostic prototype of a tool that other repositories could use. A dynamic bibliographic metadata service could identify candidate related publications from external open source corpora, evaluate likely dataset-publication relationships, and route recommendations for staff review. Other tools may be created using GenAI to mine collections like the Bibliography to reveal and highlight existing interdisciplinary collaborations and other unforeseen benefits of sharing and reusing data. GenAI may also help reveal unexpected aspects of the secondary research being conducted across the social and biomedical sciences, surfacing more ways in which those who share data or share replication packages deserve recognition for the scholarship they enable.

Measurable indicators of science that should be recognized and incentivized:

- Data creators prepare, document, and share data for reuse in trusted data repositories.
- Data repositories link shared data to works that reused the shared data, to the extent possible.
- Repositories and archives help demonstrate the scientific value generated by shared data.

- Funders and data creators can verify that their data are reused, by which researchers, and in what disciplines.
- Data repositories, data creators/agencies, data funders, and information scientists can determine whether reuse extends beyond the original research team, whether the data support new research questions, whether the data contribute to interdisciplinary science, whether the data are used in training, replication, or policy-relevant work, and whether investments in curation and metadata increase downstream reuse.

Implementation approaches:

- Trusted repositories should provide a data citation and an accompanying persistent data identifier for all deposited datasets, requiring data reusers to cite the data in publications in which the data were analyzed.
- NIH should support trusted repositories in their efforts to discover and make linkages between shared data and the literature that uses it.
- NIH should invest in the infrastructure of data repositories and build on existing tools to create additional metrics of downstream reuse, connecting datasets to researchers, institutions, grants, and publications.
- NIH should support repository efforts to enhance publication citations beyond standard citation elements by leveraging robust metadata and persistent identifiers available in open sources.
- NIH should support data repositories seeking to develop GenAI mechanisms to increase efficiency in finding data for reuse.
- NIH should invest in research to uncover networks of interdisciplinary data reuse hidden in the citing literature.

Potential benefits or unintended consequences:

- Data seekers may gain an understanding of the data by reading works that use it.
- Funders, investigators, and others have access to a metric of shared data reuse.
- High-value data may be recognized by reuse in publications, not simply by download metrics.
- Data reusers may gain a better understanding of the uses and limitations of datasets, avoid duplicating effort, ask new questions, apply new analytical methods, and extend findings from the original dataset, thereby increasing its long-term value.
- Data repositories may be able to measure the impact of NIH-funded data after grants are completed.
- Potential unintended consequences could include placing additional burdens on small data repositories. By providing links between data studies and the citing literature, data repositories risk prioritizing data reuse quantity over the quality of scholarship.

References

Apartis, S., Catalano, G., Consiglio, G., Costas, R. Delugas, E. Dulong de Rosnay, M., Grypari, I., Kárász, I., Klebel, T., Kormann, E., Manola, N., Papageorgiou, H., Seminaroti, E., Stavropoulos, P., Stoy, L., Traag, V. A., van Leeuwen, T. N., Venturini, T., Vignetti, S., … Willemse, T. (2024). *Open science impact indicator handbook.* Zenodo. 2024. https://doi.org/10.5281/zenodo.14538442.

DataCite, & Make Data Count. (2024). *Data Citation Corpus data file*. DataCite. https://doi.org/10.5281/zenodo.11216814.

Hwang, S. Y., Yon, D. K., Lee, S. W., Kim, M. S., Kim, J. Y., Smith, L., Koyanagi, A., Solmi, M., Carvalho, A. F., Kim, E., Shin, J. I., & Ioannidis, J. P. A. (2023). Causes for retraction in the biomedical literature: A systematic review of studies of retraction notices. *Journal of Korean Medical Science*, *38*(41), e333. https://doi.org/10.3346/jkms.2023.38.e333

Inter-university Consortium for Political and Social Research (ICPSR). (2026). *Data-related publications*. Retrieved August 17, 2026, from https://www.icpsr.umich.edu/sites/icpsr/find-data/citations

Inter-university Consortium for Political and Social Research (ICPSR). (2026). *ICPSR Bibliography of Data-related Literature: Data-linked Publication Citation Files, Full Download*. [Data set]. https://doi.org/10.3886/E218241V3

Lafia, S., Fan, L., & Hemphill, L. (2022a). A natural language processing pipeline for detecting informal data references in academic literature. *Proceedings of the Association for Information Science and Technology (ASIS&T)*. https://doi.org/10.1002/pra2.614

Lafia, S., Fan, L., Thomer, A., & Hemphill, L. (2022b). Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network. *Quantitative Science Studies (QSS)*. https://doi.org/10.48550/arXiv.2205.08395

Million, A. J., York, J., Lafia, S., & Hemphill, L. (2025). Data, not documents: Moving beyond theories of information‐seeking behavior to advance data discovery. *Journal of the Association for Information Science and Technology*, *76*(4), 649-664. https://doi.org/10.1002/asi.24962

Moss, E., Cave, C., & Lyle, J. (2015). Chapter 4: Sharing and citing research data: A repository's perspective. In H.K.P. Jayasuriya (Ed.), *Big data, big challenges in evidence-based policy making* (pp. 47-65). West Academic Publishing. https://hdl.handle.net/2027.42/115490

Moss, E., & Lyle J. (2018, February 20). *Opaque data citation: Actual citation practice and its implication for tracking data use*. [Poster presentation]. 13th International Digital Curation Conference, Barcelona, Spain. https://deepblue.lib.umich.edu/handle/2027.42/142393

National Aeronautics and Space Administration, Science Mission Directorate (NASA). (2022). *Scientific information policy for the Science Mission Directorate (SPD-41a)*. Retrieved August 8, 2026 from https://www.neh.gov/publicaccess

National Endowment for the Humanities (NEH). (2025). *Public access at NEH*. Retrieved August 8, 2026 from https://www.neh.gov/publicaccess

National Institutes of Health (NIH). (2023). *NIH Policy for Data Management and Sharing*. NOT-OD-21-013. Retrieved August 8, 2026 from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-013.html

National Science Foundation (NSF). (2024). *Proposal and award policies and procedures guide (NSF 24-1), Section XI.D.4: Dissemination and sharing of research results*. Retrieved August 8, 2026 from https://www.nsf.gov/policies/pappg/24-1/ch-11-other-post-award-requirements#ch11D4

Narayan, A., Million, A. J., & Kubale, J. T. (2026). *Examining the prevalence and persistence of data implicated in article retractions* [Preprint]. OSF. https://osf.io/preprints/metaarxiv/fqezp_v1/

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. [Preprint]. ArXiv. https://arxiv.org/abs/2205.01833

U.S. Department of Energy (DOE). (2023). *DOE public access plan*. Retrieved August 8, 2026 from https://www.energy.gov/sites/default/files/2023-06/doe-public-access-plan-2023.pdf

Vannan, S., Downs, R. R., Meier, W., Wilson, B. E., & Gerasimov, Irina V. (2020). Data sets are foundational to research. Why don’t we cite them?, *Eos*, *101*. https://doi.org/10.1029/2020EO151665.

250. Micah Altman and Philip N Cohen

I am responding: On behalf of myself

Name: Micah Altman and Philip N Cohen

Name of Organization: Personal

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/NIH-26-087-Response.pdf

Aug 19, 2026

Dr. Jay Bhattacharya
National Institutes of Health
Re: Response to Request for Information on Measuring and Rewarding Scientific Impact
Via electronic filing
(<https://osp.od.nih.gov/comment-form-measuring-and-rewarding-scientific-impact/>).

Dear Dr. Bhattacharya,

We have spent much of our careers studying reproducibility, reliability, and the impact of scientific research -- as scientists who study the role of science in society, and who study how scientific information is produced and communicated. This comment is based on our evaluation of the current state of research and practice in the science of science and draws specifically on our prior research and commentary on these topics [1-6].

Given the breadth of the proposed RFI, we focus our comments on two topics central to our research expertise and to the RFI's core purpose: Systematically increasing rigor and reproducibility in NIH funding, and core requirements for reaping public benefit from scientific research.

Rigor and Reproducibility

One of the most substantial barriers to improved rigor and reproducibility in science is the lack of a systematic, scientific approach to 'operational science policies', defined as the policies and practices that effectively control the conduct and communication of formal, natural, and social sciences. [1-3] As one of the largest and most important science funders, NIH is in a uniquely influential position to advance rigor and reproducibility in this area – both through how NIH updates and evaluates its own policies and through the research NIH supports.

While many policies for grant review, scholarly publication, and research transparency have served science and society well, there is an increasing need to update these policies and practices. However, changes to new policies must be designed as experiments that incorporate evaluation by both NIH and the scientific community. In particular, every policy change should include an assessment of how that policy affects the cost of science, who participates in it, and the rigor, impact, and quality of the resulting science.

Furthermore, it is vital that the selection of goals for each policy change, the design of assessments of those policy changes, and the resulting evaluations be transparent and deliberative. Science is a collective activity that works best when the community engages systematically to reach consensus based on reason and evidence. [6] Because of this, developing a collective understanding of what operational science approaches work depends on systematic monitoring of and reporting on operational science policies, interventions, and

outcomes; common standards for communicating and measuring operational interventions and assessments; and support for infrastructure needed to collect, durably store, and transparently share data about new projects in a transparent way. [1-3]

Moreover, recent systematic reviews of the state of reproducibility and reliability of science across multiple fields emphasize the need to continue advancing the general theory, methodology, and evidence base of metascience research. [4] To support rigorous biomedical sciences and public health, NIH should make significant investments in basic and multidisciplinary metascience—including grant programs to support experimental reproducibility interventions; data-sharing and replication infrastructure; evaluation and evidence-synthesis methodologies; and theory development for the science of science.

Public Impact

Transparency and dissemination are essential for both scientific rigor and impact—and can be designed to reinforce each other. Science has the greatest impact when the results are shared, and rigor and reliability depend in substantial part on the extent to which materials, methods, and results are accessible for evaluation, critique, and reproduction. Contemporary advances in information technology and infrastructure have expanded both the scope of scientific knowledge and the social impact of science. Innovation in research practices, cultures, and infrastructure has the potential to enable substantially more open and durable access to all the components of science production – and to lower the barriers facing individuals who aim to learn from and participate in science. [5]

In particular, in order to achieve its public impact potential, science depends on durable, open infrastructure for publishing results and sharing evidence. [1,5] The history of science development in the U.S. and around the world shows that this infrastructure cannot survive without public funding – especially from major science funders such as the NIH. But we cannot assume that all investments provide the same incentives and benefits. NIH should leverage the tools and infrastructure it supports to collect, analyze, and share the evidence necessary to evaluate its policy changes. And it should set the tone by incentivizing a scientific approach to studying operational science policy, and by establishing a climate in which each intervention can contribute cumulatively to the collective public and scientific knowledge.

And, to a large extent, public impact depends on public trust. The public's confidence in science grows from a web of legitimacy, sewn together by signals such as degrees, publications, affiliations, and accreditations. This network reflects the evolution of norms and practices that together comprise scientific communities. Public agencies play a key role in supporting scientific communities while allowing the independence they require to thrive. That independence, in turn, reinforces public trust and protects the reputation of science from the threat of politicization. Democratic governments take the lead in determining which goals a society pursues – and how to balance competing values through public investment – but in the face of high-stakes issues involving complex systems and uncertain approaches, scientific expert consensus serves as an honest broker to assess likely consequences of public policies. Government should not restrict scientists' freedom to engage in policy discussions, and should support the collective processes

needed for science: including supporting the costs of scientific dissemination (e.g. journal publications), deliberation (e.g. conferences) and resource sharing (e.g. data archiving). [1,6]

Summary

Interventions in science policy and practice should be treated, at least in part, as scientific experiments—designed to incrementally add to our cumulative knowledge of how, and how well, science works.

We should not expect NIH's processes to be perfect. It is worth noting, however, that NIH has previously changed its operational policies substantially -- including three major changes to peer review in the last two decades -- and still lacks clear public assessments of what was learned from prior. The public should be able to rely on NIH to change policies with the goals of improving scientific rigor and public benefit; to design those policies for systematic measurement; and to transparently share these performance measurements and conclusions drawn from them. Scientists should be able to rely on the NIH to support independent research evaluating its operational policies and to fund basic meta-science research needed to advance rigor and public impact in biomedical science.

The last decade has seen an acceleration in the rate of new initiatives to innovate in how science operates. NIH is not alone in calling for major changes to operational science policy without supporting plans to study the effects of those changes — absence of evaluation is endemic. Making progress will require that we thoughtfully design operational science policies to enable public assessment. And requires resources and attention – to support internal evaluation, independent external evaluation, and a durable public knowledge infrastructure.

Sincerely,

Micah Altman, Research Scientist, MIT
Philip N. Cohen, Professor of Sociology, University of Maryland

References:

[1] Altman, M., Cohen, P. N., & Polka, J. (2023). Interventions in scholarly communication: Design lessons from public health. First Monday, 28(8). https://doi.org/10.5210/fm.v28i8.12941
[2] Altman, M., & Cohen, P. N. (2023). Evaluating peer review at NIH. Science, 380(6649), 1022-1022. https://doi.org/10.1126/science.adi1570
[3] Altman, M., & Cohen, P. N. (2023). We are in a period of science policy innovation, yet there are major evidence gaps in evaluating their effectiveness*. LSE Impact,* https://blogs.lse.ac.uk/impactofsocialsciences/2023/07/27/we-are-in-a-period-of-science-policy-innovation-yet-there-are-major-evidence-gaps-in-evaluating-their-effectiveness/
[4] Altman M. and Cohen P.N. (2022) The Scholarly Knowledge Ecosystem: Challenges and Opportunities for the Field of Information. Front. Res. Metr. Anal. 6:751553. doi: https://doi.org/10.3389/frma.2021.751553

[5] Altman M., Boug C, Cohen P. N., et. al (2018) A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science, [Report] https://doi.org/10.21428/62b3421f
[6] Altman M. and Cohen P.N. (2026) ‘Expertise’ shouldn’t be a bad word – expert consensus guides science and society. *The Conversation*, https://theconversation.com/expertise-shouldnt-be-a-bad-word-expert-consensus-guides-science-and-society-271467

251. Wiley

I am responding: On behalf of an organization

Name: Jessica Sebeok

Name of Organization: Wiley

Type of Organization: Publisher

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Wiley is the largest U.S. scientific publisher and the largest scientific society partner publisher, as well as a global leader in research and learning. Wiley's story begins in 1807, when Charles Wiley opened a small print shop on Reade Street in New York City. In our young republic's era of invention, the shop became a conduit for pioneering American ideas, first through literature and then through technical, medical, and scientific works. After the Civil War, William H. Wiley, a former U.S. Representative oriented the family company toward the applied sciences that powered American industrial growth by putting practical, reliable knowledge into the hands of engineers, physicians, and educators. When scholarship moved online, Wiley modernized again, first to digitized journals and reference works, and more recently to integrated platforms that combine submission, screening, peer review, and the permanent curation and preservation of the Version of Record. Over more than two centuries we have published the work of U.S. Nobel Prize winners and partnered with hundreds of U.S. scientific societies, supporting thousands of federally funded authors each year.

That 200-year vantage point—spanning the full lifecycle of federally funded research, from submission and peer review through correction, preservation, and reuse—shapes how we approach NIH's questions on rigor and reproducibility. Reproducibility is not a property of a single experiment; it is a property of a research record reliable enough for others to test, correct, and build upon. As the steward of that record, the publisher sits at the center of any effort to measure and reward reproducible science.

The Version of Record (VoR) is the authoritative, final, and citable form of a scholarly work, reflecting completed peer review, verified methods and data-availability statements, disclosed conflicts of interest and funding, and any subsequent corrections or retractions. It is the only version guaranteed to be checked for accuracy and curated by experts over time, and it is what allows a finding to be located, interpreted, and independently reproduced years after publication. This is the publisher's core contribution to rigor: the peer review, integrity screening, metadata stewardship, and post-publication correction that make a result trustworthy enough to reproduce. As NIH designs indicators of reproducibility, it should treat the permanence and integrity of this record as foundational rather than ancillary, consistent with the tenets of Gold Standard Science articulated in Executive Order 14303.

To strengthen NIH's framework for rigor and reproducibility, we recommend the following approaches:

Recognize the Version of Record as the authoritative basis for reproducibility assessment. Preprints and

other non-peer-reviewed outputs play a useful role in rapid communication, but they are not substitutes for the peer-reviewed record and should not be treated as equivalent in metrics frameworks.

Credit negative and null results in funding and career-advancement review. Well-conducted studies that find no effect are essential to an honest and complete literature; their absence from the record is a root cause of the reproducibility problem.

Rely on publisher-led integrity infrastructure, including peer review, editorial assessment, research-integrity screening, corrections, retractions, and post-publication stewardship, as part of the reproducibility ecosystem rather than treating these as separate from NIH's priorities.

Recognize long-term and programmatic research contributions rather than rewarding the pace of publication. Metrics that prize rapid output can push researchers toward premature claims; NIH should credit sustained, careful work that genuine reproducibility requires.

Account for disciplinary variation, since replication looks different across fields. A flexible framework that adapts to field-specific norms and methods will be more effective than applying a single standard to all research.

Wiley and the broader publishing community welcome opportunities to collaborate with NIH on the standards, infrastructure, and incentives that make reproducible science possible.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: Advancing NIH's mission requires a stable and growing biomedical research workforce, supported by stronger measures of the impact of training and mentorship. In our experience across the U.S. research ecosystem, mentorship is not a supplementary benefit but the primary mechanism by which the tacit skills of rigorous science are transmitted from one generation to the next. Long-horizon discovery depends on giving young researchers the time, resources, and guidance to learn from lead researchers on projects that may not yield results for a decade or more.

Federal laboratories and NIH-supported training environments are uniquely positioned to serve as national training grounds, not only for technical excellence but for the standards of rigor, reproducibility, and responsible conduct that underpin Gold Standard Science. Protecting mentored time on long-duration projects is essential to sustaining the discovery pipeline and delivering public-health returns years after the initial investment. As Wiley noted in its response to the U.S. Department of Energy's Genesis Mission RFI, the durability of the U.S. research workforce also depends on distributing training opportunity broadly across institutions and regions. NIH should ensure that training and mentorship indicators capture participation from R1, R2, and R3 universities, Historically Black Colleges and Universities, Tribal Colleges, and community colleges, so that emerging talent is developed wherever it is found rather than only at the best-resourced institutions.

To assess and reward training and mentorship impact, NIH should consider the following approaches:

Track the longitudinal trajectories of researchers who receive mentored training: whether they go on to

win independent funding, publish as principal investigators, establish productive laboratories, and remain in the biomedical workforce.

Recognize peer review and editorial participation as a form of mentorship and professional development. Serving as a reviewer or early-career editor is one of the primary ways scientists learn to evaluate evidence, judge methodology, and communicate uncertainty.

Support structured mentorship programs that can be delivered in both in-person and hybrid formats to extend access to trainees who lack proximity to major research centers, strengthening geographic resilience in the national talent pipeline.

4. Collaboration: Biomedical progress increasingly depends on large, interdisciplinary teams working across institutions, sectors, and Federal agencies. However, traditional measures of scientific achievement, such as author position, publication counts, and citation totals, do not always capture the full range of contributions required to conduct complex collaborative research. Wiley has advocated for more comprehensive approaches to recognizing team-based science in prior submissions, including our response to the U.S. Department of Energy's Genesis Mission RFI, and we encourage NIH to adopt a framework that reflects the true architecture of modern biomedical collaboration.

Publishers can support this effort by preserving a clear and reliable record of who contributed to a study and how. Publication records can document contributions beyond writing and lead authorship, including data collection and curation, methodology, software development, project coordination, statistical analysis, mentoring, and research-resource sharing. This infrastructure already exists and can serve as the foundation for any NIH framework for recognizing collaborative contributions.

NIH should consider the following to assess and recognize collaboration:

Recognize outputs beyond traditional journal articles. Data, software, research methods, shared infrastructure, and reproducibility work can all provide substantial value to a collaborative project and should be considered when evaluating investigators and research teams.

Coordinate with other Federal research agencies and with publishers on common principles for documenting and attributing research contributions, using persistent identifiers such as ORCID and DOIs to link researchers to their outputs across institutions and over time.

Adopt the CRediT taxonomy, which attributes individual researcher contributions across the research pipeline in a standardized, machine-readable format at submission.

Assign weight to team-level and program-level outcomes in funding review, particularly for large-scale, inter-agency, or public-private initiatives where success depends on coordinated effort rather than the work of a single investigator.

Build safeguards against unintended incentives such as inflated authorship or exaggerated contribution claims, by relying on transparent contribution statements that are preserved as part of the published Version of Record and subject to post-publication correction.

A more complete approach to evaluating team science would allow NIH to recognize the researchers who generate, organize, validate, and translate biomedical knowledge, not only those whose names appear in the most prominent author positions.

5. Entrepreneurship and Translation: The translation of discovery from bench to bedside depends on a scientific record that industry, investors, clinicians, and regulators can trust, and that record is the publisher's product. The peer-reviewed Version of Record, with its verified methods, disclosures, and corrections, is the defensible foundation on which patents are filed, products are developed, and clinical decisions are made. A discovery that cannot be located, cited, and relied upon with confidence cannot be translated, however promising it may be.

The publisher's role in translation is therefore not incidental but enabling. Wiley's investments in metadata, persistent identifiers, taxonomy, and integrity screening allow a finding to be discovered and built upon by a start-up or research-and-development team years after it is first published. As AI accelerates the translational pipeline, rights-cleared and provenance-rich content becomes even more critical: products and regulated systems built on unverified or unlawfully obtained inputs carry commercial, regulatory, and litigation risk that deters investment. The Wiley AI Gateway demonstrates how licensed, peer-reviewed content can be embedded in AI-enabled discovery tools, for tasks such as biomarker discovery and clinical trial design, with full provenance and attribution, accelerating translation while preserving research integrity.

To assess and reward translational and entrepreneurial impact, NIH should consider these approaches:

Track patents and regulatory submissions that cite federally funded publications as a concrete, verifiable indicator of translational impact, and include these in the criteria used to evaluate investigators and programs.

Count licensing agreements and start-up formation arising from NIH-funded research as recognized outputs, with appropriate attribution to the researchers, institutions, and publishers whose work enabled them.

Monitor incorporation of NIH-funded findings into clinical guidelines, standards of care, and public health policy, recognizing that these outcomes often represent the most direct translation of research into health benefit.

Encourage Federal procurement and partnership mechanisms to favor AI systems and analytical tools built on licensed, verified, rights-cleared research inputs, providing industry the legal certainty it needs to invest in translation and ensuring that federally supported innovation rests on defensible scientific foundations.

6. Foundational Scientific Exploration: Basic research and high-risk, high-reward inquiry carry an inherent risk of failure alongside the potential for paradigm-shifting breakthroughs. Wiley is proud to sponsor the Golden Goose Award, which recognizes research that appeared obscure, or even accidental, that produced significant benefits for public health and the economy. The award illustrates a principle

that NIH's own mission embodies: the value of foundational science is often not visible at the outset and cannot be judged by short-horizon metrics. When research is evaluated only by how quickly it produces results or how many citations it earns, the slow, uncertain, exploratory work that leads to the most important discoveries tends to get left behind.

Publishers sustain this kind of exploration by preserving the complete record of inquiry, including incremental findings, detailed methods, and the negative and null results that keep a field honest and prevent duplicated effort. The permanence of the Version of Record and long-term archiving mean that a finding regarded as obscure today remains discoverable years later, when its significance becomes clear, so that other scientists can validate and build on it.

Therefore, NIH should consider the following to support the foundation of scientific exploration:

Evaluate high-risk, high-reward research over longer time frames rather than short productivity cycles and protect such programs from metrics that penalize delayed or uncertain returns. Review panels should be explicitly instructed that absence of near-term results is not evidence of failure in exploratory research programs.

Establish a recognition mechanism modeled on the Golden Goose Award that NIH administers or co-sponsors, identifying and celebrating federally funded research that produced unexpected or long-delayed breakthroughs.

Credit the publication of negative and null results and of replication studies as legitimate scientific contributions, recognizing that these outputs are the record through which foundational science accumulates over time.

Sustained federal funding is what makes this kind of research possible in the first place. NIH should make clear in its assessment framework that investing in research before its value is obvious is not a risk to be avoided, but a necessity for scientific discovery.

7. Public Impact: Publishing is the mechanism by which taxpayer investment becomes lasting public benefit: it converts funded research into a transparent, enduring, and accessible record that patients, clinicians, educators, and innovators can rely on. The returns are substantial: by widely cited estimates, every $1 invested by NIH yields roughly $2.56 in economic activity, and it is the published, curated record that makes those returns durable and verifiable.

Public trust is now inseparable from the reliability of that record, which increasingly feeds the AI systems Americans use every day. When an American turns to an AI tool to understand a family member's diagnosis and possible treatments, the value of the answer depends on whether the underlying model draws on the gold-standard, peer-reviewed content that Wiley and its society partners produce and curate. The publisher's integrity screening, corrections, and provenance tracking are, in this sense, public-trust infrastructure. Accessibility matters as well: the maintained Version of Record, together with public-access compliance, ensures that findings reach practitioners and the public in usable, citable form.

NIH should consider the following approaches to assess and recognize the impact of biomedical research:

Measure the accessibility and sustained reuse of the scholarly record and the availability of findings in PubMed Central and other trusted repositories.

Use responsible altmetrics, including media coverage, policy citations, and public engagement, as supplementary signals of broader reach. Publishers already collect this data through existing infrastructure: for example, NIH-funded papers published by Wiley generated more than 10,000 social media mentions and over 1,500 news articles worldwide in 2024. NIH should partner with publishers to leverage this data rather than building parallel collection systems.

Recognize publishers as part of the public-trust infrastructure for biomedical AI. NIH should require that AI systems used in research and clinical decisions are built on licensed, provenance-rich, peer-reviewed content rather than unverified inputs.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Wiley-Response-to-NIH-RFI-Measuring-and-Rewarding-Scientific-Impact.pdf

Description: Full Response

WILEY

Wiley Response to the National Institutes of Health Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)

Notice Number: NOT-OD-26-087

Rigor and Reproducibility. Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings.

Wiley is the largest U.S. scientific publisher and the largest scientific society partner publisher, as well as a global leader in research and learning. Wiley's story begins in 1807, when Charles Wiley opened a small print shop on Reade Street in New York City. In our young republic's era of invention, the shop became a conduit for pioneering American ideas, first through literature and then through technical, medical, and scientific works. After the Civil War, William H. Wiley, a former U.S. Representative oriented the family company toward the applied sciences that powered American industrial growth by putting practical, reliable knowledge into the hands of engineers, physicians, and educators. When scholarship moved online, Wiley modernized again, first to digitized journals and reference works, and more recently to integrated platforms that combine submission, screening, peer review, and the permanent curation and preservation of the Version of Record. Over more than two centuries we have published the work of U.S. Nobel Prize winners and partnered with hundreds of U.S. scientific societies, supporting thousands of federally funded authors each year.

That 200-year vantage point—spanning the full lifecycle of federally funded research, from submission and peer review through correction, preservation, and reuse—shapes how we approach NIH's questions on rigor and reproducibility. Reproducibility is not a property of a single experiment; it is a property of a research record reliable enough for others to test, correct, and build upon. As the steward of that record, the publisher sits at the center of any effort to measure and reward reproducible science.

The Version of Record (VoR) is the authoritative, final, and citable form of a scholarly work, reflecting completed peer review, verified methods and data-availability statements, disclosed conflicts of interest and funding, and any subsequent corrections or retractions. It is the only version guaranteed to be checked for accuracy and curated by experts over time, and it is what allows a finding to be located, interpreted, and independently reproduced years after publication. This is the publisher's core contribution to rigor: the peer review, integrity screening, metadata stewardship, and post-publication correction that make a result trustworthy enough to reproduce. As NIH designs indicators of reproducibility, it should treat the permanence and integrity of this record as foundational rather than ancillary, consistent with the tenets of Gold Standard Science articulated in Executive Order 14303.

To strengthen NIH's framework for rigor and reproducibility, we recommend the following approaches:

- Recognize the Version of Record as the authoritative basis for reproducibility assessment. Preprints and other non-peer-reviewed outputs play a useful role in rapid communication, but they are not substitutes for the peer-reviewed record and should not be treated as equivalent in metrics frameworks.

- Credit negative and null results in funding and career-advancement review. Well-conducted studies that find no effect are essential to an honest and complete literature; their absence from the record is a root cause of the reproducibility problem.
- Rely on publisher-led integrity infrastructure, including peer review, editorial assessment, research-integrity screening, corrections, retractions, and post-publication stewardship, as part of the reproducibility ecosystem rather than treating these as separate from NIH's priorities.
- Recognize long-term and programmatic research contributions rather than rewarding the pace of publication. Metrics that prize rapid output can push researchers toward premature claims; NIH should credit sustained, careful work that genuine reproducibility requires.
- Account for disciplinary variation, since replication looks different across fields. A flexible framework that adapts to field-specific norms and methods will be more effective than applying a single standard to all research.

Wiley and the broader publishing community welcome opportunities to collaborate with NIH on the standards, infrastructure, and incentives that make reproducible science possible.

Training and Mentorship. NIH recognizes that advancing its mission requires promoting the growth and stability of the biomedical research workforce. While NIH supports training and protects time through existing policies and programs, additional indicators are needed to assess and quantify the impact of these efforts.

Advancing NIH's mission requires a stable and growing biomedical research workforce, supported by stronger measures of the impact of training and mentorship. In our experience across the U.S. research ecosystem, mentorship is not a supplementary benefit but the primary mechanism by which the tacit skills of rigorous science are transmitted from one generation to the next. Long-horizon discovery depends on giving young researchers the time, resources, and guidance to learn from lead researchers on projects that may not yield results for a decade or more.

Federal laboratories and NIH-supported training environments are uniquely positioned to serve as national training grounds, not only for technical excellence but for the standards of rigor, reproducibility, and responsible conduct that underpin Gold Standard Science. Protecting mentored time on long-duration projects is essential to sustaining the discovery pipeline and delivering public-health returns years after the initial investment. As Wiley noted in its response to the U.S. Department of Energy's Genesis Mission RFI, the durability of the U.S. research workforce also depends on distributing training opportunity broadly across institutions and regions. NIH should ensure that training and mentorship indicators capture participation from R1, R2, and R3 universities, Historically Black Colleges and Universities, Tribal Colleges, and community colleges, so that emerging talent is developed wherever it is found rather than only at the best-resourced institutions.

To assess and reward training and mentorship impact, NIH should consider the following approaches:

- Track the longitudinal trajectories of researchers who receive mentored training: whether they go on to win independent funding, publish as principal investigators, establish productive laboratories, and remain in the biomedical workforce.

- Recognize peer review and editorial participation as a form of mentorship and professional development. Serving as a reviewer or early-career editor is one of the primary ways scientists learn to evaluate evidence, judge methodology, and communicate uncertainty.
- Support structured mentorship programs that can be delivered in both in-person and hybrid formats to extend access to trainees who lack proximity to major research centers, strengthening geographic resilience in the national talent pipeline.

Collaboration. As biomedical research becomes increasingly interdisciplinary and convergent, collaboration is growing in both scale and complexity. NIH seeks new approaches for assessing and recognizing contributions to team-based science.

Biomedical progress increasingly depends on large, interdisciplinary teams working across institutions, sectors, and Federal agencies. However, traditional measures of scientific achievement, such as author position, publication counts, and citation totals, do not always capture the full range of contributions required to conduct complex collaborative research. Wiley has advocated for more comprehensive approaches to recognizing team-based science in prior submissions, including our response to the U.S. Department of Energy's Genesis Mission RFI, and we encourage NIH to adopt a framework that reflects the true architecture of modern biomedical collaboration.

Publishers can support this effort by preserving a clear and reliable record of who contributed to a study and how. Publication records can document contributions beyond writing and lead authorship, including data collection and curation, methodology, software development, project coordination, statistical analysis, mentoring, and research-resource sharing. This infrastructure already exists and can serve as the foundation for any NIH framework for recognizing collaborative contributions.

NIH should consider the following to assess and recognize collaboration:

- Recognize outputs beyond traditional journal articles. Data, software, research methods, shared infrastructure, and reproducibility work can all provide substantial value to a collaborative project and should be considered when evaluating investigators and research teams.
- Coordinate with other Federal research agencies and with publishers on common principles for documenting and attributing research contributions, using persistent identifiers such as ORCID and DOIs to link researchers to their outputs across institutions and over time.
- Adopt the CRediT taxonomy, which attributes individual researcher contributions across the research pipeline in a standardized, machine-readable format at submission.
- Assign weight to team-level and program-level outcomes in funding review, particularly for large-scale, inter-agency, or public-private initiatives where success depends on coordinated effort rather than the work of a single investigator.
- Build safeguards against unintended incentives such as inflated authorship or exaggerated contribution claims, by relying on transparent contribution statements that are preserved as part of the published Version of Record and subject to post-publication correction.

A more complete approach to evaluating team science would allow NIH to recognize the researchers who generate, organize, validate, and translate biomedical knowledge, not only those whose names appear in the most prominent author positions.

Entrepreneurship and Translation. To advance the translation of discoveries into products for people, NIH fosters the maturation of discoveries from bench to bedside and promotes innovation across sectors. More can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission.

The translation of discovery from bench to bedside depends on a scientific record that industry, investors, clinicians, and regulators can trust, and that record is the publisher's product. The peer-reviewed Version of Record, with its verified methods, disclosures, and corrections, is the defensible foundation on which patents are filed, products are developed, and clinical decisions are made. A discovery that cannot be located, cited, and relied upon with confidence cannot be translated, however promising it may be.

The publisher's role in translation is therefore not incidental but enabling. Wiley's investments in metadata, persistent identifiers, taxonomy, and integrity screening allow a finding to be discovered and built upon by a start-up or research-and-development team years after it is first published. As AI accelerates the translational pipeline, rights-cleared and provenance-rich content becomes even more critical: products and regulated systems built on unverified or unlawfully obtained inputs carry commercial, regulatory, and litigation risk that deters investment. The Wiley AI Gateway demonstrates how licensed, peer-reviewed content can be embedded in AI-enabled discovery tools, for tasks such as biomarker discovery and clinical trial design, with full provenance and attribution, accelerating translation while preserving research integrity.

To assess and reward translational and entrepreneurial impact, NIH should consider these approaches:

- Track patents and regulatory submissions that cite federally funded publications as a concrete, verifiable indicator of translational impact, and include these in the criteria used to evaluate investigators and programs.
- Count licensing agreements and start-up formation arising from NIH-funded research as recognized outputs, with appropriate attribution to the researchers, institutions, and publishers whose work enabled them.
- Monitor incorporation of NIH-funded findings into clinical guidelines, standards of care, and public health policy, recognizing that these outcomes often represent the most direct translation of research into health benefit.
- Encourage Federal procurement and partnership mechanisms to favor AI systems and analytical tools built on licensed, verified, rights-cleared research inputs, providing industry the legal certainty it needs to invest in translation and ensuring that federally supported innovation rests on defensible scientific foundations.

Foundational Scientific Exploration. Basic research and high-risk, high-reward initiatives represent foundational scientific exploration with an inherent risk of failure but the potential for paradigm-shifting breakthroughs. Further embedding these approaches within the academic enterprise is essential to ensuring continued support for bold and innovative research directions.

Basic research and high-risk, high-reward inquiry carry an inherent risk of failure alongside the potential for paradigm-shifting breakthroughs. Wiley is proud to sponsor the Golden Goose Award, which recognizes research that appeared obscure, or even accidental, that produced significant benefits for public health and the economy. The award illustrates a principle that NIH's own mission embodies: the value of foundational science is often not visible at the outset and cannot be judged by short-horizon metrics. When research is evaluated only by how quickly it produces results or how many citations it earns, the slow, uncertain, exploratory work that leads to the most important discoveries tends to get left behind.

Publishers sustain this kind of exploration by preserving the complete record of inquiry, including incremental findings, detailed methods, and the negative and null results that keep a field honest and prevent duplicated effort. The permanence of the Version of Record and long-term archiving mean that a finding regarded as obscure today remains discoverable years later, when its significance becomes clear, so that other scientists can validate and build on it.

Therefore, NIH should consider the following to support the foundation of scientific exploration:

- Evaluate high-risk, high-reward research over longer time frames rather than short productivity cycles and protect such programs from metrics that penalize delayed or uncertain returns. Review panels should be explicitly instructed that absence of near-term results is not evidence of failure in exploratory research programs.
- Establish a recognition mechanism modeled on the Golden Goose Award that NIH administers or co-sponsors, identifying and celebrating federally funded research that produced unexpected or long-delayed breakthroughs.
- Credit the publication of negative and null results and of replication studies as legitimate scientific contributions, recognizing that these outputs are the record through which foundational science accumulates over time.
- Sustained federal funding is what makes this kind of research possible in the first place. NIH should make clear in its assessment framework that investing in research before its value is obvious is not a risk to be avoided, but a necessity for scientific discovery.

Public Impact. NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.

Publishing is the mechanism by which taxpayer investment becomes lasting public benefit: it converts funded research into a transparent, enduring, and accessible record that patients, clinicians, educators, and innovators can rely on. The returns are substantial: by widely cited estimates, every $1 invested by NIH yields roughly $2.56 in economic activity, and it is the published, curated record that makes those returns durable and verifiable.

Public trust is now inseparable from the reliability of that record, which increasingly feeds the AI systems Americans use every day. When an American turns to an AI tool to understand a family member's diagnosis and possible treatments, the value of the answer depends on whether the underlying model draws on the

gold-standard, peer-reviewed content that Wiley and its society partners produce and curate. The publisher's integrity screening, corrections, and provenance tracking are, in this sense, public-trust infrastructure. Accessibility matters as well: the maintained Version of Record, together with public-access compliance, ensures that findings reach practitioners and the public in usable, citable form.

NIH should consider the following approaches to assess and recognize the impact of biomedical research:

- Measure the accessibility and sustained reuse of the scholarly record and the availability of findings in PubMed Central and other trusted repositories.
- Use responsible altmetrics, including media coverage, policy citations, and public engagement, as supplementary signals of broader reach. Publishers already collect this data through existing infrastructure: for example, NIH-funded papers published by Wiley generated more than 10,000 social media mentions and over 1,500 news articles worldwide in 2024. NIH should partner with publishers to leverage this data rather than building parallel collection systems.
- Recognize publishers as part of the public-trust infrastructure for biomedical AI. NIH should require that AI systems used in research and clinical decisions are built on licensed, provenance-rich, peer-reviewed content rather than unverified inputs.

252. University of California, San Francisco | Academic Senate Committee on Library and Scholarly Communication and University Library

I am responding: On behalf of an organization

Name: Michael C. Oldham & Chris Shaffer

Name of Organization: University of California, San Francisco | Academic Senate Committee on Library and Scholarly Communication and University Library

Type of Organization: Not Applicable

Role: Other

Role - Other: Chair, Committee on Library and Scholarly Communication (Oldham) and University Librarian (Shaffer)

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Biomedical research increasingly depends on complex datasets, software, and computational workflows that cannot be fully reproduced from publications alone. NIH's Replication and Reproducibility Initiative presents an opportunity to redefine reproducibility around persistent, reusable research objects. Community-governed metadata standards, persistent identifiers, provenance, executable workflows, and interoperable packaging would allow analyses to be reproduced directly rather than reconstructed from narrative methods. Reproducibility should be viewed not as the endpoint of validation, but as the prerequisite for continuous optimization of scientific methods. NIH can accelerate this transition by investing in shared infrastructure while recognizing reproducible datasets, software, and workflows as first-class scholarly outputs.

Actions or initiatives to support and reward rigor and reproducibility:
- Require researchers to openly share all relevant research outputs, including research methods, protocols, and code; and to cite resources and activities using platforms like Research Resource Identification (RRID) (https://www.rrids.org/) and Research Activity Identifer (RAiD) (https://www.raid.org/) in all relevant outputs. Platforms should be identified and prioritized to ensure standardization amongst researchers (see PID-Optimized Research Lifecyle (https://doi.org/10.5281/zenodo.4991732)).
- Adopt or create a platform for NIH-funded researchers to share null and negative results at no cost, following a standardized template. Such a platform would benefit reproducibility efforts and provide a record of funded research results.

2. Data, Software, and Model Sharing: NIH has successfully established expectations for data sharing. The next step is rewarding quality rather than deposition alone. Curated datasets, software, and workflows should be standardized, richly annotated, uniquely identifiable, interoperable, and reusable. Shared community infrastructure – including metadata standards, registries, persistent identifiers, and governance – should be sustained centrally, while investigators build reusable scientific assets upon this

foundation. Assessment should emphasize reuse, interoperability, reproducibility, and downstream scientific influence rather than simple download counts.

Actions or initiatives to support and reward data, software, and model sharing:
- Support the Make Data Count (MDC) initiative (https://makedatacount.org/), implement interoperable MDC practices in NIH data repositories, and promote these practices in external databases to enable standardized, measurable indicators of data usage and interoperability.

3. Training and Mentorship: Future investigators require training in data stewardship, metadata harmonization, computational reproducibility, and collaborative software development alongside training in experimental science. NIH can recognize mentors who build laboratories that consistently produce reusable scientific assets and who prepare trainees to participate in open, reproducible, data-intensive research. NIH can also provide dedicated training on these topics through the Responsible Conduct of Research requirement and other channels.

Actions or initiatives to support and reward training and mentorship:
- The Researcher Impact Framework (https://www.tara.tcd.ie/tara8/server/api/core/bitstreams/ea53964f-1f06-459d-ba0d-4f4fab9352c8/content)from Trinity College in Dublin provides a model to help researchers cite their relevant scholarly activity and write an outcome statement and impact narrative for the impact area, Development of Individuals and Collaborations.

4. Collaboration: Modern biomedical science depends on interdisciplinary teams whose contributions extend far beyond manuscript authorship. Persistent identifiers and contributor taxonomies should be used to recognize contributions to datasets, software, workflows, metadata standards, and shared infrastructure. Assessment systems should value the creation and maintenance of reusable community resources alongside traditional publications.

Actions or initiatives to support and reward collaboration:
- Encourage recognition of all research outputs and those who contribute to them in grant applications and reports, including statements by awardees addressing the impact of these contributions to the research project. These contributors may include research staff and community members. For example, the HiddenREF (https://hidden-ref.org/) initiative in the UK hosts a competition and festival to build evidence about assessing research outputs and to develop a better understanding of inclusive approaches to research recognition.
- Researchers can demonstrate their collaborations and disciplinary focuses using free tools such as Google Scholar Profiles and campus profiling systems, as well as subscription resources such as Web of Science and Dimensions Analytics.
- Researchers' organizations can also help drive collaboration and team science approaches. Providing opportunities in grant applications and reports for researchers to cite team science-oriented groups and resources within their organization, and how they have been utilized in the research project to enhance collaboration, would emphasize their importance and provide evidence of their impact.

5. Entrepreneurship and Translation: Successful translation and entrepreneurship depend upon reproducible and interoperable scientific knowledge. Standardized datasets and executable workflows reduce barriers for biotechnology, regulatory science, and clinical implementation by making research

outputs easier to validate, compare, integrate, and extend. Investments in reusable research infrastructure, therefore, accelerate both innovation and translation.

Actions or initiatives to support and reward entrepreneurship and translation:

- Contributions to innovation and entrepreneurship through patents, clinical and research products on the market, and entrepreneurial endeavors contribute towards a researcher's impact just as publications and other research outputs. A narrative description of these contributions is a fitting representation, as well as citations to their research outputs from patents, clinical guidelines, and policy documents. Citations to a researcher's outputs from others' patents, clinical trials, and policy documents may also demonstrate the impact of the researcher's work.
- Researchers may use free tools such as the U.S. patents database (https://www.uspto.gov/patents/search) and free Altmetric records (https://help.altmetric.com/en/categories/2060609-sources-of-attention) for publication citations, as well as subscription sources such as Dimensions Analytics, Web of Science, and Altmetric Explorer for publication citations.
- The Researcher Impact Framework (https://www.tara.tcd.ie/tara8/server/api/core/bitstreams/ea53964f-1f06-459d-ba0d-4f4fab9352c8/content) from Trinity College in Dublin provides a model to help researchers cite their relevant scholarly activity and write an outcome statement and impact narrative for the respective impact area, Generation of Knowledge.

6. Foundational Scientific Exploration: High-risk research often produces enduring datasets, analytical methods, and software even when the original biological hypothesis is not confirmed. Assessment systems should recognize these reusable assets because they frequently become the foundation for future discoveries. Benchmark datasets and community workflows also enable continuing methodological innovation.

7. Public Impact: Scientific impact should increasingly be measured not by where investigators publish, but by the degree to which durable research objects can be discovered, reused, reproduced, validated, and built upon to improve health, accelerate discovery, strengthen public trust, and increase research efficiency. Whereas publications enable period reporting of scientific knowledge, durable research objects enable cumulative scientific progress by fostering continuous optimization of scientific knowledge.

Actions or initiatives to support and reward entrepreneurship and translation:
- Researcher contributions to public impact can be assessed through a number of methods. Attention to research articles on social media, blogs, Wikipedia, and in the news can be tracked and measured via publisher webpages and through services such as Altmetric and PlumX Metrics. While these services tally counts, the researcher still needs to tell the impact story of the public attention their work has received.
- Databases such as Dimensions and Scopus Author Profiles map publications to corresponding United Nations Sustainable Development Goals (SDG). These goals, such as No Hunger; Good Health and Wellbeing; and Gender Equality, when applied to publications, reflect a researcher's focus on work that impacts society.

8. Other Areas: Better measures of scientific impact require better scientific infrastructure. NIH is uniquely positioned to invest in shared national resources—such as metadata standards, registries, persistent identifiers, and interoperability frameworks—that individual projects cannot sustainably create. These public goods would enable responsible assessment based on creation, quality, dissemination, reuse, scientific influence, and public impact while supporting a long-term transition from a publication-centered enterprise to a cumulative knowledge ecosystem. Standardized datasets, software, and computational workflows can become measurable scholarly contributions once they are accompanied by common metadata standards, persistent identifiers, and infrastructure that tracks discovery, reuse, and reproducibility.

Other actions or initiatives:
- Metrics for article publication citations can be useful when used responsibly. A move towards field-normalized scoring systems, such as the Relative Citation Ratio (RCR) at the article level (instead of the ranking for the journal where the article was published), should be emphasized. Field-normalized scores simplify score interpretations and ensure that articles are ranked in comparison to their field of research.
- Diverse forms of contributions should be supported, so that a social sciences researcher's activities can be accommodated and evaluated just as the basic sciences researcher's are.
- Using metrics alone for any type of research output is insufficient for assessing the impact of the research. Researchers should be expected to accompany counts and rankings with descriptions of the work's impact.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/UCSF-COLASC-Library-response-to-NIH-RFI-on-Measuring-Scientific-Impact-Final.pdf

Description: UCSF-COLASC-Library response to NIH RFI on Measuring Scientific Impact

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO

SANTA BARBARA • SANTA CRUZ

UCSF Academic Senate

UCSF Library

August 18, 2026

Re: NIH Request for Information on Measuring and Rewarding Scientific Impact

To Whom It May Concern,

Pursuant to NIH Notice Number NOT-OD-26-087 (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html), "Request for Information on Measuring and Rewarding Scientific Impact", the UCSF Academic Senate Committee on Library and Scholarly Communication (COLASC) and the UCSF Library have prepared a joint response, which is included below. We thank you for drawing attention to this critical topic and the implications it carries for the future of biomedical research.

Sincerely,

[redacted]

Michael C. Oldham, PhD
Professor, Department of Neurological Surgery
Chair, Academic Senate Committee on Library and Scholarly Communication
University of California, San Francisco

[redacted]

Chris Shaffer, MS, FMLA, AHIP
University Librarian and Assistant Vice Chancellor
Adjunct Professor, Department of Medicine
University of California, San Francisco

Response to NIH Notice Number NOT-OD-26-087: Request for Information on Measuring and Rewarding Scientific Impact

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-087.html

Rigor and Reproducibility

- Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings.

Biomedical research increasingly depends on complex datasets, software, and computational workflows that cannot be fully reproduced from publications alone. NIH's Replication and Reproducibility Initiative presents an opportunity to redefine reproducibility around persistent, reusable research objects. Community-governed metadata standards, persistent identifiers, provenance, executable workflows, and interoperable packaging would allow analyses to be reproduced directly rather than reconstructed from narrative methods. Reproducibility should be viewed not as the endpoint of validation, but as the prerequisite for continuous optimization of scientific methods. NIH can accelerate this transition by investing in shared infrastructure while recognizing reproducible datasets, software, and workflows as first-class scholarly outputs.

Actions or initiatives to support and reward rigor and reproducibility:

- Require researchers to openly share all relevant research outputs, including research methods, protocols, and code; and to cite resources and activities using platforms like Research Resource Identification (RRID) (https://www.rrids.org/) and Research Activity Identifer (RAiD) (https://www.raid.org/) in all relevant outputs. Platforms should be identified and prioritized to ensure standardization amongst researchers (see PID-Optimized Research Lifecyle (https://doi.org/10.5281/zenodo.4991732)).
- Adopt or create a platform for NIH-funded researchers to share null and negative results at no cost, following a standardized template. Such a platform would benefit reproducibility efforts and provide a record of funded research results.

Data, Software, and Model Sharing

- Sharing scientific data and tools accelerates biomedical research discovery, in part, by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies. NIH seeks to incentivize and reward these activities to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.

NIH has successfully established expectations for data sharing. The next step is rewarding quality rather than deposition alone. Curated datasets, software, and workflows should be standardized, richly annotated, uniquely identifiable, interoperable, and reusable. Shared community infrastructure – including metadata standards, registries, persistent identifiers, and governance – should be sustained centrally, while investigators build reusable scientific assets upon this foundation. Assessment should emphasize reuse, interoperability, reproducibility, and downstream scientific influence rather than simple download counts.

Actions or initiatives to support and reward data, software, and model sharing:

- Support the Make Data Count (MDC) initiative (https://makedatacount.org/), implement interoperable MDC practices in NIH data repositories, and promote these practices in external databases to enable standardized, measurable indicators of data usage and interoperability.

Training and Mentorship

- NIH recognizes that advancing its mission requires promoting the growth and stability of the biomedical research workforce. While NIH supports training and protects time through existing policies and programs, additional indicators are needed to assess and quantify the impact of these efforts.

Future investigators require training in data stewardship, metadata harmonization, computational reproducibility, and collaborative software development alongside training in experimental science. NIH can recognize mentors who build laboratories that consistently produce reusable scientific assets and who prepare trainees to participate in open, reproducible, data-intensive research. NIH can also provide dedicated training on these topics through the Responsible Conduct of Research requirement and other channels.

Actions or initiatives to support and reward training and mentorship:

- The Researcher Impact Framework (https://www.tara.tcd.ie/tara8/server/api/core/bitstreams/ea53964f-1f06-459d-ba0d-4f4fab9352c8/content)from Trinity College in Dublin provides a model to help researchers cite their relevant scholarly activity and write an outcome statement and impact narrative for the impact area, Development of Individuals and Collaborations.

Collaboration

- As biomedical research becomes increasingly interdisciplinary and convergent, collaboration is growing in both scale and complexity. NIH seeks new approaches for assessing and recognizing contributions to team-based science.

Modern biomedical science depends on interdisciplinary teams whose contributions extend far beyond manuscript authorship. Persistent identifiers and contributor taxonomies should be used to recognize contributions to datasets, software, workflows, metadata standards, and shared infrastructure. Assessment systems should value the creation and maintenance of reusable community resources alongside traditional publications.

Actions or initiatives to support and reward collaboration:

- Encourage recognition of all research outputs and those who contribute to them in grant applications and reports, including statements by awardees addressing the impact of these contributions to the research project. These contributors may include research staff and community members. For example, the HiddenREF (https://hidden-ref.org/) initiative in the UK hosts a competition and festival to build evidence about assessing research outputs and to develop a better understanding of inclusive approaches to research recognition.
- Researchers can demonstrate their collaborations and disciplinary focuses using free tools such as Google Scholar Profiles and campus profiling systems, as well as subscription resources such as Web of Science and Dimensions Analytics.
- Researchers' organizations can also help drive collaboration and team science approaches. Providing opportunities in grant applications and reports for researchers to cite team science-oriented groups and resources within their organization, and how they have been utilized in the research project to enhance collaboration, would emphasize their importance and provide evidence of their impact.

Entrepreneurship and Translation

- To advance the translation of discoveries into products for people, NIH fosters the maturation of discoveries from bench to bedside and promotes innovation across sectors. More can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission.

Successful translation and entrepreneurship depend upon reproducible and interoperable scientific knowledge. Standardized datasets and executable workflows reduce barriers for biotechnology, regulatory science, and clinical implementation by making research outputs easier to validate, compare, integrate, and extend. Investments in reusable research infrastructure, therefore, accelerate both innovation and translation.

Actions or initiatives to support and reward entrepreneurship and translation:

- Contributions to innovation and entrepreneurship through patents, clinical and research products on the market, and entrepreneurial endeavors contribute towards a researcher's impact just as publications and other research outputs. A narrative description of these contributions is a fitting representation, as well as citations to their research outputs from patents, clinical guidelines, and policy documents. Citations to a researcher's outputs from others' patents, clinical trials, and policy documents may also demonstrate the impact of the researcher's work.
- Researchers may use free tools such as the U.S. patents database (https://www.uspto.gov/patents/search) and free Altmetric records (https://help.altmetric.com/en/categories/2060609-sources-of-attention) for publication citations, as well as subscription sources such as Dimensions Analytics, Web of Science, and Altmetric Explorer for publication citations.

- The Researcher Impact Framework (https://www.tara.tcd.ie/tara8/server/api/core/bitstreams/ea53964f-1f06-459d-ba0d-4f4fab9352c8/content) from Trinity College in Dublin provides a model to help researchers cite their relevant scholarly activity and write an outcome statement and impact narrative for the respective impact area, Generation of Knowledge.

Foundational Scientific Exploration

- Basic research and high-risk, high-reward initiatives represent foundational scientific exploration with an inherent risk of failure but the potential for paradigm-shifting breakthroughs. Further embedding these approaches within the academic enterprise is essential to ensuring continued support for bold and innovative research directions.

High-risk research often produces enduring datasets, analytical methods, and software even when the original biological hypothesis is not confirmed. Assessment systems should recognize these reusable assets because they frequently become the foundation for future discoveries. Benchmark datasets and community workflows also enable continuing methodological innovation.

Public Impact

- NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.

Scientific impact should increasingly be measured not by where investigators publish, but by the degree to which durable research objects can be discovered, reused, reproduced, validated, and built upon to improve health, accelerate discovery, strengthen public trust, and increase research efficiency. Whereas publications enable period reporting of scientific knowledge, durable research objects enable cumulative scientific progress by fostering continuous optimization of scientific knowledge.

Actions or initiatives to support and reward entrepreneurship and translation:

- Researcher contributions to public impact can be assessed through a number of methods. Attention to research articles on social media, blogs, Wikipedia, and in the news can be tracked and measured via publisher webpages and through services such as Altmetric and PlumX Metrics. While these services tally counts, the researcher still needs to tell the impact story of the public attention their work has received.
- Databases such as Dimensions and Scopus Author Profiles map publications to corresponding United Nations Sustainable Development Goals (SDG). These goals, such as No Hunger; Good Health and Wellbeing; and Gender Equality, when applied to publications, reflect a researcher's focus on work that impacts society.

Other Areas

Better measures of scientific impact require better scientific infrastructure. NIH is uniquely positioned to invest in shared national resources—such as metadata standards, registries, persistent identifiers, and interoperability frameworks—that individual projects cannot sustainably create. These public

goods would enable responsible assessment based on creation, quality, dissemination, reuse, scientific influence, and public impact while supporting a long-term transition from a publication-centered enterprise to a cumulative knowledge ecosystem. Standardized datasets, software, and computational workflows can become measurable scholarly contributions once they are accompanied by common metadata standards, persistent identifiers, and infrastructure that tracks discovery, reuse, and reproducibility.

Other actions or initiatives:

- Metrics for article publication citations can be useful when used responsibly. A move towards field-normalized scoring systems, such as the Relative Citation Ratio (RCR) at the article level (instead of the ranking for the journal where the article was published), should be emphasized. Field-normalized scores simplify score interpretations and ensure that articles are ranked in comparison to their field of research.
- Diverse forms of contributions should be supported, so that a social sciences researcher's activities can be accommodated and evaluated just as the basic sciences researcher's are.
- Using metrics alone for any type of research output is insufficient for assessing the impact of the research. Researchers should be expected to accompany counts and rankings with descriptions of the work's impact.

253. James Weatherall

I am responding: On behalf of myself

Name: James Weatherall

Name of Organization: University of California, Irvine

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: As has been widely discussed, and as acknowledged in the recently announced NIH initiative on replication and reproducibility, low rates for replication attempts reflect the current incentive structure for scientists. In most fields, scientists do not get credit for replications in the way they do for new findings, and universities do not reward scientists for replicating other studies. In many fields, there is also an asymmetry, where a successful study is perceived to reflect investigator skill, whereas a failed replication (or failed initial study) may be seen as evidence that the investigator conducting the replication is less skilled. It is challenging to directly intervene on scientific culture and biases among P&T committees, ranking organizations, and peer reviewers. But there are other incentives – most notably, large federal research grants – that influence perceived stature and provide independent financial benefits, which may affect investigator behavior. Grant RFPs specifically targeting systematic replication efforts would likely impact on the number of replications attempted, as well as the reputation of scientists who focus primarily on replicating other studies. Ultimately, this will require a significant and focused financial investment.

Additionally, a mechanism should be developed to highlight and reward especially careful replications conducted by scientists with strong reputations for excellent analysis and experimental skill. The best way to do this would be to identify or develop a new publication venues with very high editorial standards that could play a role comparable to Cell, Science, and Nature for replications.

Finally, I note that replication is valuable even in cases where replications succeed, and efforts to support replication should not focus entirely on areas that are, for one reason or another, perceived to be suspect or likely to have replication attempts fail. Failed replications, and incorrect findings, should also be recognized as a normal part of science, and the choice of whether to attempt replication within some field should not be presented, or treated, as a potential indictment of results or researchers in that field. Instead, the NIH should consider an inductive risk approach, where the primary consideration behind replication efforts is an assessment of the overall strength of existing evidence along with the potential harms of adopting policies or practices based on incorrect findings. Emphasis should also be placed on publicly-funded, highly transparent replications of high-impact research previously funded by private groups with financial conflicts.

2. Data, Software, and Model Sharing: Norms around data stewardship and open data are important, and rapidly evolving. In addition to requiring data and sharing management plans, the NIH can promote

improved data practices by better auditing past practices as part of the grant proposal process, making availability of past data a criterion for assessment of future proposals, providing support for data maintenance, and providing ombudsman resources for cases where requests for data from NIH-supported research are not granted in a timely manner.

3. Training and Mentorship: The single-most important factor in building and maintaining a strong scientific workforce is funding stability and longitudinal support for early-career and untenured researchers. Funding uncertainty makes research careers less attractive, and disruptions to funding often force researchers who have already benefited from past public investments to leave the field, with minimal prospect of returning. This reduces the overall skilled workforce and it blunts the possible impact of past investment. The NIH should perform an internal audit of grant cancelations, award delays, and other funding disruptions over the past several years to identify the impact of those actions on the size and quality of the scientific workforce, especially by age cohort, to identify ways in which its own past practices have harmed its stated goals. This audit should then be the basis for new safeguards to protect the long-term stability of the scientific workforce.

4. Collaboration: There are challenging issues related to how to assess the size and importance of individual contributions to team science, and to assign credit and other resources on the basis of those contributions. I do not have specific proposals for how to do this, but I do note that similar issues arise in physics, related to very large experiments (such as particle accelerator experiments and large astronomical instruments), and that it may be possible to adapt best-practices from that context to the biomedical sciences.

5. Entrepreneurship and Translation: Translational science is crucial for realizing the public impacts of federal science funding, and it is right for the NIH to support entrepreneurship and the bench-to-bedside pipeline. Currently, federal law incentivizes institutional recipients of federal grants to commercialize federally funded research, via the Bayh-Dole act, which permits non-profits to hold patents on inventions developed with federal funds. However, there is a gap between development of the underlying technology / intellectual property to the patent stage, and further development of viable commercialization strategies that can attract private funding and ultimately reach consumers.

Moreover, the current risk/reward structure for developing private companies to commercialize research is the same as for other start-ups, where early assessments of long-term market demand determine what IP is commercialized. This does not reflect the past public investment in that IP nor the public benefits of its broad availability. Structures should be developed to permit prolonged incubation under, or in close partnership with, non-profit entities, and the NIH should provide mechanisms for early-stage venture capital funding in cases where public benefit is high but market demand may be insufficient to recover costs. Public investment in later-stage commercialization could be reflected in reduced prices, or preferred pricing for federal buyers (such as Medicare). The existing SBIR and STTR programs are good models for this, but funding levels are very low compared to private capital.

That being said, independent market mechanisms exist for rewarding successful commercialization, and the NIH should maintain its focus on basic science. Ultimately, the need for public support is greater the earlier in the development pipeline. Basic science, with no obvious commercial applications, is where public funding is most justified (as supporting a collective good) and most needed.

6. Foundational Scientific Exploration: I fully agree with the NIH's statement on this topic. One particular concern is that, while expert peer review remains the best way to assess the quality of proposals, it can select for relatively safe / conservative projects, especially in a resource-strained environment where reviewers are asked to comment on the likelihood that an award will result in specific benefits. For this reason, the NIH should consider lottery-based funding mechanisms, for high-risk proposals from researchers with an established track record who meet other criteria of quality, such as past research success.

7. Public Impact: Public Impact must be assessed holistically, with no expectation that individual projects will result in easily quantified, short-term effects. Instead, the NIH should conceive of itself as supporting a large portfolio of activities that, on the whole, will make a significant public impact over time. The history of scientific research shows that broad support for basic science will yield significant public benefits. Narrow focus on short-term indicators and individual projects, without tolerance for failure (a consequence of risk-taking) or long horizons, will ultimately reduce the public impact of the NIH's investments.

8. Other Areas: Overall, scientific researchers should be assessed holistically, with easily quantified measures like citation counts, patents, previous grants, and publications balanced against more subjective and difficult to measure qualities such as creativity, quality of past research, indirect impact, and intellectual humility. This is necessary to properly reward activities such as replications (which are less likely to be cited or published in especially high-status journals, especially when successful), commercialization (which does not show up in citations or publications), risk-taking (which may not have immediate benefits), and highly original contributions that are not immediately recognized as such. However, holistic assessment is often unstable, and can quickly collapse into either bean-counting or lowered standards. One mechanism to protect against instability is to target grant opportunities for researchers who are especially strong in valuable but difficult to quantify dimensions, and either focus evaluation of proposals on specific criteria or use lottery-based mechanisms to avoid biases.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/NIH-feedback-final.pdf

Description: Memo addressing questions with additional context

UNIVERSITY
OF
CALIFORNIA

Academic Senate

Office of the President
1111 Franklin Street
Oakland, CA 94607

senate.universityofcalifornia.edu

CAMPUSES
Berkeley
Davis
Irvine
UCLA
Merced
Riverside
San Diego
San Francisco
Santa Barbara
Santa Cruz

MEDICAL CENTERS
Davis
Irvine
UCLA
San Diego
San Francisco

NATIONAL LABORATORIES
Lawrence Berkeley
Lawrence Livermore
Los Alamos

UNIVERSITY COMMITTEE ON RESEARCH POLICY (UCORP)
James Weatherall, Chair
Email: james.owen.weatherall@uci.edu

August 19, 2026

RE: Request for Information on Measuring and Rewarding Scientific Impact

Dear Colleagues,

I write this memo in a personal capacity. For identification and contextualization purposes, I note that I am Chancellors Professor in the Department of Logic and Philosophy of Science at UC Irvine and I currently serve as Chair of the University Committee on Research Policy (UCORP), a standing committee of the systemwide Academic Senate of the University of California. In that capacity, and in my prior capacity as Chair of UC Irvine's Council on Research, Computing, and Libraries, I have had opportunities to participate in many discussions across disciplines and across UC campuses on topics related to this request for information. Additionally, my primary academic field is philosophy of science, and I have previously written and taught on issues related to science in public discourse, the credit economy of science, and scientific best-practices. This submission is partially informed by feedback from colleagues on UCORP and others in my local community whose research concerns these topics.

Rigor and Reproducibility.
As has been widely discussed, and as acknowledged in the recently announced NIH initiative on replication and reproducibility, low rates for replication attempts reflect the current incentive structure for scientists. In most fields, scientists do not get credit for replications in the way they do for new findings, and universities do not reward scientists for replicating other studies. In many fields, there is also an asymmetry, where a successful study is perceived to reflect investigator skill, whereas a failed replication (or failed initial study) may be seen as evidence that the investigator conducting the replication is less skilled. It is challenging to directly intervene on scientific culture and biases among P&T committees, ranking organizations, and peer reviewers. But there are other incentives – most notably, large federal research grants – that influence perceived stature and provide independent financial benefits, which may affect investigator behavior. Grant RFPs specifically targeting systematic replication efforts would likely impact on the number of replications attempted,

as well as the reputation of scientists who focus primarily on replicating other studies. Ultimately, this will require a significant and focused financial investment.

Additionally, a mechanism should be developed to highlight and reward especially careful replications conducted by scientists with strong reputations for excellent analysis and experimental skill. The best way to do this would be to identify or develop a new publication venues with very high editorial standards that could play a role comparable to Cell, Science, and Nature for replications.

Finally, I note that replication is valuable even in cases where replications succeed, and efforts to support replication should not focus entirely on areas that are, for one reason or another, perceived to be suspect or likely to have replication attempts fail. Failed replications, and incorrect findings, should also be recognized as a normal part of science, and the choice of whether to attempt replication within some field should not be presented, or treated, as a potential indictment of results or researchers in that field. Instead, the NIH should consider an inductive risk approach, where the primary consideration behind replication efforts is an assessment of the overall strength of existing evidence along with the potential harms of adopting policies or practices based on incorrect findings. Emphasis should also be placed on publicly-funded, highly transparent replications of high-impact research previously funded by private groups with financial conflicts.

Data, Software, and Model Sharing.
Norms around data stewardship and open data are important, and rapidly evolving. In addition to requiring data and sharing management plans, the NIH can promote improved data practices by better auditing past practices as part of the grant proposal process, making availability of past data a criterion for assessment of future proposals, providing support for data maintenance, and providing ombudsman resources for cases where requests for data from NIH-supported research are not granted in a timely manner.

Training and Mentorship.
The single-most important factor in building and maintaining a strong scientific workforce is funding stability and longitudinal support for early-career and untenured researchers. Funding uncertainty makes research careers less attractive, and disruptions to funding often force researchers who have already benefited from past public investments to leave the field, with minimal prospect of returning. This reduces the overall skilled workforce and it blunts the possible impact of past investment. The NIH should perform an internal audit of grant cancelations, award delays, and other funding disruptions over the past several years to identify the impact of those actions on the size and quality of the scientific workforce, especially by age cohort, to identify ways in which its own past practices have harmed its stated goals. This audit should then be the basis for new safeguards to protect the long-term stability of the scientific workforce.

Collaboration.
There are challenging issues related to how to assess the size and importance of individual contributions to team science, and to assign credit and other resources on the basis of those contributions. I do not have specific proposals for how to do this, but I do note that similar issues arise in physics, related to very large experiments (such as particle accelerator experiments and large astronomical instruments), and that it may be possible to adapt best-practices from that context to the biomedical sciences.

Entrepreneurship and Translation.
Translational science is crucial for realizing the public impacts of federal science funding, and it is right for

the NIH to support entrepreneurship and the bench-to-bedside pipeline. Currently, federal law incentivizes institutional recipients of federal grants to commercialize federally funded research, via the Bayh-Dole act, which permits non-profits to hold patents on inventions developed with federal funds. However, there is a gap between development of the underlying technology / intellectual property to the patent stage, and further development of viable commercialization strategies that can attract private funding and ultimately reach consumers.

Moreover, the current risk/reward structure for developing private companies to commercialize research is the same as for other start-ups, where early assessments of long-term market demand determine what IP is commercialized. This does not reflect the past public investment in that IP nor the public benefits of its broad availability. Structures should be developed to permit prolonged incubation under, or in close partnership with, non-profit entities, and the NIH should provide mechanisms for early-stage venture capital funding in cases where public benefit is high but market demand may be insufficient to recover costs. Public investment in later-stage commercialization could be reflected in reduced prices, or preferred pricing for federal buyers (such as Medicare). The existing SBIR and STTR programs are good models for this, but funding levels are very low compared to private capital.

That being said, independent market mechanisms exist for rewarding successful commercialization, and the NIH should maintain its focus on basic science. Ultimately, the need for public support is greater the earlier in the development pipeline. Basic science, with no obvious commercial applications, is where public funding is most justified (as supporting a collective good) and most needed.

Foundational Scientific Exploration.
I fully agree with the NIH's statement on this topic. One particular concern is that, while expert peer review remains the best way to assess the quality of proposals, it can select for relatively safe / conservative projects, especially in a resource-strained environment where reviewers are asked to comment on the likelihood that an award will result in specific benefits. For this reason, the NIH should consider lottery-based funding mechanisms, for high-risk proposals from researchers with an established track record who meet other criteria of quality, such as past research success.

Public Impact.
Public Impact must be assessed holistically, with no expectation that individual projects will result in easily quantified, short-term effects. Instead, the NIH should conceive of itself as supporting a large portfolio of activities that, on the whole, will make a significant public impact over time. The history of scientific research shows that broad support for basic science will yield significant public benefits. Narrow focus on short-term indicators and individual projects, without tolerance for failure (a consequence of risk-taking) or long horizons, will ultimately reduce the public impact of the NIH's investments.

Other Areas.
Overall, scientific researchers should be assessed holistically, with easily quantified measures like citation counts, patents, previous grants, and publications balanced against more subjective and difficult to measure qualities such as creativity, quality of past research, indirect impact, and intellectual humility. This is necessary to properly reward activities such as replications (which are less likely to be cited or published in especially high-status journals, especially when successful), commercialization (which does not show up in citations or publications), risk-taking (which may not have immediate benefits), and highly original contributions that are not immediately recognized as such. However, holistic assessment is often

unstable, and can quickly collapse into either bean-counting or lowered standards. One mechanism to protect against instability is to target grant opportunities for researchers who are especially strong in valuable but difficult to quantify dimensions, and either focus evaluation of proposals on specific criteria or use lottery-based mechanisms to avoid biases.

Thank you for the opportunity to comment.

Sincerely,



Chair, University Committee on Research Policy

Chancellor's Professor and Department Chair
Logic and Philosophy of Science
University of California, Irvine

254. University of Illinois Urbana-Champaign

I am responding: On behalf of an organization

Name: Susan Martinis

Name of Organization: University of Illinois Urbana-Champaign

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: As much as feasible, biomedical research should rely on accepted and standardized methods for experimental design. Such methods should be developed by experts within disciplines, as a partnership between educational institutions and the private sector; these efforts are often coordinated through professional societies. This will enhance reproducibility by standardizing methods and data handling, through supporting disciplinary norms and best practices.

It will be essential for NIH to recognize and reward investments in open infrastructure. Local repositories make outputs like articles, data, and code openly available, with persistent identifiers. The University of Illinois Urbana-Champaign (Illinois) has invested in such infrastructure, including our institutional repository IDEALS and the Illinois Data Bank. These are two examples of efficient and open means to make federally funded research available. They are critical infrastructure, not just in fields where no readily available disciplinary repositories exist. For example, they support reproducibility where the costs of open access publication in commercial journals are prohibitive (and becoming more so). Assessment models that measure and reward the impact of research distributed through these open access repositories can and should be part of processes that assess impact and support rigor.

As a related point, the winners of NIH's recent S-index competition demonstrate growing interest in developing better measures of data-sharing value and impact. NIH should build on this momentum by developing richer measures of dataset value. Download counts are increasingly difficult to interpret as AI agents and automated systems generate substantial repository traffic. Metrics should capture citations, reuse, downstream publications and software/models, FAIRness, and other evidence of scientific impact—and should recognize the institutional infrastructure and professional curation that make such reuse possible.

A supercomputing center, like the National Center for Supercomputing Applications (NCSA), can support reproducible and replicable biomedical approaches by providing version-controlled code, documented computational environments, reproducible/archived workflows and pipelines, containerized software, and persistent storage. These practices will help independently rerun analyses and arrive at the same result. Small research groups may be overburdened if they need to do those things at the group level. Centralized infrastructure makes these practices accessible across disciplines and institutions.

2. Data, Software, and Model Sharing: Academic library data services, such as the Illinois Research Data Service, support researchers in identifying and using the most appropriate data-sharing venue. However, NIH's data-sharing policies encompass data for which no disciplinary repository exists. Domain repositories are also notoriously difficult to sustain and fund over the long term; there will never be a repository for every data type or research area subject to NIH's policies. Universities, including Illinois, make up for this structural gap by investing in library services and infrastructure to provide long-term access, preservation, and curation for data that have no disciplinary repository home.

Principal Investigators should be incentivized to publish original data and software via public repositories, including those maintained by individual institutions. Education and Professional Development of Open Science best practices should be encouraged at all levels of education for students and faculty.

The potential benefits of data sharing include democratization of science and greater access to data. This is especially important for early career scientists or researchers with limited resources. Sharing data in an open way can engage a broader community in research and increase collaborative opportunities.

Some of the measurable indicators can be the number of datasets, software packages, models deposited in recognized repositories, number of downloads. A supercomputer center like NCSA can provide secure data storage and data engineering, mechanisms for metadata standardization, computing environments, software management, etc.

Institutional data repositories are vital infrastructure for the long-term stewardship and reuse of research data (Darragh et al., 2024, Science 385(6714):1174; DOI: 10.1126/science.adr0789). The value of institutional repositories is demonstrable at scale. For example, a dataset in the University of Illinois, Illinois Data Bank, ranks 4th among approximately 45 million datasets indexed by S-index winner, scholarly.data.io. This illustrates the reach and impact that professionally curated institutional repositories can provide and underscores why the infrastructure and expertise supporting them should be recognized as an essential part of the research enterprise.

3. Training and Mentorship: NIH should incentivize non-academic mentors for trainees, including those in non-academic institutions and industry. As part of this, NIH should develop mentor training resources for those in non-academic institutions. This will provide a broader perspective on career options and development.

Measurable indicators could be the number of researchers and students trained, repeat use of training resources, etc.. Implementation approaches should offer tiered training, from introductory computing to advanced AI & HPC workflows. This will broaden access to advanced infrastructure.

4. Collaboration: As research becomes more collaborative, it will be helpful if NIH moves away from using proxies to assess impact. Historically, these proxy measures have included publication counts and journal prestige (e.g. impact factor), which are at best directional, fuzzy indicators. They are widely recognized as poor substitutes for assessing research quality, rigor, and real-world impact. Accordingly, assessing the impact or quality of research using these metrics can incentivize poor research practices, such as rewarding volume over quality, while disincentivizing replication and the adoption of open

research practices that increase impact and rigor.

Metrics that capture citations, reuse, downstream publications and software/models, FAIRness, and other evidence of scientific impact—and the open institutional infrastructure and professional curation that make such reuse possible through library and other services—are better measures of impact.

Similarly, it will be helpful for NIH to make clear that contribution-based credit models, as opposed to author-order driven credit, reflect the very real and deep collaboration that underlies much scientific research and innovation. Contribution-based credit also better supports replication and reproducibility. Metrics should recognize different roles and levels of contribution (not just rely on # of grants or publications). For example, when researchers collaborate with a supercomputer center, the contributions from computational scientists, research software engineers, data scientists, and infrastructure experts are often overlooked in the publication of IP invention disclosures. We need to look into the number and diversity of disciplines, number of new interdisciplinary collaborations, computational/data contributions to publications and funded grants, etc.

NIH should encourage and incentivize programs for academic institutions to partner and collaborate with industry to enhance interdisciplinary innovation. As part of this, NIH should develop template collaboration guides that outline best practices for team science, authorship, data management, sharing and ownership, and handling of intellectual property. Such materials should include guidance on obtaining oversight and administrative approval.

The benefits and unintended consequences of interdisciplinary team science have been studied for large and small teams (e.g., https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0322943). The evidence shows that access to interdisciplinarity and team science improves the diversity of collaboration without impacting a researcher's productivity.

5. Entrepreneurship and Translation: NIH should encourage and incentivize programs for academic institutions to partner and collaborate with industry to enhance interdisciplinary innovation.

Metrics could include the number of invention reports and the number of interdisciplinary creators/inventors on those reports, to an office like Illinois' Office of Technology Management (OTM). OTM protects and licenses intellectual property to bring Illinois innovations to the world. One could also track computational contributions to translational projects and provide access to HPC, AI, data, and technical expertise early in the process.

6. Foundational Scientific Exploration: NIH should encourage the dissemination of all results and outcomes from scientific research, regardless of successful findings.

The Beckman Institute for Advanced Science and Technology at the University of Illinois Urbana-Champaign shows how foundational, high-risk science can be embedded in the academic enterprise. Its model — a theme-organized institute offering seed grants for exploratory interdisciplinary projects, independent postdoctoral fellowships, and shared open-access instrumentation — suggests implementation approaches NIH could scale: institute-based set-asides for no-preliminary-data seed awards; person-based fellowships freeing early-career scientists from PI-defined aims; and physical co-

location of disciplines. Measurable indicators: share of seed projects maturing into extramural awards; new cross-departmental collaborations formed; fellows' time to independent positions; publication of null results; and field-normed novelty metrics tracked alongside citations. Benefits: paradigm-shifting discovery at disciplinary interfaces, faster course correction through published failure, and retention of creative early-career scientists. Unintended consequences to guard against: "risk-washing" of incremental work, metric gaming, and concentration of such institutes at wealthy universities — NIH-funded institute set-asides for MSIs and emerging institutions are essential. Feasibility spans career stages if fellows earn tenure credit for exploratory work and senior investigators face renewal criteria tolerant of partial failure.

7. Public Impact: NIH should encourage the dissemination of academic and private sector partnerships that have led to successful development of clinical innovation, diagnoses, and health therapeutics.

8. Other Areas: N/A

255. Datavant

I am responding: On behalf of an organization

Name: Samuel Roods

Name of Organization: Datavant

Type of Organization: Other

Type of Organization - Other: Commercial Entity (Health Information Technology Company)

Role: Other

Role - Other: Director, Government Affairs & Policy

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: NIH should establish the reproducibility of patient-level data linkage as a measurable dimension of research rigor and invest in shared infrastructure that makes reproducible linkage a standard capability rather than a burden each laboratory must shoulder alone.

A frequently overlooked dimension of reproducibility in research that relies on real-world data is whether the same patient population, defined using the same criteria, can be identified consistently across datasets and research teams. Privacy-preserving record linkage (PPRL) makes this measurable. Consistent, de-identified tokens generated from standardized inputs produce linkage that is documentable, independently validatable, and reproducible across research groups, in contrast to ad hoc data pulls in which cohort construction is undocumented and cannot be replicated. This is a proven approach at national scale: Datavant processes roughly 2.2 trillion records for linkage each year and has supported more than 180 clinical trials and observational programs covering over 800,000 patients.

To institutionalize this as a discipline, NIH could require or incentivize a small set of measurable practices: documentation of linkage methodology in study protocols, including the provenance of the real-world data source and the token-generation approach; reporting of false-positive and false-negative match rates and their effect on study conclusions; validation of the linkage method through peer-reviewed publication, performance testing against known datasets, or independent expert review; and use of shared, centrally governed data enclaves built on common data models.

Infrastructure to support this already exists within NIH's own portfolio. The N3C Data Enclave and the All of Us Center for Linkage and Acquisition of Data demonstrate reproducible, harmonized, privacy-preserving linkage at scale, using linkage infrastructure that Datavant provides and governs. NIH could designate them as reference implementations and as a central locus for reproducible-linkage methods. Recognizing this infrastructure also creates a natural set of incentives: NIH could reward documented, validated linkage in application scoring, recognize teams that contribute reusable linkage methods or replication studies conducted on shared enclaves, and engage the research community in selecting data-intensive areas that are ripe for replication.

Because the cost of building linkage infrastructure falls hardest on smaller institutions and early-career investigators, providing reproducible linkage as a centralized, shared service rather than a capability each laboratory must fund is what keeps reproducibility feasible across career stages and institution types. The N3C enclave is an existing example of this shared-service model.

One caution: mandating a single method or platform would risk entrenching incumbent providers and crowding out novel approaches. NIH can avoid this by specifying the reproducibility outcomes it wants, namely documented, validated, auditable, and interoperable linkage, rather than a named technology, keeping the standard method-neutral.

2. Data, Software, and Model Sharing: NIH should reward investment in the privacy infrastructure that makes data sharing compliant and reusable, and should measure the downstream reuse of shared data rather than only its initial deposit.

The barriers to compliant data sharing are rarely philosophical; they are technical and legal. Researchers who want to share data often cannot, because they lack the means to de-identify it at sufficient quality, link it to other datasets in a privacy-compliant manner, or certify that downstream use meets HIPAA's De-Identification Standard. Incentives should therefore reward investment in privacy-enabling infrastructure, not only the act of deposit.

First, NIH should recognize HIPAA Expert Determination as a measurable indicator of data-sharing readiness. Expert Determination is the statistical, risk-based method for certifying that a dataset meets the Privacy Rule's De-Identification Standard, and datasets certified this way can be shared, reused, and linked across institutions without exposing patient identity. NIH could reward this certification as a condition of data-sharing credit. The capacity to perform it already exists in the ecosystem, including Datavant's disclosure-risk analysis practice, which certifies de-identification across dozens of jurisdictions.

Second, NIH should measure downstream reuse rather than deposit alone. A dataset that has been tokenized and linked to other datasets, enabling patient-level follow-up, external validation, or meta-analysis, represents a higher scientific contribution than one that is deposited but never accessed, and it directly serves NIH's interest in enabling validation of results. Reuse of this kind is quantifiable: token-based overlap analyses yield counts of patients matched across datasets. This is established practice; Datavant has tokenized nearly 300 clinical trials to link them to real-world data without unblinding studies or exposing participant identity.

Third, NIH should sustain shared, standardized data infrastructure and hold it up as a benchmark. The N3C model combines electronic health record, imaging, mortality, genomic, Medicare and Medicaid, and social-determinants data, all de-identified and linked through common privacy-preserving infrastructure, as part of an NIH-CMS interagency effort led by NCATS and a reference implementation for the NAIRR pilot. NIH could position it as a reference framework for cross-institutional, disease-area data sharing.

As with reproducibility, feasibility varies across institutions: Expert Determination and high-quality de-identification are unevenly distributed. To keep the incentive from rewarding only the already-capable, NIH should pair it with funded, shared de-identification and linkage services so that a researcher at a

smaller institution can meet the standard without building it alone.

Finally, two consequences to manage. A reuse metric can be gamed through superficial or low-value linkages, and a certification requirement can itself become a barrier if it is treated as a precondition rather than a supported service. NIH should combine quantitative reuse measures with qualitative review of scientific contribution, and fund the shared infrastructure rather than only mandating the standard.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: We appreciate the opportunity to comment on how to better define, measure, recognize, and incentivize the qualities of modern biomedical research, particularly on the topics of Rigor and Reproducibility as well as Data, Software, and Model Sharing. Datavant supports the Agency's efforts to create measurable indicators and incentives as part of NIH's broader Replication and Reproducibility Initiative and its Plan to Drive Gold Standard Science.

Datavant is the data collaboration platform trusted for healthcare, working with more than 80,000 hospitals and clinics, every health insurance company in the U.S., all 20 of the top 20 pharmaceutical companies, government agencies, and more than 300 real-world data partners. We are a company of health information interoperability, privacy, security, and pharmacoepidemiology experts, on a mission to make the world's health data secure, accessible, and actionable. Through our acquisition of Aetion, we also provide scalable, transparent, and scientifically rigorous analysis of real-world data (RWD), applying causal-inference and epidemiologic methods that support reproducible, auditable research.

Datavant appreciates the Agency's leadership in incentivizing and rewarding rigorous research that delivers value to the American public, and we stand ready to serve as a resource as NIH advances replication and reproducibility.

Please contact us at governmentaffairs@datavant.com.

Thank you.

Description: Attached please find comments from Datavant.

256. Noam Y Harel

I am responding: On behalf of myself

Name: Noam Y Harel

Name of Organization: James J. Peters VA Medical Center; Icahn School of Medicine at Mount Sinai

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: I write from the perspective of a PI performing investigator-initiated clinical research. I have a prior history of basic bench science, but this proposal applies mostly to translational and clinical research.

I propose that NIH lead the way toward UNTETHERING IDEA GENERATION FROM PROTOCOL EXECUTION.

Research HUBS would focus on protocol execution and reproducibility, while SPOKES focus on idea generation and pilot testing.

An Idea Spoke-Execution Hub model would counter the disincentives that drive irreproducibility:

SECRECY: Researchers currently hoard their ideas in secret until they can get funded to actually test those ideas. That day only rarely comes.

CREDIT: Academic credit currently focuses on first/senior-author papers (too easy to achieve in the swamp of crappy publishers) and grant indirects (too difficult to earn in the current funding landscape).

HOLDING THE FORT: Once academic researchers get funded for a particular line of research, they're subconsciously pressured to defend the hypotheses underlying that line of research (and the results of initial experiments) forever. Academia incentivizes us to maintain funding, not to get it right.

PART-TIME: Most academic researchers, myself included, lose research time to clinical, teaching, and/or administrative duties (not to mention frequent grant writing/editing). Less time devoted to research (especially when individual labs with varying levels of experience, resources, and time attempt to execute complicated research protocols) means less focus on research reproducibility.

Academic institutions have invested too much in the irreproducible status quo to lead the way towards necessary change. Hopefully, NIH can lead this effort, like it has done with so many other initiatives over the years.

Research Spokes would be the descendants of current academic research centers. Research Hubs would be regional centers staffed by experienced FULL-TIME RESEARCHERS with well-equipped facilities and

stable rather than cyclical funding. Kind of like the ALLEN INSTITUTE.

Rather than managing all logistical, administrative, and regulatory aspects of taking a project from idea to full translation, Spoke investigators would focus on generating ideas and early-stage testing. This would relieve financial pressure on each Spoke to obtain all the space and equipment necessary for advanced protocol execution, thereby democratizing the idea landscape to give lesser-resourced Spokes a more equal footing.

More importantly, focusing on (and giving sufficient credit for) earlier stage research would also reduce Spoke investigators' incentive to massage data and exaggerate results to survive in the publish-grant-perish model.

This system could also reduce duplicative administrative infrastructures by enabling smaller-sized Grants and Contracts Offices, Offices of Sponsored Research, etc. Therefore, the government's overall bill for indirects would decrease without requiring an overnight imposition of greatly reduced indirect rates.

Meanwhile, Hubs would be freed of the grant hamster wheel and the pressures of academic credit. Stable long-term staffing and greater focus on protocol execution would likely increase efficiency and reliability of sophisticated outcome testing.

For patient stakeholders, participation in clinical trials run by Hubs will be more likely to contribute to reliable progress (and participant travel and lodging could easily be covered by savings on administrative costs).

Obviously, this is an idealized vision. Executing it will be very disruptive and difficult.

There would be far fewer Hubs than Spokes. Academic institutions driven by indirects will obviously oppose giving up any aspects of the current research funding model towards full-time protocol execution Hubs. Conversely, a centralized network of fewer Hubs executing large government-sponsored projects will be more susceptible to political interference.

NIH needs to discuss some key issues with stakeholders:

How to fund Spokes? Whereas the process for funding Hubs seems self-evident, Idea-generators will still need funding for their time and for pilot testing. Inevitably this will be at least partially grant-driven. But ideally, grants would fund people, not projects.

Notably, this model would also push academic institutions to (finally) evolve towards giving credit for SHARING ideas rather than hoarding them in a rarely used file cabinet.

How to develop graduate students and postdocs in this landscape? Presumably, they would still be based at academic Spokes, but perhaps they could each have sponsored years at a Hub as well.

How to protect IP for ideas that eventually succeed in the real world? I am delighted to not have any

expertise or opinions on this key issue.

How, and by whom, would Spoke ideas and preliminary data be vetted for potential Hub execution? One approach could be via depositing ideas into a publicly accessible centralized proposal repository (CPR).

The CPR would function like a “Common App” for college or AMCAS for medical school. It would allow proposal submission at any time (no deadlines). One generalized format. Funders (not just NIH) could search through submissions for proposals matching their missions. If a funder(s) like(s) a proposal, they contact that PI to develop a secondary application (red tape only when necessary).

This would mean one portal for one idea (instead of repackaging ideas across multiple grant mechanisms), no deadlines, less red tape. Applicants would have a larger audience. Funders would have a larger pool of ideas from which to pick.

Unfortunately, Funders would also surely have more spam to sift through. And peer review panels do not currently work on a rolling basis. But this would open up jobs for full-time paid reviewers (hopefully mostly human!?) and thereby free up time spent by excessively burdened volunteer reviewers.

It would be great if NIH would convene a stakeholder meeting to discuss how a pilot Hub might look?

For more detailed expression of some of these arguments, please see past editorials at these PMIDs: 18368097, 30776062, 36113101. Also attached.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: Thank you to NIH for understanding the inherently fluid nature of spending research budgets. At the VA, we have great difficulty spending VA research dollars due to impossibly dumb HR practices, as you are surely aware. VA researchers have to use separate nonprofit Foundations to more easily spend non-VA dollars. Most of us would rather go for non-VA than VA grants.

If only the NIH could somehow take over control of VA ORD administrative logistics (not its mission or scope). And if only there were a way to bring some sense into VA research HR and onboarding…I’d be happy to provide some absurd examples, but surely you’re already aware of many. The rumors are true.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/jamaneurology_harel_2019.pdf

Description: JAMA editorial

VIEWPOINT

Untethering Idea Generation From Protocol Execution in Neurorehabilitation Research

Noam Yehezkel Harel, MD, PhD
Icahn School of Medicine at Mount Sinai, New York, New York; and James J. Peters Veterans Affairs Medical Center, New York, New York.

Proficient conduct of human research is difficult. The difficulty is magnified in neurorehabilitation research; function is difficult to quantify, large numbers of participants are required to distinguish statistically significant effects of an intervention from rates of natural recovery, nonpharmacological interventions garner less industry funding and are harder to apply in a standardized fashion, and so on.[1,2] Overcoming these challenges requires resources and trial experience that are simply not available to most would-be clinician-investigators who have limited protected research time.

As a result, when an investigator has a novel idea but lacks the time or resources to experimentally pursue it, what happens? Often, nothing. To preserve the opportunity for academic credit, the idea is privately hoarded for a future implementation that may never occur. When an investigator with suboptimal experience does get the opportunity to test a novel idea, what happens? Usually, an unreliable pilot study that muddies the field. This writer can attest to the latter scenario through personal experience. Neurorehabilitation studies often suffer from serious flaws in study design, conduction, analysis, and/or reporting.[2,3] It is no surprise that rehabilitation manuscripts have significantly lower odds of acceptance in *Stroke*.[4]

Unexplored ideas and inexpert investigation contribute to a bottleneck in effective clinical translation and a crisis of irreproducible results in neurorehabilitation and beyond. Commendable initiatives like the National Institutes of Health Network for Excellence in Neuroscience Clinical Trials (NeuroNEXT) and StrokeNet are not enough to surmount the immense barriers facing most clinical investigators. To reap more health benefits from basic biomedical advances, clinical research needs radical reengineering.

Ironically, to facilitate greater idea propagation and participation in clinical research, fewer investigators should conduct their own studies. Each investigator should not need to initiate, own, operate, and maintain all aspects of a project to earn full academic credit. Societal resources would be more efficiently used if academic investigators—and their institutions—specialized into complementary roles. Untether idea generation from protocol execution.

Rather than shelve untested ideas or inadequately perform early-stage clinical neurorehabilitation studies, idea generators in a synergistic system would collaborate remotely with experimentalists at high-volume protocol execution hubs. Governments, foundations, and industries will be motivated to generously fund protocol execution hubs, as further outlined below. In this fashion, the best ideas would be tested by the most experienced practitioners at properly equipped and staffed facilities. Investigators at idea-generating and protocol execution hubs would share academic credit. Their institutions would share potential royalties. This would increasingly democratize the scientific enterprise among academics in developing and wealthy countries alike.

Of course, not all biomedical research can or should be dichotomized into idea generation and protocol execution at separate locations. Clinical neurorehabilitation research, with its difficulty in standardizing procedures, would be an obvious place to start. Rather than focusing only on randomized clinical trials, neurorehabilitation protocol execution hubs would be essential for performing reliable early-phase studies—points at which funders make critical decisions on investing further resources.[3] Conducting studies exclusively through regional hubs would increase savings through economies of scale; improve study quality, objectivity, and reliability; and provide multifaceted opportunities for mentoring in protocol execution methodology—goals shared by all stakeholders.

Consolidating protocol execution at high-throughput centers would also reduce competitive recruiting among disparate teams for studies with overlapping inclusion criteria. Instead, potential participants could enlist with protocol execution centers and undergo thorough baseline characterization. They would then be offered participation in studies that best match their inclusion criteria. With suitable washout periods, participants could choose to participate in more than 1 study at the same center. Execution teams would thereby collect multidimensional data and metadata on each participant. Housing a wider selection of studies under one or a few roofs would thus improve experimental reliability and participant equity.

Patients would benefit the most from synergistic specialization in clinical research. A larger crop from which to efficiently harvest the best ideas for testing at the most experienced clinical research centers would make medical progress faster, more reliable, and more impactful. This approach could enhance progress in all medical fields, not just neurorehabilitation.

Talk of reorganizing institutional resources could raise concerns of socialist idealism. Idealist, yes; socialist, no. Rather than imposing a plan from the top down, multiple interesting market-based scenarios could transpire in a synergistically specialized ecosystem. How and by whom would proposals be vetted? How and by whom would idea-generating scholars and their institutions be compensated? Would Uber-like intermediaries arise to match proposals with execution teams? Would career prospects improve for legions of discouraged postdocs and junior faculty? In one envisioned model, investigators would deposit proposals into a publicly accessible centralized repository[5,6]—perhaps bioRxiv could ex-

Corresponding Author: Noam Yehezkel Harel, MD, PhD, James J. Peters Veterans Affairs Medical Center, 130 W Kingsbridge Rd, Ste 7A-13G, New York, NY 10468 (noam.harel@mssm.edu).

© 2019 American Medical Association. All rights reserved.

Downloaded from jamanetwork.com by Mount Sinai School of Medicine user on 02/19/2019

pand its current scope of archiving completed study preprints. Trial hubs would bid to execute the best proposals. Academic and financial bonuses would be given for meritorious ideas and reproducible study results. Stable contracts underwriting personnel at idea-generating and protocol execution hubs would partially replace cyclical grant writing. This would reduce pressure on investigators to aggrandize or even distort results of premature pilot studies. Market-based interplay between academia, governments, and industries at multiple nodes in this environment could stimulate innovative approaches to technology transfer, a goal of governments and industries around the world.

Talk of unleashing market forces on academia may raise concerns of capitalist idealism and libertarian dystopia. Might disputes arise when similar research designs are enacted in parallel by competing experimental hubs? Yes. But many such disputes already rage today. With broader exchange of ideas independent of execution, disputes would occur in the context of a more transparent system, with financial and academic incentives that better align with the societal imperative of reproducibility.[7]

Inevitably, many academic institutions will resist abandoning aspirations to maintain a full in-house clinical trial apparatus. Most of the academic biomedical community continues to function analogously to brick-and-mortar stores, which fell behind Amazon by stubbornly trying to produce, warehouse, and sell products all by themselves. Forward-thinking retailers prosper by specializing in what they do best without investing massive effort on infrastructure activities they do less well. Likewise, a network of clinical research hubs would provide the optimal platform for the best products (ie, ideas) to be sold (ie, tested).

Gradually, the advantages of institutional specialization and the disadvantages of clinging to an older research model will become more evident. Nonhub institutions would expend fewer resources on administrative overhead and capital. Funders will need to help these institutions revise spending priorities and soften the blow of lost indirect costs. Funds reinvested in teaching, idea development, and mentoring would be far more beneficial than funds wasted on duplicative research administrative duties. This would further incentivize funders to financially support all academic institutions, not just those that become experimental hubs.

Disruption almost never occurs suddenly. Many key aspects of the new research ecosystem have been advancing for years—for example, greater emphasis on research transparency, data reproducibility, and credit attribution for team science.[7] Forward-thinking funders need to play a financially reassuring, decisive role in pushing many institutions to take the next leap of ceding clinical research efforts.

As with so many debates, everyone knows the system for translating biomedical research into clinical advances needs to improve, and everyone has different opinions on how to do so. In neurorehabilitation and beyond, untethering idea generation from protocol execution could allow investigators, institutions, and funders to find common ground for patients' benefit.

ARTICLE INFORMATION

Conflict of Interest Disclosures: None reported.

Published Online: February 18, 2019. doi:10.1001/jamaneurol.2018.4918

Disclaimer: The views expressed in this article are those of the author and do not necessarily reflect the position or policy of the US Department of Veterans Affairs or the US government.

Additional Contributions: I thank Stuart C. Sealfon, MD (Icahn School of Medicine at Mount Sinai, New York, New York), and John P. A. Ioannidis, MD, DSc (Stanford School of Medicine, Stanford, California), for constructive suggestions on drafts of this Viewpoint. Neither was compensated for their contributions.

REFERENCES

1. Fawcett JW, Curt A, Steeves JD, et al. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: spontaneous recovery after spinal cord injury and statistical power needed for therapeutic clinical trials. *Spinal Cord*. 2007;45(3):190-205. doi:10.1038/sj.sc.3102007

2. Kwakkel G, Lannin NA, Borschmann K, et al. Standardized measurement of sensorimotor recovery in stroke trials: consensus-based core recommendations from the Stroke Recovery and Rehabilitation Roundtable. *Int J Stroke*. 2017;12(5):451-461. doi:10.1177/1747493017711813

3. Dobkin BH. Progressive staging of pilot studies to improve phase III trials for motor interventions. *Neurorehabil Neural Repair*. 2009;23(3):197-206. doi:10.1177/1545968309331863

4. Sposato LA, Ovbiagele B, Johnston SC, Fisher M, Saposnik G; Stroke Outcome Research Working Group (www.sorcan.ca). A peek behind the curtain: peer review and editorial decision making at *Stroke*. *Ann Neurol*. 2014;76(2):151-158. doi:10.1002/ana.24218

5. Harel NY. Call for a centralized grant proposal repository. *Nature*. 2008;452(7186):409. doi:10.1038/452409b

6. Harel NY. Funding: some sense in overlapping grants. *Nature*. 2013;495(7440):174. doi:10.1038/495174b

7. Nosek BA, Alter G, Banks GC, et al. Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science*. 2015;348(6242):1422-1425. doi:10.1126/science.aab2374

© 2019 American Medical Association. All rights reserved.

Downloaded from jamanetwork.com by Mount Sinai School of Medicine user on 02/19/2019

257. American Educational Research Association

I am responding: On behalf of an organization

Name: Tabbye M. Chavous

Name of Organization: American Educational Research Association

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: AERA supports the ongoing efforts that NIH has taken to promote the importance of reproducibility and replication research as part of NIH's research portfolio. Some additional actions that NIH can take to build incentives and provide infrastructure for reproducibility research include:

• Provide funding support for research activities that explicitly support replication and reproducibility studies, as well as secondary data analyses, of prior NIH-funded research. Such activities also could include using technological innovations to develop new methodological approaches for conducting reproducibility and replication studies. These activities would also align with recent priorities from the White House Office of Science and Technology (OSTP) to invest in replication work. To achieve such a rigorous replication research ecosystem, it is also critical to support the foundational research necessary to establish baselines for replication studies.

• Through its positionality, collaborate with scientific societies and academic publishers to pursue strategies to promote publication of research that note null or negative effects for different education and health interventions, including through replication and reproducibility studies. Such studies would be able to describe barriers to implementation among other contexts that may affect replicable results.

• Invest in training and professional development opportunities for graduate students and early career researchers that would promote and develop skills for innovative methods to facilitate replication and reproducibility studies that build upon prior NIH-funded research.

2. Data, Software, and Model Sharing: We appreciate the ongoing work that NIH has undertaken to promote the characteristics of repositories for data sharing that reflect the Findable, Accessibility, Interoperability, and Reuse (FAIR) principles as well as efforts across federal scientific agencies – including NIH – to enable public access to data from federally funded research publications for future use. At the same time, we recognize the challenges and barriers that exist in the broader scientific enterprise, including in education research, to fully facilitate data sharing practices.

Recent research related to the sharing of data in education research has indicated several barriers that include the lack of shared data management frameworks in addition to lack of training on data sharing and concerns on participant identification and being scooped, among other concerns. At the same time,

researchers have noted the importance of developing and maintaining shareable datasets in education research to maximize the value of longitudinal data in important topics such as examining factors that lead to understanding teacher effectiveness.

Even with NIH's efforts to promote and educate the importance of data sharing with NIH-funded researchers, more can be done to recognize the value of how shared data and other products are integrated into future basic and applied research activities:

• Promote the NIH data management and sharing policy and implementation of the FAIR principles, alongside training efforts for peer reviewers, grant applicants, and grant recipients on how NIH's data management and sharing policies should be considered during the review of grant proposals and as a result of funded research.

• Invest in infrastructure to enable the storage and sharing of data from NIH-funded research with considerations of privacy and confidentiality for research involving human subjects. The OSTP Desirable Characteristics of Data Repositories for Federally Funded Research serves as one source of core guidance to develop secure and accessible data repositories.

• Invest in predoctoral and postdoctoral professional development and training to support data sharing – including developing codebooks, incorporating metadata, and ensuring privacy and confidentiality of human subjects.

• Develop procedures that would enable the use of clinical trial data from NIH-funded research under a restricted use basis. Frequently used and cited datasets could be made available for further research use through the National Secure Data Service. Practices undertaken by the National Center for Education Statistics to provide restricted-use access to evaluation data from evaluations conducted by the Institute of Education Sciences could be a model for NIH procedures.

• Recognize the curation and accessibility of NIH-funded datasets through citations of direct object identifiers submitted into data repositories and used in future research.

3. Training and Mentorship: In addition to the training and mentorship opportunities noted previously for rigor and reproducibility and for data, software, and model sharing, NIH is encouraged to undertake activities to examine impact of its own funded training and mentorship opportunities:

• Collaborate with NSF on work Congress authorized under the CHIPS and Science Act to conduct a study examining the impact of different funding mechanisms on graduate student experiences and outcomes, including whether such mechanisms have differential impacts on sub-sets of the student population.

• Develop measures that describe the impact of NIH-funded training and mentorship activities for participants in assessing longitudinal success that include:

o Number of training and mentorship participants.
o Future training and mentorship activities developed or provided by prior participants in NIH-funded training and mentorship activities.

o Linkages between NIH-funded training and mentorship programs and citations and success in receiving future federal research grants
o Recognition for engagement with practitioners, policymakers, and other public audiences in advancing NIH-funded research.

4. Collaboration: We recognize the tradition – even within AERA's own activities – of honoring individual accomplishments and contributions to advancing scientific knowledge. At the same time, researchers increasingly must engage in interdisciplinary partnerships to address complex problems that require different expertise (e.g., disease treatment that requires a deep understanding of both biological/physiological mechanisms and behavioral/social science mechanisms that determine health outcomes). Further, research funding entities may be requiring interdisciplinary teams in ways that do not align with institutional rewards and advancement for "independent" scientific contributions. NIH could serve as a leader in collaboration with the broader scientific community and practice/policy-oriented partners to develop recognitions of high quality, high impact team-based science in work developed through NIH funding.

Such an effort would require consideration of how to recognize both the contributions that individual team members make to knowledge and practice and the synergistic impacts that would not be possible without the team members' collaboration. An additional consideration is how to best assess and honor high-quality team coordination leadership and the infrastructure needed to foster effective, impactful collaboration among teams. This could be an opportunity for NIH to partner with other federal agencies to engage with scholars who are the forefront of interdisciplinary "team science" to understand its value and the different factors that lead to or challenge successful research and its application for societal impacts.

5. Entrepreneurship and Translation: The National Center for Advancing Translational Sciences (NCATS) has as its mission "to transform the translational process so that new treatments and cures for disease could be delivered to patients faster." NCATS funds translational science but could also play a significant role in synthesizing evidence on effective health and education practices that can be produced in accessible formats for practitioners.

As one example, NCATS would be an appropriate partner with NICHD to both support NIH-funded syntheses and to produce evidence-based recommendations in plain language from NICHD-funded research that would support the cognitive and behavioral development of students with intellectual and developmental disabilities. NCATS would have the capability to monitor and track use and adoption of recommendations to assess broader impact of NIH-funded research.

6. Foundational Scientific Exploration: AERA has had a longstanding appreciation for the role that basic and fundamental research has in building the evidence base for its application in evidence-based policy and practice as well as the role that basic research has in generating cumulative knowledge. We encourage NIH to ensure continued investment in basic and fundamental research as it is essential to generating tools, interventions, and strategies that improve health outcomes, including those at the intersection of education.

We recommend that NIH build upon prior research and evidence-based strategies that have shown evidence of effectiveness in incorporating high-risk, high-reward activities within the NIH portfolio. In

addition to ongoing work within the Advanced Research Projects Agency-Health, other federal research agencies are taking steps to implement high-risk, high-reward activities. The Institute of Education Sciences (IES) has funded the first round of Seedlings to Scale, which promotes the development of quick-turnaround high-reward, scalable solutions intended to improve education outcomes for all learners. Seedlings to Scale projects are focused on an identified need with funding conducted in three phases. This is an example of one model that NIH could pursue.

7. Public Impact: NIH could model actions that institutions, including scientific societies, are taking to recognize high-quality translation, synthesis, knowledge mobilization, and other public engagement activities that are results of NIH-funded research:

• The NIH Office of Behavioral and Social Science Research honors a distinguished researcher who has made transformative impact on NIH research as a Matilda White Riley Honoree. A criterion for this honor is conducting "behavioral and social sciences research that improvs the lives of individuals and society." This award is an example that could be a model for similar awards and honors for NIH-funded researchers – either within specific institutes or NIH-wide – that would recognize public impact of NIH-funded research to improving overall societal outcomes.

• As an example from a scientific society, AERA established the Exemplary Contributions to Practice-Engaged Research Award in 2016, which recognizes an education research scholar or scholars in recognition of collaborative project(s) between researchers and practitioners that have had sustained and observable effects on contexts of practice. Recipients of this award have demonstrated impact of their work in connection to the communities with which they conduct research and provide guidance for developing and implementing evidence-based practice. NIH could honor similar connections of researchers with patients, health practitioners, and other communities that use NIH-funded research in their work.

NIH could also incorporate non-publication impact, dissemination, and knowledge mobilization as part of consideration of grants in the peer review process and for highlighting long-term grant success. One resource for assessing public impact that could be a model for NIH from education research is a recent report from IES describing the development and validation of measures for research evidence use, research capacity, and partner experiences with the Regional Educational Laboratories. Constructs that could be applied to NIH-supported research-practice partnerships include usefulness of research evidence, relevance of evidence and services, and accessibility of evidence and services. Additional metrics that NIH could use to measure public impact include:

• The number of people who are served from the interventions, diagnostics, and treatments (among other tools) funded by NIH.

• The number of practitioner-oriented resources (e.g., curricular materials for medical schools, professional development for health practitioners, assessments, evidence-based health and/or education practices) that are developed as a result of NIH investment, along with the number of people using those resources.

• The number of research-practice partnerships that enable intentional connections between basic

health and behavioral research with intermediaries that translate and use research findings in practice. These types of partnerships would help promote continuous improvement among communities of practice and inform practice-informed research questions within basic biomedical and behavioral research.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/AERA-Comments-on-NIH-RFI-Measuring-and-Rewarding-Impact-FINAL.pdf

Description: Please see attached PDF with fuller response with footnotes to the RFI.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

August 19, 2026

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive #750
Bethesda, MD 20817
Comments submitted through NIH website

Re: NOT-OD-26-087, Request for Information on Measuring and Rewarding Scientific Impact

On behalf of the American Educational Research Association (AERA), thank you for the opportunity to provide comments on the Request for Information (RFI) on Measuring and Rewarding Scientific Impact. AERA is the major national scientific association of 22,000 faculty, researchers, graduate students, and other distinguished professionals dedicated to advancing knowledge about education, encouraging scholarly inquiry related to education, and promoting the use of research to improve education and serve the public good. AERA members include NIH principal investigators who receive funding for research on human development and cognition, including from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), the National Institute of Mental Health (NIMH), and the National Institute on Aging (NIA).

We appreciate NIH's consideration of innovative ways to measure and reward the scientific impact of NIH-funded research. Overall, we encourage NIH to consider investment in professional development activities and infrastructure to educate and promote best practices that would enable assessment and measurement of the impact of NIH-funded research beyond publication citations and patents.

NIH has been a leading federal scientific agency in advancing rigorous research and practices for facilitating public access to NIH-funded research and datasets that are the basis behind research publications. As NIH considers new measures and rewards for scientific impact, AERA also encourages NIH to adopt practices that highlight the public impact of the agency's funded research and to reward public engagement of its work with communities and practitioners that benefit from the application of NIH-funded research. This response reflects the categories included in the online form for comment.

• ***Rigor and Reproducibility***. *Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings.*

AERA supports the ongoing efforts that NIH has taken to promote the importance of reproducibility and replication research as part of NIH's research portfolio. Some additional actions that NIH can take to build incentives and provide infrastructure for reproducibility research include:

- Provide funding support for research activities that explicitly support replication and reproducibility studies, as well as secondary data analyses, of prior NIH-funded research. Such activities also could include using technological innovations to develop new methodological approaches for conducting reproducibility and replication studies. These activities would also align with recent priorities from the White House Office of Science and Technology (OSTP) to invest in replication work. To achieve such a rigorous replication research ecosystem, it is also critical to support the foundational research necessary to establish baselines for replication studies.
- Through its positionality, collaborate with scientific societies and academic publishers to pursue strategies to promote publication of research that note null or negative effects for different education and health interventions, including through replication and reproducibility studies. Such studies would be able to describe barriers to implementation among other contexts that may affect replicable results.[1]
- Invest in training and professional development opportunities for graduate students and early career researchers that would promote and develop skills for innovative methods to facilitate replication and reproducibility studies that build upon prior NIH-funded research.

• ***Data, Software, and Model Sharing****. Sharing scientific data and tools accelerates biomedical research discovery, in part, by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies. NIH seeks to incentivize and reward these activities to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.*

We appreciate the ongoing work that NIH has undertaken to promote the characteristics of repositories for data sharing that reflect the Findable, Accessibility, Interoperability, and Reuse (FAIR) principles as well as efforts across federal scientific agencies – including NIH – to enable public access to data from federally funded research publications for future use. At the same time, we recognize the challenges and barriers that exist in the broader scientific enterprise, including in education research, to fully facilitate data sharing practices.

[1] Kim, J. S. (2019). Making Every Study Count: Learning From Replication Failure to Improve Intervention Research. *Educational Researcher*, *48*(9), 599-607.

Recent research related to the sharing of data in education research has indicated several barriers that include the lack of shared data management frameworks[2] in addition to lack of training on data sharing and concerns on participant identification and being scooped, among other concerns.[3] At the same time, researchers have noted the importance of developing and maintaining shareable datasets in education research to maximize the value of longitudinal data in important topics such as examining factors that lead to understanding teacher effectiveness.[4]

Even with NIH's efforts to promote and educate the importance of data sharing with NIH-funded researchers, more can be done to recognize the value of how shared data and other products are integrated into future basic and applied research activities:

- Promote the NIH data management and sharing policy and implementation of the FAIR principles,[5] alongside training efforts for peer reviewers, grant applicants, and grant recipients on how NIH's data management and sharing policies should be considered during the review of grant proposals and as a result of funded research.
- Invest in infrastructure to enable the storage and sharing of data from NIH-funded research with considerations of privacy and confidentiality for research involving human subjects. The OSTP *Desirable Characteristics of Data Repositories for Federally Funded Research*[6] serves as one source of core guidance to develop secure and accessible data repositories.
- Invest in predoctoral and postdoctoral professional development and training to support data sharing – including developing codebooks, incorporating metadata, and ensuring privacy and confidentiality of human subjects.
- Develop procedures that would enable the use of clinical trial data from NIH-funded research under a restricted use basis. Frequently used and cited datasets could be made available for further research use through the National Secure Data Service. Practices undertaken by the National Center for Education Statistics to provide restricted-use access to evaluation data from evaluations conducted by the Institute of Education Sciences could be a model for NIH procedures.
- Recognize the curation and accessibility of NIH-funded datasets through citations of direct object identifiers submitted into data repositories and used in future research.

---

[2] Bowers, A. J., & Choi, Y. (2023). Building School Data Equity, Infrastructure, and Capacity Through FAIR Data Standards: Findable, Accessible, Interoperable, and Reusable. *Educational Researcher*, *52*(7), 450-458.

[3] Logan, J. A. R., Hart, S. A., & Schatschneider, C. (2021). Data Sharing in Education Science. *AERA Open*, *7.* (note that this was supported with funding from NICHD)

[4] Sajjad, T., Hartig, J., Lösch, T., & Köhler, C. (2026). Investigating Data Sharing Practices in Teaching Effectiveness Research. *AERA Open*, *12*.

[5] https://www.go-fair.org/fair-principles/

[6] The National Science and Technology Council, *Desirable Characteristics of Data Repositories for Federally Funded Research*, 2022, DOI: https://doi.org/10.5479/10088/113528

• ***Training and Mentorship****. NIH recognizes that advancing its mission requires promoting the growth and stability of the biomedical research workforce. While NIH supports training and protects time through existing policies and programs, additional indicators are needed to assess and quantify the impact of these efforts.*

In addition to the training and mentorship opportunities noted previously for rigor and reproducibility and for data, software, and model sharing, NIH is encouraged to undertake activities to examine impact of its own funded training and mentorship opportunities:

- Collaborate with NSF on work Congress authorized under the CHIPS and Science Act to conduct a study examining the impact of different funding mechanisms on graduate student experiences and outcomes, including whether such mechanisms have differential impacts on sub-sets of the student population.[7]
- Develop measures that describe the impact of NIH-funded training and mentorship activities for participants in assessing longitudinal success that include:
    - Number of training and mentorship participants.
    - Future training and mentorship activities developed or provided by prior participants in NIH-funded training and mentorship activities.
    - Linkages between NIH-funded training and mentorship programs and citations and success in receiving future federal research grants
    - Recognition for engagement with practitioners, policymakers, and other public audiences in advancing NIH-funded research.

• ***Collaboration****. As biomedical research becomes increasingly interdisciplinary and convergent, collaboration is growing in both scale and complexity. NIH seeks new approaches for assessing and recognizing contributions to team-based science.*

We recognize the tradition – even within AERA's own activities – of honoring individual accomplishments and contributions to advancing scientific knowledge. At the same time, researchers increasingly must engage in interdisciplinary partnerships to address complex problems that require different expertise (e.g., disease treatment that requires a deep understanding of both biological/physiological mechanisms and behavioral/social science mechanisms that determine health outcomes). Further, research funding entities may be requiring interdisciplinary teams in ways that do not align with institutional rewards and advancement for "independent" scientific contributions. NIH could serve as a leader in collaboration with the broader scientific community and practice/policy-oriented partners to develop recognitions of high quality, high impact team-based science in work developed through NIH funding.

Such an effort would require consideration of how to recognize both the contributions that individual team members make to knowledge and practice and the synergistic

---

[7] 42 USC 18993(c)

impacts that would not be possible without the team members' collaboration. An additional consideration is how to best assess and honor high-quality team coordination leadership and the infrastructure needed to foster effective, impactful collaboration among teams. This could be an opportunity for NIH to partner with other federal agencies to engage with scholars who are the forefront of interdisciplinary "team science" to understand its value and the different factors that lead to or challenge successful research and its application for societal impacts.

• ***Entrepreneurship and Translation****. To advance the translation of discoveries into products for people, NIH fosters the maturation of discoveries from bench to bedside and promotes innovation across sectors. More can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission.*

The National Center for Advancing Translational Sciences (NCATS) has as its mission "to transform the translational process so that new treatments and cures for disease could be delivered to patients faster." NCATS funds translational science but could also play a significant role in synthesizing evidence on effective health and education practices that can be produced in accessible formats for practitioners.

As one example, NCATS would be an appropriate partner with NICHD to both support NIH-funded syntheses and to produce evidence-based recommendations in plain language from NICHD-funded research that would support the cognitive and behavioral development of students with intellectual and developmental disabilities. NCATS would have the capability to monitor and track use and adoption of recommendations to assess broader impact of NIH-funded research.

• ***Foundational Scientific Exploration****. Basic research and high-risk, high-reward initiatives represent foundational scientific exploration with an inherent risk of failure but the potential for paradigm-shifting breakthroughs. Further embedding these approaches within the academic enterprise is essential to ensuring continued support for bold and innovative research directions.*

AERA has had a longstanding appreciation for the role that basic and fundamental research has in building the evidence base for its application in evidence-based policy and practice as well as the role that basic research has in generating cumulative knowledge. We encourage NIH to ensure continued investment in basic and fundamental research as it is essential to generating tools, interventions, and strategies that improve health outcomes, including those at the intersection of education.

We recommend that NIH build upon prior research and evidence-based strategies that have shown evidence of effectiveness in incorporating high-risk, high-reward activities within the NIH portfolio. In addition to ongoing work within the Advanced Research Projects Agency-Health, other federal research agencies are taking steps to implement high-risk, high-reward activities. The Institute of Education Sciences (IES) has funded the first round of Seedlings to Scale, which promotes the development of quick-

turnaround high-reward, scalable solutions intended to improve education outcomes for all learners.[8] Seedlings to Scale projects are focused on an identified need with funding conducted in three phases. This is an example of one model that NIH could pursue.

• ***Public Impact****. NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.*

NIH could model actions that institutions, including scientific societies, are taking to recognize high-quality translation, synthesis, knowledge mobilization, and other public engagement activities that are results of NIH-funded research:

- The NIH Office of Behavioral and Social Science Research honors a distinguished researcher who has made transformative impact on NIH research as a Matilda White Riley Honoree. A criterion for this honor is conducting "behavioral and social sciences research that improvs the lives of individuals and society." This award is an example that could be a model for similar awards and honors for NIH-funded researchers – either within specific institutes or NIH-wide – that would recognize public impact of NIH-funded research to improving overall societal outcomes.
- As an example from a scientific society, AERA established the Exemplary Contributions to Practice-Engaged Research Award in 2016, which recognizes an education research scholar or scholars in recognition of collaborative project(s) between researchers and practitioners that have had sustained and observable effects on contexts of practice.[9] Recipients of this award have demonstrated impact of their work in connection to the communities with which they conduct research and provide guidance for developing and implementing evidence-based practice. NIH could honor similar connections of researchers with patients, health practitioners, and other communities that use NIH-funded research in their work.

NIH could also incorporate non-publication impact, dissemination, and knowledge mobilization as part of consideration of grants in the peer review process and for highlighting long-term grant success. One resource for assessing public impact that could be a model for NIH from education research is a recent report from IES describing the development and validation of measures for research evidence use, research capacity, and partner experiences with the Regional Educational Laboratories.[10]

---

[8] https://ies.ed.gov/funding/grants/2025/seedlings-scale-84-305j

[9] https://www.aera.net/About-AERA/AERA-Awards/Exemplary-Contributions-to-Practice-Engaged-Research-Award

[10] Bleeker, M., Crigler, F., Hurwitz, F., Carmichael, K., Poznyak, D., Neild, R. C. (2026). *Developing and validating measures of research evidence use, research capacity, and partner experiences with the Regional Educational Laboratories* (REL 2026–016). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/sites/default/files/rel-central/document/2026/08/2026016r.pdf

Constructs that could be applied to NIH-supported research-practice partnerships include usefulness of research evidence, relevance of evidence and services, and accessibility of evidence and services. Additional metrics that NIH could use to measure public impact include:

- The number of people who are served from the interventions, diagnostics, and treatments (among other tools) funded by NIH.
- The number of practitioner-oriented resources (e.g., curricular materials for medical schools, professional development for health practitioners, assessments, evidence-based health and/or education practices) that are developed as a result of NIH investment, along with the number of people using those resources.
- The number of research-practice partnerships that enable intentional connections between basic health and behavioral research with intermediaries that translate and use research findings in practice. These types of partnerships would help promote continuous improvement among communities of practice and inform practice-informed research questions within basic biomedical and behavioral research.

Thank you again for the opportunity to provide comments. Please do not hesitate to contact us with any questions or need additional information as NIH takes up this important opportunity on behalf of the broader federal research enterprise.

Sincerely,



Tabbye M. Chavous, PhD
Executive Director
tchavous@aera.net

258. Alyson Eggleston

I am responding: On behalf of myself

Name: Alyson Eggleston

Name of Organization: Penn State University, CTSI

Type of Organization: Academic Institution

Role: Other

Role - Other: Faculty Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: NIH has already advanced beyond raw publication and citation counts through tools such as iCite and the Relative Citation Ratio (RCR), which provide useful field-normalized indicators of scientific influence. The next opportunity is not to replace these measures with another universal score, but to consider portfolios of contribution that capture how research, resources, people, and collaborations generate downstream scientific and public value.
Rigor, reproducibility, and reuse
NIH should distinguish compliance with reproducibility requirements from evidence that research products are actually reusable or independently replicated. Particularly useful indicators include:
• availability of protocols, analytic documentation, code, metadata, provenance, and versioning; and
• successful reuse, adaptation, or independent replication by researchers outside the originating team.
Independent replication is especially important but often outside the control of the original investigator. NIH could strengthen incentives by supporting dedicated opportunities to replicate influential findings and protocols and by recognizing successful replication as a scientific contribution.

2. Data, Software, and Model Sharing: For datasets, software, models, instruments, and other shared resources, deposit alone should not constitute impact. Greater weight should be given to whether resources are discoverable, documented, maintained, and reused. NIH could also promote minimum metadata and implementation standards that make shared research products easier to compare and adopt.

3. Training and Mentorship: Training outcomes should be assessed longitudinally and should not be reduced to publications or attainment of an independent R01. Relevant indicators could include:
• progression from career-development or pilot support to subsequent research, implementation, collaboration, or funding.
This approach would recognize scientific contribution without assuming that becoming an independent principal investigator is the only meaningful outcome of research training.

4. Collaboration: Coauthorship counts alone are insufficient measures of collaboration. NIH should consider both the structure and consequences of collaborative work. Useful evidence could include:
• formation and persistence of cross-disciplinary or cross-organizational collaborations; and
• documented contributor roles and evidence that methods, data, expertise, or perspectives were

meaningfully integrated.
Network analysis can help identify collaboration patterns, but network breadth should not automatically be interpreted as quality. NIH should also recognize the contributions of connectors, translators, infrastructure builders, and boundary-spanners who enable team science.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Translation should be evaluated as progression through multiple possible pathways rather than solely commercialization or immediate population-health change. Frameworks such as the Translational Science Benefits Model illustrate how research may generate clinical, community, economic, and policy impacts.
Examples include:
• adoption of research findings, methods, tools, or workflows in clinical, community, organizational, or policy settings; and
• patents, products, partnerships, follow-on funding, cost savings, increased efficiency, or sustained use of research-derived resources.
Follow-on funding and return on research investment can demonstrate leverage and stewardship, but they should not become stand-alone proxies for scientific value. Foundational, qualitative, prevention-oriented, and community-engaged research may produce substantial benefit without generating large near-term financial returns.

8. Other Areas: N/A

259. Peace-Work (peace-work.org)

I am responding: On behalf of an organization

Name: David J Corliss

Name of Organization: Peace-Work (peace-work.org)

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: (NB: these comments refer to only to the section on Public Impact. Please see tayh section for my comments. Thank you. DJC)

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: It is proposed that a metric be implemented in the review process to assess the beneficial impact both to individuals and society as a whole. This proposed metric follows the methodology of the United Nations' Human Development Index (HDI). For further information on that index, please see https://hdr.undp.org/data-center/human-development-index#/indicies/HDI.

The metric I am proposing in these comments utilizes five factors, beginning with median life expectancy and years of education which are included in the HDI. The poverty rate replaces HDI factor for Income. To these factors, the maternal mortality rate and incarceration rate are added.

The rubric for the proposed metric begins with the percent change expected in the base numbers that could result from the research. These numbers are then normalized to 100% and then combined using the geometric mean, in the same manner as the HDI. Most factors will of course be 1 for most most research. Use of this metric provides a data-driven assessment of human and societal benefit from scientific research.

8. Other Areas: N/A

260. American Association of Colleges of Pharmacy

I am responding: On behalf of an organization

Name: Dorothy Farrell

Name of Organization: American Association of Colleges of Pharmacy

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Rigor, reproducibility, replicability and transparency are baseline expectations of all research and scholarship. However, standards and practices that contribute to these essential characteristics, including reporting guidelines, are not always available or in wide use, especially for emerging technologies and disciplines. NIH should support dedicated research studies on the development of reproducible, replicable and transparent approaches in the use of new technology in biomedical and behavioral research, e.g., AI-guided drug discovery, new approach methodologies (NAMs) for drug testing, LLM-guided analysis of electronic health records. Such studies would elucidate how research design and disclosure practices affect the rigor, reproducibility, replication and robustness of research results, improve methodology and reporting, and enable development of rubrics and evidence-based practices for use of new technologies. Supplements to these or other NIH awards could support testing and validation of the practices in different research studies. NIH funding and supplements for these studies would reward the efforts of those who devote energies to developing rigorous research approaches and provide incentives for other researchers to utilize those approaches.

Development of evidence-based practices should be accompanied by support for dissemination and implementation studies, to ensure these practices are translated into active research settings, including academic, industry and clinical sites. Studies that can reveal the barriers and facilitators to rigorous and reproducible research approaches will provide further information on what actions NIH and institutions can take to incentivize these approaches.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: Recipients of NIH training and career development awards will pursue varied career paths, and the appropriate indicators of success and performance metrics for these individuals will vary depending on desired career path. Reflecting this, it has become standard practice to use Individual Development Plans (IDPs) to set personalized outcomes and indicators for graduate students, while training and career development grants require plans for didactic and experiential training to advance stated career objectives. Mentor history is evaluated during grant review, but difficulties in matching outcomes for prior trainees with their career objectives mean success is generally defined after the fact. That is, a former training who remains a research scientist is considered a success, even if they aren't content in their position, weren't well trained for it or if their career path was torturous. This suggests assessing workforce mentorship may require aggregate indicators, where trainee outcomes for programs are judged not only as individual placements but by the mix of career

outcomes and its relation to the stated goals of the program and judged needs of the research ecosystem.

For research training outside of dedicated training grants, inconsistent use of IDPs and adherence to their goals and misalignment of training goals and practices have been identified as persistent challenges. [Coop] NIH could and perhaps should collect outcomes for all trainees supported through its grants and assess PI mentorship readiness in evaluating proposals. However, these issues may reflect a mismatch between the model of a trainee supported by the focused research project of an academic advisor and the career development needs of trainees and workforce training needs of the biomedical research system. Considering alternative models to support research training may better serve NIH's mission of preparing the biomedical research workforce than adding or altering indicators and metrics of research program grants. Models could include institutional training grants offered in partnership with states, regional alliances or industry, such as NSF's pilot Industry-Integrated PhD Program, with indicators established in advance for each training model for both individual trainees, e.g., career placement satisfaction, and the program as a whole, e.g., workforce needs met, local economic stimulus. Any new model or program trial should include support for program evaluation and implementation studies, so that barriers, facilitators and outcomes are identified and tracked.

Even in traditional modes of research training, such as graduate student support through an R grant, NIH support for scholarship of teaching and learning studies applied to research training, scientific mentorship and graduate education would help provide an evidence base for the indicators and metrics that correlate with desired outcomes.

Coop A, Early J, et al. Supporting Research and Graduate Education in a Troubled Landscape: Report of the 2025-2026 AACP Research and Graduate Affairs Committee. Am J Pharm Educ. 2026 Jul 16;90(9):102056. doi: 10.1016/j.ajpe.2026.102056.

4. Collaboration: Recognizing the breadth of and the unique contributions made by different disciplines and research endeavors is critical to activate collaborations and spur innovation. To illuminate the relationships between types of research and enable identification of opportunities for collaboration, the 2024-2025 American Association of Colleges of Pharmacy (AACP) Research and Graduate Affairs Committee developed a unified definition of research and scholarship for pharmacy that captured different research pursuits, e.g., the scholarships of discovery, integration, application, and teaching, while supporting the purposes such a definition would serve, e.g., regulatory and ethical oversight, research impact and career development assessment. [Chahine] Relevant to this RFI, the Committee developed distinct definitions for research and scholarship to capture the broadest range of scholarly activities that should be recognized as advancing or contributing to the body of knowledge and improvement of health and wellbeing.

Both research and scholarship demand high standards of intellectual honesty, methodological rigor, ethical consideration, and transparency and should be subject to peer-review or recognition and publicly shared. Traditional research assessment focuses on basic discovery and the publications, citations and funding that accrue to it. The Committee recognized that scholarship, often centered on practice and including integrative, systemic or scoping reviews, commentaries/editorials, invited lectures/podium presentations, case reports/series, community engagement and real-world studies, is crucial to

translating discoveries to practice and ensuring successful implementation of evidence-based interventions. Scholarship will often represent the collaborative effort of many individuals and organizations spanning research, clinical practice and social engagement. Metrics to capture impact of scholarship can include changes in operating procedure in a clinic or health center due to case reports, numbers of community members who benefit from a public health campaign, or number of patients or cost savings due to a successful implementation study. These metrics are often collected at an institutional level, but individual effort can be assigned and credited. However, traditional research metrics may accord more impact to the outputs of basic or translational science researchers, e.g., publications, than to the work of community liaisons and clinical practitioners, e.g., clinical trial accrual or patient explainers. Both are crucial to project outcomes, though, so that impact may be measured at the project or team level and individual outputs mapped to project outcomes to assess individual contributions.

As research increasingly occurs in the context of collaborations, both near and distant, NIH should recognize the value of scholarship of integration and the key contributions made by research center and community managers, particularly with larger and multidisciplinary teams. NIH should support studies of successful interdisciplinary and interprofessional collaborations to build on the existing evidence base for effective practices in research collaboration management and community engagement. Studies of collaborations would also provide insights into metrics and indicators of successful integration that could be used to assess future research collaborations and provide insight into the conditions under which individual efforts can be recognized and accounted for in the context of collaborations and team science.

Chahine E, Cole J, Ogunsanya M et al. Advancing Research and Scholarship: Report of the 2024–2025 AACP Research and Graduate Affairs Committee. Am J Pharm Educ. 2025 Sep;89(9):101471. doi: 10.1016/j.ajpe.2025.101471. Epub 2025 Jul 20. PMID: 40692110.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Broader impact of biomedical and behavioral research is already being recognized in faculty evaluations at many institutions and by NIH, e.g., through inclusion in the “Contributions to Science” section of the biosketch. Examples include including regulatory approvals or commercialization of new therapies or technologies, business startups, creation of intervention protocols and their adoption in health care, educational or other settings, and creation of data resources or research tools for use by the research community. In many instances, though, these appropriately capture effort but not necessarily impact. Metrics that rely on counting, e.g., number of patients who are treated with a new product or protocol or number of researchers who access a resource, are valuable but may undervalue innovations in rare diseases or emerging research areas. NIH should support studies of research translation that explore the value and weighting of different indicators of success and impact and how contributions to successful translation can be identified and recognized.

The AACP Research and Graduate Committee has recognized that the education of the public, including policymakers, and clinicians on how to evaluate and use evidence is a core scientific contribution that is frequently underappreciated. [Coop] Although some metrics for evaluating efforts in communication, community engagement and related endeavors exist, e.g., altmetrics including media mentions, social

media footprint and references in policy documents, further development of metrics is needed. NIH should support studies of scientific communications, public health campaigns and community engagement to elucidate measures that go beyond counting contacts or impressions, as valuable as those may be, and indicate positive effects on health outcomes and societal wellbeing.

Coop A, Early J, et al. Supporting Research and Graduate Education in a Troubled Landscape: Report of the 2025-2026 AACP Research and Graduate Affairs Committee. Am J Pharm Educ. 2026 Jul 16;90(9):102056. doi: 10.1016/j.ajpe.2026.102056.

8. Other Areas: Given the considerable interest in research and researcher assessment and the current lack of consensus metrics and indicators beyond traditional bibliometrics and funding, NIH should share its findings through this RFI or follow-on efforts, to help the community identify and investigate those measures that may be pertinent to their own assessment needs.

261. N/A

I am responding: On behalf of an organization

Name:

Name of Organization:

Type of Organization: Non-Profit Research Organization

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/NOT-OD-26-087.pdf

We appreciate the opportunity to provide comments in response to NOT-OD-26-087, *Request for Information on Measuring and Rewarding Scientific Impact*.

As NIH develops such approaches, we encourage the agency to distinguish between established practices for conducting high-quality research which already exist and measures of scientific impact. Rigor, reproducibility, transparent reporting, appropriate experimental design, and data sharing are essential components of responsible biomedical research and are already reflected in NIH policies and guidance. Rather than creating additional layers of requirements to document these practices, NIH should focus on appropriately recognizing investigators and teams, and the organization's environment and capacity that make meaningful contributions to reproducibility, resource sharing, collaboration, training, translation, and public benefit.

Biomedical advances frequently result from decades of cumulative, collaborative work involving numerous investigators, institutions, funders, and sectors. Models that facilitate sharing of data, methods, resources, and expertise can amplify scientific impact and enable validation and extension of findings. NIH should therefore recognize collaborative infrastructure and shared resources while remaining cautious about assigning a specific economic or health outcome to an individual award or investigator.

NIH has established substantial expectations concerning rigor and reproducibility in grant applications, including consideration of the rigor of prior research (published and unpublished), rigorous experimental design, relevant biological variables, and authentication of key resources. Discipline-specific standards, such as the ARRIVE guidelines for animal research, similarly provide detailed expectations concerning experimental design, statistical analysis, minimization of bias, and transparent reporting. We therefore encourage NIH to build upon and harmonize existing frameworks rather than introduce duplicative requirements. Additional requirements should have a clearly demonstrated benefit and should avoid increasing administrative burden without a corresponding improvement in research quality.

Importantly, reproducibility should not be reduced to a binary indicator or a requirement that every study be independently replicated. NIH could instead recognize meaningful contributions to a reproducible research ecosystem, including:

- publication of sufficiently detailed methods and protocols;
- availability and documented reuse of datasets, software, models, and other research resources;
- development and validation of methods or resources used by other investigators;
- well-designed replication or validation studies, including studies that do not confirm an original finding; and
- adoption of and training in rigorous experimental-design and statistical practices appropriate to the field.

Because judgments about rigor, reproducibility, and scientific impact are inherently field- and context-dependent, NIH should also ensure that review panels include sufficient breadth of scientific, methodological, and statistical expertise to evaluate these dimensions appropriately.

NIH should minimize new reporting requirements. Wherever possible, indicators should be derived from information already available through NIH systems, publications, repositories, persistent identifiers, or other established sources rather than requiring investigators and institutions to create additional reports.

We encourage NIH to pilot proposed indicators before incorporating them into funding decisions and to evaluate those pilots for validity, administrative burden, unintended incentives, and effects across disciplines, career stages, and institution types.

262. Stanford University

I am responding: On behalf of an organization

Name: Stanford Program on Research Rigor and Reproducibility (SPORR), , composed of Steven N. Goodman, MD, MHS, PhD, Ruth O'Hara, John Borghi, PhD, and Dr. Mario Malicki, MD, MA, PhD.

Name of Organization: Stanford University

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: The Stanford Program on Research Rigor and Reproducibility (SPORR) is a Stanford School of Medicine hub for promoting and implementing best practices in research rigor and reproducibility (hereafter abbreviated "R&R"). SPORR has been supported by CTSA grants UL1TR003142 and UM1TR004921 (PI: Ruth O'Hara, PhD). Our philosophy is that measurement and rewards, or incentives, cannot just be at the end of the process (e.g. at promotion), but must be embedded at multiple stages throughout a scientific career. And we must build incentives that are aligned with the ultimate purpose of biomedical research, the advancement of human health.

Promotion incentives at Stanford:

1) Our Stanford CTSA SPORR program has developed a new optional Rigor and Reproducibility (R&R) CV template for researchers (https://med.stanford.edu/sporr/rewards/CVTemplate.html) that allows the highlighting of R&R practices, making it possible to evaluate faculty not just on research reports, but on how they do the research. The new additions include listing (and linking) the availability of protocols; analysis plans; analytic code; data; public accessibility of results for research outputs; registration number and result reporting for clinical trials; details of peer review activities; and the candidate's persistent identifier (e.g., ORCIDiD). We also have a list of R&R practices that can be described in a candidate's narrative statement. Details of the template can be found in our recent publication (PMCID: PMC13239307, https://doi.org/10.1111/cts.70575).

We worked closely with the Office of Academic Affairs and Academic Vice Dean to incorporate this CV template into promotion review. This format was enthusiastically received and this coming year we have permission to promote the use of the new CV template amongst department chairs and promotion committees, as well as highlighting the practices in early faculty orientations. This is critical, as the practices highlighted therein need to be adopted early in a research career; a candidate up for promotion cannot go back in time to employ procedures that can be featured on an academic CV.

2) These R&R CV practices are also embedded in the education and training of graduate students, which in turn affects the faculty. This is done via a suite of courses offered to these audiences (some encouraged by the NIH T32 requirements). This year we will try to incorporate reporting on these practices in PhD students' Individual Development Plans (IDP), a formal annual review of progress for

each student that is filled out after a structured meeting with their research advisor. This ensures that both faculty and trainees are continually aware of and accountable for the kinds of practices needed to support R&R, encouraging its incorporation into the normal work flow.

3) SPORR issues annual Rigor and Reproducibility Researcher awards (RaRe) to individuals and teams who have demonstrated the use of R&R practices in a manner that demonstrably accelerates the research process, with a clear connection to improving human health. The list of winners and the reasons for the awards can be found at https://med.stanford.edu/sporr/rewards/RaRe.html.

We have also awarded seed grants for Stanford researchers to improve their R&R practices with support from the Office of the Senior Associate Dean for Research. The first round of these grants has just concluded, and the initial reports are that research processes have been demonstrably improved.

4) We are working with AI scientists to extract open science indicators (e.g., sharing of data, meta-data, code, protocols, statistical analysis plans, etc.) from all published work published by Stanford scientists and developing dashboards that could display them. Furthermore, as part of the effort of our CTSA we are also developing methods to assess the computational reproducibility of research by Stanford scientists.

We are monitoring all corrections and retractions from biomedical publications by Stanford researchers, not as an indicator of problematic research, because these can exemplify laudable self-correction practices, but as the basis for root-cause analyses, so we can see if these arise from scientific or data management procedures that the institution can help improve.

5) We are instituting an institution-wide monitoring program of the fate of all IRB-approved research. This will provide key evidence related to selective reporting. We are initiating a program to encourage all Stanford researchers conducting human subjects research to report the IRB protocol number in their publication, a practice that required by few journals. This will provide the institution, journals, and the funders and ability to confirm that any reported research is an authentic product of research that was actually approved and conducted at an institution. In the current environment where a large and rapidly increasing proportion of biomedical research reports are AI-produced fakes, this could be a critical tool to weed out such false reports, which undermine confidence in the entire enterprise. We are also working with journals to require this reporting.

Part of the Stanford CTSA networking mission is to disseminate across the CTSA Consortium, and it is our goal to do so for these practices.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

263. N/A

I am responding: On behalf of myself

Name:

Name of Organization:

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: Recognizing team-based science contributions is critical - also wondering about innovative ways to support the key role of community partners and groups as collaborators in team-based science.

5. Entrepreneurship and Translation: Yes -- and in addition to translation, what is the role of co-production with key audiences and users of research evidence, including those affected by the problem or the leaders/practitioners in systems in which evidence would be embedded?

6. Foundational Scientific Exploration: N/A

7. Public Impact: Community-engaged forms of research including participatory research are key pathways towards building public impact and trust in science. How will NIH support such approaches?

8. Other Areas: N/A

264. DataSeer

I am responding: On behalf of an organization

Name: Tim Vines

Name of Organization: DataSeer

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Other

Role - Other: Service provider

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Under half of NIH-funded research articles are reproducible – not because of methodological flaws, but because the datasets needed for the re-analysis are not publicly available*. One cannot check the reproducibility of an article without access to the underlying datasets.

In other words, the biggest single piece of infrastructure supporting 'Rigor and Reproducibility' is the same infrastructure supporting 'Data, Software, and Model Sharing'. Having access to the datasets that underpin an article allows others to begin checking reproducibility, so any actions that improve data and code sharing necessarily improve Reproducibility and Rigor. Analysis code is particularly important: our earlier study (https://peerj.com/articles/1137/) found that reproducibility was much higher when a) author provided their data and code, and b) the code reads in the provided datasets.

Only once researchers' data and code are routinely available alongside NIH-funded articles will other incentives that encourage researchers to test reproducibility of others' work be worthwhile.

*see 'Data, Software, and Model Sharing' section for more explanation

2. Data, Software, and Model Sharing: Step 1 in incentivizing the sharing of datasets, code, and software is monitoring compliance. NIH-funded researchers currently make promises about data sharing in their grant proposals, around half break those promises when the resulting research is published or when the grant ends.

How do we know? DataSeer has the most advanced Open Science monitoring system available, and we supply Open Science metrics to publishers like PLOS, Taylor & Francis, and AAAS, along with funders like the Michael J Fox Foundation and UK Research & Innovation.

Some preliminary results for data and code sharing among 1213 NIH funded articles can be found here:

https://dataseer.ai/nih/

For articles that generate new data, only 48% share any data from their article online. For articles that use command line software (and hence generate shareable code), only 38% share code on an online

repository.

The current absence of compliance monitoring has generated a downward spiral – researchers know that their data sharing is not being monitored, and hence there is no incentive to share data from subsequent articles or subsequent grants. Institutions know that they will continue to receive NIH grant money regardless of their researchers' data sharing efforts and make no effort to promote data sharing.

The widespread availability of compliance monitoring data rapidly inverts this spiral. First, monitoring generates accountability – it is immediately clear which institutions and which researchers are not complying with the data sharing expectations of their NIH grants. Complacency at the institutional level is replaced with action, particularly if the funders themselves gets in touch to ask for an explanation.

Second, monitoring also empowers intervention, as then stakeholders (including the NIH and the grantee institutions) can target their efforts to promote better data and code sharing with the groups that most need help.

Lastly, as pointed out above, the upward spiral in the sharing of research data created by compliance monitoring efficiently enables the NIH to push for and achieve progress towards greater Reproducibility and Rigor.

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

265. TK Kozai

I am responding: On behalf of myself

Name: TK Kozai

Name of Organization: Pitt

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Score rigor from what a paper says, not from where it was published. A Scientific Reasoning Quality Index scores the nine steps of the scientific method, from observation through reporting. At the analysis step, fixed statistical rules do the checking. Does the test match the data type? Does the analysis account for repeated measures on the same animals or cells? Are the reporting elements the field requires present? Those checks need no training corpus, because statistics already fixes the right answer.

Reproducibility needs a reward, so a calibration axis credits a correctly-reported negative result and a candidly-stated limitation as much as a positive finding. Citations accumulate for clean positive results, and that is how nulls get buried and claims get stretched. Crediting honest reporting supplies what citation counts leave out. On feasibility, these scores read the paper text, so they work in any field. Where a literature is thinly indexed, the score comes with lower confidence attached.

2. Data, Software, and Model Sharing: A shared dataset, reagent, or tool is invisible to current metrics unless someone wraps it in a citable paper. A researcher who releases a widely-used resource therefore earns less credit than one who publishes small papers steadily. We propose crediting shared artifacts by how much independent reuse they generate, counting use by people who have never co-authored with the originator, so reuse inside one's own group cannot inflate the number. Credit then follows the use an artifact earns in other laboratories. This needs persistent identifiers and a way to track reuse. Both work well for data and software and less well for physical reagents, so the score reports its coverage rather than pretending to be uniform.

3. Training and Mentorship: Mentorship is hard to measure because the product is a person, not a paper. Doing science means running the scientific method, all nine steps from observation to reporting. A trainee grows as a scientist as they run more of those steps without help. So measure progression toward independent fundability. Track how many steps a trainee executes alone, and track movement toward a body of work a review panel would judge independent. This makes the middle steps visible, meaning design, execution, and analysis. Those are the steps that connect ideas to written results, and many trainees never watch a mentor perform them. A mentor who produces independent scientists then scores above a mentor who produces skilled hands. Because the measure reads change over time rather than total output, it works at every career stage.

4. Collaboration: Team science breaks author-level credit, because a citation counts a paper the same way no matter who did what. A Co-Author Contribution Index credits each author for the scientific-method steps that author actually engaged. It reads contribution structure, not author position and not demographics, so credit follows the distribution of real work. An Independent Recognition Indicator then separates citations from strangers from citations by former collaborators. It computes this per paper and takes the median, so one heavily-cited paper cannot dominate. Both report a distribution rather than a rank, so a researcher with a large network is described, not punished.

5. Entrepreneurship and Translation: Translation matters, and this framework credits it by rewarding methods and tools that other groups adopt. One unintended consequence needs attention. NIH is moving toward holistic, priority-aligned funding. If a measure treats near-term translational appeal as impact, it will over-reward work that pays off soon and cut support for the basic research that later translation depends on. So translational consequence is scored on the same forward-looking axis as foundational consequence, with wide uncertainty and an explicit allowance for work a field has not absorbed yet. Translation gets credit without becoming the only kind of impact the system can see.

6. Foundational Scientific Exploration: This category is the one the framework addresses most directly, and current funding conditions make it urgent. Before NIH stopped publishing cutoffs, institute paylines sat near the 5th percentile. Success rates for R01 grants now run below 8%, and the data are still incomplete. In November 2025, NIH replaced fixed score-based paylines with holistic decisions that weigh how well an application aligns with agency priorities.

Under that combination, scarcity plus priority weighting, review favors work that pays off soon and reads as mission-relevant. Support for deep, slow work then disappears at the priority step rather than at a published cutoff, where nobody can see it happen. Three measurable indicators address this:

A Sleeping-Beauty Trajectory Index scores the chance that a paper will be recognized late. Katalin Karikó's messenger RNA work is the anchor case, since the field ignored it for two decades.
A consequence axis predicts how much a contribution will change the field, and reports that prediction with wide uncertainty. Under the rule that incomputability scores high, work too far ahead of its field is flagged as possibly dormant, not as minor.
A Lifetime Foundational Impact Indicator, scored once at the end of a career, credits origination, delayed payoff, risk-taking, and infrastructure. Annual metrics never measure those things.

The design principle is that evaluation should push against the translational preference the funding system already applies, rather than adding to it. Foundational capacity is also a competitive position. A bipartisan national-security commission has told Congress that China's biotechnology expansion is a direct security concern. Cutting foundational funding degrades an ecosystem of trained scientists, and a competitor cannot rebuild that quickly, so the loss compounds.

CRISPR shows what these indicators would credit. In the late 1980s, federally funded scientists studied strange repeated sequences in bacterial genomes for no applied reason. The National Science Foundation gave Jennifer Doudna's laboratory more than 1.5 million dollars over her career. That work produced the gene-editing methods behind Casgevy, the first CRISPR medicine, approved in December 2023 for sickle cell disease. A metric that treated impact as near-term translational appeal would have

scored the 1990 version of that work low and ended its support. The two-axis design and the delayed-recognition index exist to prevent that outcome.

7. Public Impact: Public impact includes whether the public understands why federally funded research matters, because that understanding is what sustains funding across political cycles. A public-engagement indicator credits substantive research communication and treats it as protection for everything else, not as a vanity score. It measures verifiable engagement with the research and its consequences rather than follower counts. This is the one place the framework credits work outside the paper record. Health outcomes, clinical practice, and economic growth enter through the same forward-looking consequence axis, reported with the wide uncertainty those slow outcomes deserve. Return on public investment belongs here too, and the consequence axis is what lets a review system credit value that arrives late and compounds, rather than only the near-term output a translation-appeal proxy can register.

8. Other Areas: Two design choices apply everywhere. The framework has two arms. One diagnoses, scoring quality that current metrics do not measure. The other sets incentives, using metrics built so that the cheapest way to score well is to do the wanted thing. Every indicator advises human reviewers rather than deciding for them. Human review comes last and handles only what the automated checks cannot reach. Every score arrives with its supporting evidence and a confidence flag, so a reviewer can inspect it and overrule it.

One limit belongs in the record. The consequence axis protects basic research, and it also holds the widest uncertainty in the framework, because at review time contribution magnitude is a prediction rather than a measurement. It is reported with that uncertainty, and it needs both retrospective and forward validation before anyone uses it in a funding or promotion decision.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/rfi_response_NOT-OD-26-087.pdf

Response to NIH Request for Information NOT-OD-26-087

Measuring and Rewarding Scientific Impact

A Framework of AI-Enabled Indicators for 21st-Century Biomedical Research Assessment

Submitted by Takashi D. Y. Kozai, PhD

Ernest E. Roth Professor of Bioengineering, University of Pittsburgh

Director, B.I.O.N.I.C. Lab and UP NExT (University of Pittsburgh Neural Engineering Cross-Translation)

Date: August 18, 2026

Preface: What This Framework Is Weighed Against

In 1995, the University of Pennsylvania demoted Katalin Karikó. Her grants were unsuccessful and her papers were not cited. Over the next two decades, the messenger RNA chemistry she developed became the COVID-19 vaccines and won the 2023 Nobel Prize. The metrics used at her evaluation predicted failure at the moment her most consequential work was beginning. Her case is not an anomaly in an otherwise sound system. It is one visible instance of a baseline that routinely misvalues the science with the longest payoff. The case for a new framework rests on how bad that baseline is.

The case does not rest on the proposed metrics being good in the abstract. It rests on what they are weighed against. Three observations, taken together, decide whether building this helps or harms.

First, the status quo is not thoughtful human judgment that a flawed metric might displace. Evaluators under time pressure reach for the two most misleading numbers available. One is the raw h-index. The other is a count of papers in high-prestige journals such as Nature, Science, and Cell. The h-index is the largest number h for which a researcher has published at least h papers, each cited at least h times. It is a single figure, and it rewards steady citation accumulation. It also excludes delayed, foundational, and corrective work. Both defaults suppress the slow, foundational, corrective, and exploratory work with the longest eventual payoff. Removing a computable score does not produce careful qualitative evaluation. It produces the h-index. The choice is therefore not metric versus judgment. The choice is which computable default a busy committee reaches for.

Second, and deeper, human committee judgment is itself biased by relationship, and the current system rewards that bias. Outcomes are shaped by who trained whom, who publishes with whom, who sits on study section, and which institution and nation a researcher belongs to. A convenient number is then used to ratify a judgment already made on social grounds. This is the Matthew Effect working through networks, and its errors are not random. They are structured by who you know. Structured errors accumulate across a career and concentrate advantage in the already-advantaged, rather than averaging out the way random noise would. A content scorer reads only what is in the paper. It deliberately excludes the collaboration graph, institutional prestige, and training lineage. That trade replaces structured relationship bias with unstructured noise, which lowers total unfairness across many decisions even when any single score is imperfect.

Third, the bar a new metric must clear is not high absolute accuracy. It has two parts. The metric must predict real contribution better than the h-index does. It must also fail more gracefully, meaning its errors must not fall hardest on foundational, slow, and corrective work, as the incumbent metrics do. The incumbent is a scorer with almost no accuracy for long-term contribution, trusted mainly because habit has made it familiar. Beating that while failing more gracefully is an achievable bar, and it is the bar this framework sets for itself.

Against this baseline, the framework proposes two arms. One arm diagnoses. It scores the content of papers to detect quality the current metrics miss. The second arm sets incentives. Its metrics reward foundational exploration, replication, correction, data sharing, mentorship, public understanding, and lifetime contribution, and they are built so that gaming them produces the behavior the enterprise requires. Many of these incentive reforms need no artificial intelligence at all. They could run on human judgment and verified counts. The AI content scorers are the more powerful layer, and the riskier one.

One risk survives every argument in favor, and it must be stated. A content scorer inherits the biases of its training data. If the relationship-biased system it would replace generates its labels, it will reproduce that bias in a form that is harder to contest, because it looks objective. Whoever controls the scorer's training and calibration would then control scientific careers. The people who benefit from relationship bias now are the senior figures best placed to capture that calibration, and they could use it to confirm the existing hierarchy while appearing neutral. Governance language does not answer that risk. Architecture does. Scorers must be open-source and independently reproducible. The computation must not build lasting dossiers on individuals. No scorer should be deployed unless its training and calibration can be independently audited and reproduced.

The downside is asymmetric, and that shapes how any of this should be deployed. A poorly built or captured content scorer does not fail by doing nothing. It fails by doing harm with false authority. It encodes bias or noise and presents the result as objective computation, which is harder to contest than a human judgment. History warns that the discipline needed to keep such a system advisory is unreliable. Both the h-index and the journal impact factor arrived as one input among many. The impact factor is the average citation rate of a journal's recent articles, and it is widely used as a stand-in for the quality of any single paper published there. Habit and workload then made both of them decisive. A promise to deploy only after accuracy and openness gates are met is therefore not self-enforcing. The gates must be architectural, meaning reproducibility, no lasting dossiers, and network-blindness, rather than procedural assurances. Where a gaming or bias mode cannot be closed, this document names it rather than presenting the metric as sound, so evaluators can weight it appropriately and later work can address it.

The recommendation is deliberate and staged. The incentive-structure reforms are the primary contribution and should come first. Most need no AI and can run on human judgment and verified counts. They credit high-risk work including its failures, replication, correction, verified reuse of data and tools, trainee independence, and, at the end of a career, origination and delayed impact. The content-scoring layer is more powerful and riskier. It warrants deployment only as a gated pilot, and only where the graceful-failure and open-reproducibility preconditions are demonstrably met. If those conditions are not met, it should not be deployed at all. Section 8 states the staged recommendation in full.

The comparison that matters is against the applicable baseline, not an imagined one. Consider deploying a network-blind, gracefully-failing, openly-built scorer alongside long-horizon incentive metrics. That is more likely than not a positive contribution, because it partly dismantles the system that currently converts relationships into judgments and misallocates careers and resources. The contribution is real. The principal risk is capture of the scorer's training, and open, reproducible construction answers that risk better than promises do. This framework should be judged on that comparison, not against a baseline of neutral expert judgment that the evidence does not support.

1. Problem Statement

Current impact metrics count papers and citations for individual researchers. They do not measure what modern biomedical research actually looks like, which is collaborative, long-horizon, and mission-driven. Two failure types occur.

False negatives are work whose impact arrives decades after publication, where short-window indicators cannot see it. Karikó's demotion, described in the preface, is one instance. Tom Brock sampled Yellowstone's thermal pools in 1966 because one question interested him. How could anything live above 70 C? He isolated the microbe Thermus aquaticus. The Taq polymerase from that organism made PCR possible, and PCR in turn made the Human Genome Project possible. Richard Normann built the Utah intracortical microelectrode array at the University of Utah through the late 1980s, with NIH and DARPA support. It sat in academic laboratories for two decades before the BrainGate clinical trials began in 2004. The same architecture now restores speech in patients with ALS, and Pat Bennett reached 62 words per minute in 2023. Carl Sagan published roughly 500 peer-reviewed papers over 39 years, about one a month. Harvard denied him tenure in 1968, and the National Academy of Sciences blocked his 1992 nomination, because colleagues treated his public communication as evidence that he was not a serious scientist.

False positives are volume-first publication patterns. They produce high paper counts without durable contribution. The strategies include single-test analyses, papers that reuse public databases, papers with no primary data, and first-author output above ten papers a year. These accumulate counts that look like productivity while generating nothing the field builds on. The current proxy counts them equally with substantive work.

Both failure modes share one root cause. The metrics are pseudo-metrics, meaning indirect proxies for something no one can observe directly, long-term scientific impact. Using a proxy is not the problem. Mistaking the proxy for the thing is the problem. This proxy excludes delayed translation, deep mechanistic work without translational hooks, failure analysis, replication, and public engagement, and it rewards the gameable volume patterns described above.

The most favorable version of the incumbent does not rescue it. Consider the m-index, which divides a researcher's h-index by the number of years since their first publication. That normalization accounts for career length, so a productive early-career researcher is not penalized against a senior one. It is the closest existing metric to a rate of impact rather than a cumulative count. It still fails on delayed impact, because it depends on citations that have already accumulated. Katalin Karikó's m-index at her 1995 demotion would have been near zero, since her foundational messenger RNA work had not yet been cited. The metric would

have delivered the same verdict the raw h-index did. Normalizing for time does not help when the impact itself is delayed by decades. The failure is not that incumbent metrics ignore career stage. Every citation-based metric, whether cumulative or rate-based, cannot register impact that has not yet arrived.

Funding pressure compounds this failure. NIH funding notices fell from 756 in 2024 to about 120 in 2025, and to 14 by mid-March 2026, with competitive awards down 74%. Under that pressure, principal investigators pursue short-term translational framings and positive results to stay publishable. That shift defunds the deep characterization, failure analysis, and replication that future translation depends on.

Funding pressure has since changed form. In November 2025, NIH moved away from fixed, score-based paylines. Funding decisions are now holistic, weighing how well an application aligns with agency priorities alongside its peer-review score. Before NIH stopped publishing cutoffs, institute paylines sat near the 5th percentile. R01 success rates on incomplete current data run below 8%, down from the low 20s several years earlier. Under a holistic regime the translational preference relocates rather than disappearing. The priority-alignment step rewards payoff that is soon and mission-legible over work whose value is deep but distant. Whatever proxy stands in for impact and priority now determines funding directly. A proxy that treats near-term translational legibility as impact therefore removes support for basic research at the priority-alignment step, where no published cutoff makes the loss visible. Two separated axes address this. A consequence axis predicts contribution magnitude with wide uncertainty, and it scores work the field has not yet absorbed as unresolved rather than as minor. A calibration axis cannot raise the consequence read, so foundational research is not penalized for lacking a near-term application.

Foundational capacity is also a competitive position. Every intracortical brain-computer interface now in human trials rests on neuroscience from the 1990s and early 2000s. United States investment in extending that base slowed while the devices that depend on it remain in clinical use. The durable advantage is the research ecosystem rather than the publication count. That ecosystem means trained scientists, interdisciplinary density, and cross-institutional capacity, and a centralized, state-directed program replicates those slowly. A bipartisan national-security commission has reported to Congress that China's biotechnology expansion presents a direct security concern. An impact measure that reduces foundational funding degrades that ecosystem rather than one budget line, and ecosystem losses take a scientific generation to reverse. Measuring impact so that foundational capacity stays funded is therefore an economic and security question as well as a scientific one.

Venue prestige is the dominant informal proxy in hiring, promotion, and award decisions, and it is decoupling from scientific contribution under two pressures at once. The first is cost. Publishing gold open access in Nature currently costs $12,850, or 9,390 pounds, in article-processing charges, so venue access increasingly reflects institutional wealth and funder policy rather than the quality of the work. United States public-access policy and current grant-cost restrictions are moving to limit whether federal grants can pay those charges. The second pressure is national policy. Some countries have rewarded publication in high-impact venues directly. China offered institutional cash bonuses for Nature and Science papers until 2020, when the Ministry of Education and the Ministry of Science and Technology ordered institutions to stop the practice and to stop evaluating researchers on publication count or venue alone. As

a result, evaluation based on impact factor measures who can pay and which national policy regime a researcher sits inside, not what the work contributes. Indicators that read the paper directly would let hiring, promotion, and award committees evaluate the science rather than the venue. That is among the principal justifications for the framework proposed here.

A deeper problem sits beneath the measurement failures. Study sections reward near-term clinical translation, because it is fundable and demonstrable within a grant cycle. Public sentiment rewards the same thing, because visible cures arrive during the appropriations window that funds them. Both forces push toward the short horizon. Basic and foundational science has the opposite profile, the longest impact horizon and the weakest political constituency. Its payoff arrives after the appropriations cycle, after the administration, and sometimes after the researcher's lifetime. An evaluation system that scores all work on the same near-term axis does not stop at missing foundational science. It systematically under-invests in it, because it adds to a pressure that already over-rewards translation instead of counterbalancing it. Correcting this takes more than better measurement. It takes metrics that actively reward long-horizon, high-risk work whose returns no one can capture privately, which is the work only public funding will ever support.

This response proposes a two-arm framework, responsive to both jobs the RFI names, measuring scientific impact and rewarding it. The diagnostic arm has fourteen top-level AI-enabled indicators that detect quality the existing metrics miss. They score content rather than metadata, through computational analysis of full text and citation networks. Two of the fourteen are meta-indices, the Scientific-Reasoning Quality Index and the Engineering-Method Index, and each composes named step-level sub-indices. The incentive arm has indicators built so that the cheapest way to score high is to perform the wanted behavior, which makes them resistant to the gaming that degrades proxy metrics. This arm leads with foundational and exploration incentives. It ends with career-terminal metrics, scored once and late, that measure what no annual metric measures. Those are the fields a researcher originated, the delayed payoffs realized, the risks taken, the scientific lineage launched, the errors corrected, and the shared infrastructure built across an entire career. The incentive arm should receive heavier evaluative weight than the diagnostic arm, and the foundational and career-terminal metrics within it should receive the most. Weighting long-horizon incentives corrects a standing imbalance toward near-term impact. Only a public funder with a multi-decade mission has the mandate to correct it.

The diagnostic arm is organized into four operational groups. First, foundational-exploration indicators, meaning sleeping-beauty trajectory, mechanism-graph centrality, field-define, and gap. Second, rigor and reasoning indicators, meaning assumption-addressing, the paper-substance score, and two meta-indices. The Scientific-Reasoning Quality Index composes eleven scientific-method step-indices, and the Engineering-Method Index composes six engineering-method step-indices, since much NIH-funded work builds things rather than discovering them. Third, collaboration and integration indicators, meaning cross-field integration, co-author composition, and independent recognition. Fourth, impact-direction indicators, meaning verifiable public engagement, citing breadth, and the data-driven innovation composite. Section 5 specifies the incentive arm.

2. Notation

Throughout this document, p denotes a paper, a denotes an author, and t denotes time, typically a year. The term c(x, y) denotes a citation from x to y, and refs(p) denotes the reference list of p.

AI components are of three kinds, distinguished throughout because they differ in what they require. Rule-based checks have a fixed ground truth and need no training corpus, and the statistical-validity checks are of this kind. Corpus-relational pattern detectors have the existing literature's structure as their ground truth, and they need the corpus for retrieval but no annotation, covering novelty, field-distance, provenance, and design-overlap. Graded scorers map computed features to a scale anchored to the reference-population distribution of those same features, and the meta-index step-indices are of this kind, with structural feature detection and a scale mapping that uses no expert quality-labels. Model architecture is noted where it constrains implementation.

Each standalone indicator reports on the scale native to what it measures. That may be a probability where a probability is interpretable, as with SBTI and DDIC. It may be a field-relative z-score for standing against a reference distribution. It may be a bounded 0-to-1 factor for component scores that combine. Or it may be a raw or log-transformed count where magnitude is informative. These standalone outputs are deliberately not rescaled onto one common band. They feed calibrated committee decisions and an ensemble model rather than a single-number comparison, and forcing a fixed bounded band would compress genuine outliers, discard tail information, and reintroduce the false precision of the one legible index this framework replaces. Preserving native scale and reported uncertainty is more informative than cosmetic uniformity.

One class of indices is an explicit exception. The step-level sub-indices of the two meta-indices, meaning SRQI in Section 3.5 and EMI in Section 3.5E, must be commensurable with one another. A meta-index composite is a committee-weighted average across its step-scores, and an average is coherent only when the terms share a scale. Otherwise whichever sub-score has the largest numeric range dominates the weighting. Each step-index therefore reports on a common anchored 1-to-10 rubric, where 3 is deficient, 7 is solid, and 10 is excellent within the field's normal range. The top is left uncapped, so documented work beyond the field's normal range scores above 10 and the composite average preserves the outlier rather than compressing it. The 1-to-10 anchoring matches the range reviewers already reason in, and the uncapped top keeps a ceiling from flattening the foundational-outlier signal this framework preserves.

3. Indicator Specifications

This section specifies the diagnostic arm, the AI content-scoring layer. Per the recommendation in the preface, this arm is the more powerful and the riskier component and should be treated as a gated, optional experiment, deployed only where the graceful-failure and open-reproducibility preconditions of Sections 6 and 7 are demonstrably met. The incentive-structure reforms of Section 5, most of which require no artificial intelligence, are the primary contribution and do not depend on any indicator in this section being built.

Each indicator section presents seven parts. These are a definition of what the indicator measures, the motivation explaining why current metrics fail at this measurement and what gap the new indicator fills, a mathematical formulation, the AI specification including model architecture and training requirements, numbered implementation steps, failure modes and mitigations, and the RFI category alignment.

One distinction governs how every indicator should be read, and it is the distinction the current metrics collapse. Quality and impact are independent axes.

The quality axis is whether the work was done as science. That means the observation was real, the hypothesis falsifiable, the design controlled, the execution reproducible, the analysis sound, and the reasoning traceable. All six are assessable at publication time from the paper itself. Impact is whether the work eventually changed the field, which is knowable only later and only from what the field does with it. A careful scientist can produce high-quality work that turns out wrong or inconsequential, and careful-but-wrong is still good science, just not consequential science. An incautious one can stumble onto high-impact results through low-quality work.

The scientific-reasoning indices of Section 3.5 and the paper-substance score measure quality, meaning the reasoning was sound at publication. The trajectory and recognition indices, meaning SBTI, IRI, and the citation-structure measures, together with the career-terminal incentives, measure impact, meaning the work mattered over time. Separating the two axes is the point. The current citation-and-venue metrics conflate them, treating eventual citation as if it certified quality and treating quality as if it guaranteed impact, when neither implication holds. The reasoning score must therefore never be read as a claim about truth or importance. It measures only that the work was done as science, which is the most any at-publication metric can measure, and the thing citation and venue do not measure at all.

3.1 Sleeping-Beauty Trajectory Index (SBTI)

Definition.

A sleeping beauty, in bibliometrics, is a paper that goes largely uncited for years or decades and then accumulates citations rapidly once the field catches up to it. This index estimates the probability that a paper will follow that trajectory, computed at publication time from early-trajectory features and content-level scoring rather than waiting for the delayed citations to appear.

Formally, it is the probability that a paper's downstream citation accumulation will exceed a field-specific threshold beyond a horizon Δt years from publication, computed from early-trajectory features and content-level scoring.

Motivation.

Current metrics rely on cumulative citations within a 5-to-7-year window, which dominates h-index calculation (Hirsch 2005) and grant-review evaluation. Karikó's 1995 mRNA chemistry work had near-zero early citation despite eventually founding the COVID-19 vaccine platform, and the current metric system would have predicted her career failure at her 1995 evaluation. The Sleeping-Beauty Trajectory Index fills this gap by scoring publication-time features that predict long-horizon citation accumulation, providing decision-relevant signal during evaluation rather than waiting decades for the citation pattern to emerge.

Formulation.

$$SBTI(p,\ t) = P(\ C(p,\ t + \Delta t) > k \mid F(p,\ t)\ )$$

where $C(p, t + \Delta t)$ is cumulative citations to p by year $t + \Delta t$, k is the 90th percentile of the field-specific citation distribution at the chosen horizon, and $F(p, t)$ is the feature vector at publication time including early citation trajectory shape, mechanism-graph centrality, methodological specificity, and gap-filling status.

AI specification.

A gradient-boosted decision tree or LSTM model, trained on historical late-blooming papers. Training targets come from computed disruption indices rather than human annotation. Those indices measure whether a paper eclipses the prior work it builds on, since a disruptive paper is cited in place of its own references while an incremental one is cited alongside them. The targets are the Funk and Owen-Smith 2017 CD index and the Wu and colleagues 2019 disruption score in Nature, both computed from citation structure. Retrospective validation requires reproducing the late-blooming signal on three papers held out of training, Karikó and Weissman 2005 in Immunity, Brock and Freeze 1969 in J Bacteriol, and Maynard and colleagues 1997 in Electroencephalogr Clin Neurophysiol.

The feature vector combines content-scored inputs with early-trajectory shape. The content-scored inputs are mechanism-graph centrality, methodological specificity, and gap-filling status. The trajectory inputs are citation velocity and citing-field breadth at two and five years. The model predicts the disruption label from these features rather than from author, institution, or venue identity, which are excluded so a prestigious byline cannot lift the score. This index is difficult to game directly, because its target realizes 10 or more years out and its content features depend on other researchers' independent downstream choices rather than on anything the author writes. The only path to a high score is producing work with genuine mechanistic centrality and specificity that later matures. The output is a calibrated probability with a 90% confidence interval.

Implementation.

1. Construct training corpus from Web of Science, OpenAlex, or Microsoft Academic with full-text access.
2. Label papers by disruption score at 20-year horizon (where possible) or by curated panels of foundational and non-foundational exemplars.
3. Extract early-trajectory features at t+2 and t+5 years post-publication.

4. Train model to predict 10-year and 20-year citation accumulation conditional on early features.
5. Apply to new submissions with calibrated probability output and 90% confidence intervals.

Failure mode and mitigation.

The model can over-fit to obscurity itself, generating false positives for unimportant work with low early citation. Mitigation requires joint conditions with mechanism-graph centrality and paper-substance score above field-median thresholds. A paper qualifies as a sleeping-beauty candidate only when low early citation pairs with high content scoring on independent dimensions.

RFI category alignment.

Foundational Scientific Exploration.

3.2 Mechanism-Graph Centrality Index (MGCI)

Definition.

Centrality of a paper's contributed concepts within a knowledge graph of mechanistic relationships extracted from biomedical literature.

Motivation.

Current metrics treat all citations equally, counting a citation to a deep mechanism paper the same as a citation to a translational hook paper. This treats foundational mechanistic characterization as equivalent to shallow translation when in fact the mechanism work is the substrate downstream applications require. The Mechanism-Graph Centrality Index fills this gap by scoring how foundational a paper's contributed concepts are within the knowledge graph of mechanistic relationships, distinguishing structural contributions from incremental elaboration.

Formulation.

$$MGCI(p) = \Sigma \text{ over } c \in Concepts(p) \text{ of } PageRank(c, G_mech) \cdot attribution(p, c)$$

where *G_mech = (V, E)* is the knowledge graph with vertices *V* (mechanistic concepts extracted from full-text papers via biomedical LLM) and edges *E* (causal and dependency relationships extracted from text), and *attribution(p, c)* is the fraction of paper *p*'s contribution to concept *c* (1 if origination, decreasing for elaboration).

AI specification.

A two-stage system. First, a biomedical LLM (BioGPT, PubMedBERT, or domain-tuned successor) performs semantic relation extraction on full text to identify causal and dependency edges between mechanistic concepts. Second, a graph neural network computes centrality with PageRank or eigenvector-centrality variants. Edge extraction uses dependency-parsing plus relation-classification, with confidence thresholds for edge inclusion.

Parameters and thresholds.

Edges enter the graph only above a relation-classification confidence of 0.7. That threshold is set so extraction precision against a held-out gold-reference edge set exceeds 0.8, since a false edge distorts centrality more than a missed one, so precision is favored. The gold reference checks factual extraction correctness, meaning whether the stated mechanistic relation is actually present in the text, rather than any judgment of the relation's importance. PageRank uses the standard 0.85 damping factor over the directed mechanism graph, recomputed quarterly as new papers add edges.

Attribution is set to 1.0 for the originating paper of a concept, and decays for elaborating papers by their semantic distance from the originating definition, so centrality credit concentrates on origination rather than restatement. Paper-level centrality is the sum of attributed centrality across the concepts a paper contributes, field-normalized to a z-score against same-field same-era papers. An audit samples 10% of extracted subgraphs each year against a gold reference, which bounds drift in edge extraction and checks the factual correctness of extracted relations rather than scoring importance. Centrality outputs are reported with confidence intervals derived from edge-confidence propagation.

Implementation.

6. Build candidate concept dictionary from UMLS, MeSH, Gene Ontology, and field-specific ontologies.
7. Run LLM relation extraction across full-text corpus to populate edges with confidence scoring.
8. Compute centrality scores per concept at quarterly intervals.
9. Aggregate concept-level centrality back to papers as their concept-weighted contribution.
10. Update graph as new papers add edges, with concept disambiguation across emerging fields.
11. Calibrate centrality outputs against the NIH-funded reference population stratified by field per Section 3.5.0 Element 4, so a paper's mechanism-graph centrality is scored relative to its own field's connectivity norms rather than a global standard.

Failure mode and mitigation.

Concept extraction errors propagate to graph structure, particularly in emerging fields where terminology has not stabilized. Two steps mitigate this. Audit extraction accuracy each year against a gold reference on a 10% sample of subgraphs, and report confidence intervals on centrality outputs. Flag low-confidence extractions and down-weight them rather than letting a human scorer override them, so the mitigation stays computational and adds no per-paper judgment.

Direct gaming is limited, because a concept's centrality is determined by how the rest of the literature connects to it rather than by anything the author writes, so an author cannot inflate the score by asserting importance. The residual vector is manufactured citation to a coined concept by a coordinated group. That is handled as in the field-define indicator, by weighting concept connections toward independent sources with no co-authorship tie, so within-clique reinforcement does not raise centrality.

RFI category alignment.

Foundational Scientific Exploration.

3.3 Assumption-Addressing Indicator (AAI)

Definition.

A publication-time score of how well a paper identifies an incorrect or unexamined assumption in the prior literature and addresses it, across three modes, demonstrating that a prior approach or premise fails, correcting a wrong assumption directly, or building a new approach past a documented failure. It replaces the separate failure-analysis and corrective-contribution indicators of earlier framework versions, whose cores were the same act viewed two ways, and it scores the quality of the engagement rather than any downstream effect on the field.

Motivation.

Current metrics systematically penalize the work that confronts a field's mistakes. High-impact journals select against negative and corrective findings, and the resulting citation metrics follow, so incorrect assumptions go unexamined and failure modes go uncharacterized. Framing this contribution as field-killing, meaning whether the work ended a line of inquiry, sets the wrong and unmeasurable bar. The right bar is how well the work addresses an incorrect assumption, which is the ordinary and essential mechanism of scientific self-correction, and which no current metric credits. This indicator is deliberately publication-time and corpus-relational. It reads the paper's own engagement with the prior literature rather than waiting on downstream citations, and it credits the common high-value case that pure-correction framing missed, a paper that takes a documented failure and builds a new design past it.

Formulation and AI specification.

This indicator applies the Section 3.5.0 Element 2 machinery to the paper's target assumption, scores three features, and then classifies the mode.

Assumption identification locates a specific assumption, premise, or approach in the same-year-and-prior corpus that the paper treats as incorrect or unexamined. Credit requires showing the assumption was actually present and supported in the prior corpus, rather than characterized as prevailing by the paper itself, which blocks the strawman vector. Confrontation quality scores whether the paper engages that assumption directly with evidence, by testing it, showing where it breaks, or identifying the variable that invalidates it, reusing the conflict-resolution and dogma-engagement machinery, as against asserting the assumption is wrong. Resolution scores what the paper does about it and classifies the mode as one of three. Failure-demonstration shows empirically that the premise does not hold. Correction overturns the assumption with a replacement. Building-past is the therefore-we-tried-this-new-design case, where the paper takes the documented failure as a premise and constructs a new approach.

The building-past mode is first-class, since building forward on a corrected assumption is the common workhorse the field advances on. The strength of the addressed assumption is weighted by its recency and support in the corpus, so confronting a well-supported recent assumption scores above knocking down a fringe one.

Implementation.

12. Extract the paper's target assumption, its confrontation, and its resolution using the Element 1 extraction model.
13. Verify the assumption is actually present and supported in the same-year-and-prior corpus via Element 2 retrieval, weighting by its recency-and-citation support and rejecting assumptions the corpus shows were already doubted.
14. Score confrontation quality by the conflict-resolution and dogma-engagement checks, requiring direct evidence-based engagement rather than assertion.
15. Classify the resolution mode (failure-demonstration, correction, building-past) and score the resolution, crediting building-past as a first-class mode.
16. Map the three features to the score and normalize against the Element 4 reference stratum, reporting the mode alongside the score. The correctness of the challenge is not scored here and is deferred to the career-terminal Assumption-Correctness Indicator (Section 5.4.2).

Failure mode and gaming defense.

Manufacturing a strawman assumption that nobody actually held is the dominant gaming vector. It is blocked because assumption identification uses the corpus support for the targeted assumption rather than the paper's characterization of it, so an assumption the corpus shows was already contested cannot be dressed as prevailing dogma. Framing an ordinary incremental disagreement as addressing a wrong assumption is the second vector, defended by the confrontation-quality requirement of direct evidence-based engagement. One boundary is intrinsic and stated. This indicator measures the quality of the assumption-addressing act, not whether the assumption was in fact wrong, since a paper can confront an assumption well and still be mistaken. Correctness is therefore deferred to the career-terminal metric and to eventual replication.

RFI category alignment.

Rigor and Reproducibility (primary), Foundational Scientific Exploration (secondary).

3.4 Paper-Substance Score (PSS)

Definition.

A composite score of methodological complexity, data novelty, statistical rigor, and primary-data ratio, calibrated to distinguish substantive contributions from volume-first publication patterns.

Motivation.

Current metrics count papers regardless of underlying methodological substance. A paper with a single statistical test on a reused public dataset therefore counts the same as a paper with extensive primary data collection, multiple independent analyses, and rigorous statistical correction.

That produces the volume-first failure mode, where high-output strategies accumulate paper counts, ten or more first-author papers a year, without contributing substantive work the field builds on. This score measures methodological complexity, data novelty, statistical rigor, and

primary-data ratio as content-level features, which distinguishes substantive contribution from publication-velocity gaming.

Formulation.

$$PSS(p) = w_1 \cdot M(p) + w_2 \cdot D(p) + w_3 \cdot S(p) + w_4 \cdot R(p)$$

where *M(p)* is methodological complexity (LLM-evaluated breadth and depth of methods in the Methods section), *D(p)* is data novelty (proportion of dataset newly generated versus reused from public sources or collaborator data), *S(p)* is statistical rigor (number of independent tests, multiple-comparison correction reporting, effect-size reporting, and pre-registration where applicable), *R(p)* is the primary-data ratio (fraction of analyses based on newly collected versus retrieved data), and weights $w_1 \ldots w_4$ are tuned empirically against retrospective disruption signal.

AI specification.

A composite whose four components are computed from structural evidence and fixed rules rather than from a human-labeled quality corpus, so this score detects the volume-first pattern computationally rather than reproducing expert taste.

Data novelty is computed by matching the paper's datasets against public-repository fingerprints and collaborator-data provenance. A reused GEO or dbGaP dataset is therefore detected as reuse regardless of how the methods describe it, and the primary-data ratio follows from the same provenance matching rather than from the paper's self-description. Methodological complexity counts the number and interdependence of distinct experimental or analytic procedures actually executed, which is a structural count from the methods. Statistical rigor reuses the analysis-rigor checks of Section 3.5.8 in full, meaning the test-to-data-type match, dependence-structure and pseudoreplication detection, claim-alignment, and the silence-counts-as-deficiency disclosure audit, rather than re-specifying a weaker parallel version.

The volume-first archetype is a single test on a reused public dataset with no primary collection. It is caught by the conjunction of low data novelty, low primary-data ratio, low methodological complexity, and a thin disclosure profile. None of the four components requires an annotated good-versus-bad training corpus, since each reads provenance, structure, or the fixed rules of statistics. For a theory, review, or perspective paper the primary-data-ratio and statistical-rigor components are not scored, per the analytical-mode specification of Section 3.5A, so a synthesis paper's substance is measured by its gap, novelty, and integration signals rather than penalized for primary-data collection it did not involve.

Implementation.

17. Compute data novelty and primary-data ratio by matching the paper's datasets against public-repository fingerprints and collaborator-data provenance (Element 2 corpus).
18. Compute methodological complexity as the structural count of distinct executed procedures extracted from the methods.
19. Compute statistical rigor by invoking the analysis-rigor checks of Section 3.5.8, reusing that index's output rather than a separate scorer.

20. Combine the four components, publish per-component sub-scores for transparency, and apply PSS as a calibrator across other indicators.
21. Calibrate each component against the NIH-funded reference stratum per Section 3.5.0 Element 4, so substance is scored relative to field norms rather than a global standard.

Failure mode and mitigation.

Engineering a paper to score well without substance is the gaming vector, and the components constrain it because they read provenance and structure. Raising data novelty or the primary-data ratio requires actually generating primary data rather than describing it. Raising methodological complexity requires actually executing more interdependent procedures. Raising statistical rigor requires actually satisfying the Section 3.5.8 checks, including full disclosure. The residual vector is manufacturing apparent methodological complexity through procedures that add count without contribution. Weighting complexity by interdependence rather than raw count mitigates that, as does conjunction with the other components, so complexity without novelty or primary data does not lift the composite. Periodic external audit by an independent bibliometric panel, together with public per-component sub-scores, keeps the composite inspectable.

RFI category alignment.

Cross-cutting calibrator applicable to all RFI categories, with primary alignment to Rigor and Reproducibility and Foundational Scientific Exploration.

3.5 Scientific-Reasoning Quality Index (SRQI), a meta-index over scientific-method steps

Definition.

A meta-index that composes eleven named step-level indices into a reported profile of the scientific reasoning a paper demonstrates. Nine of the eleven correspond to the canonical steps of the scientific method, meaning observation, question, hypothesis, prediction, experimental design, execution, analysis, interpretation, and reporting. Two capture reasoning qualities that cut across those steps and are also scored at the network level by standalone indicators. Those two are gap identification, paired with the Gap Index of Section 3.7, and synthesis, paired with the Cross-Field Integration Index of Section 3.9.

This meta-index stands to its step-indices as the Data-Driven Innovation Composite stands to the diagnostic arm, meaning a composite over independently-specified sub-indices. It is reported as the vector of its eleven step-scores rather than collapsed into a single number.

Motivation.

Current metrics measure the outcome of reasoning, meaning the citations a paper eventually receives, but never the reasoning itself. A paper with rigorous hypothesis structure, adequate controls, and calibrated interpretation is therefore scored identically to a paper with the same eventual citation count and weaker reasoning. Venue prestige operates as the informal proxy for reasoning quality, and venue prestige is decoupling from contribution under the pay-to-play pressures described in Section 1. This meta-index scores the reasoning operations directly from the paper text.

Decomposing into named step-indices rather than a flat multi-component score is deliberate. Each step of the scientific method is a distinct operation that a paper can perform well or poorly independent of the others. Each therefore warrants its own definition, formulation, evidence rubric, and gaming defense, rather than being buried as an unnamed component.

A single collapsed number would reward proximity to one template and homogenize the field toward one reasoning style. Step-level reporting instead preserves the legitimacy of different scientific modes. An exploratory foundational study can score high on gap identification and synthesis while scoring lower on immediate analytic closure.

Formulation.

$$SRQI(p) = (SObs, SGap, SQ, SHyp, SPred, SDes, SExec, SAna, SInt, SSyn, SRep)$$

reported as the vector of the eleven step-indices defined in Sections 3.5.1 through 3.5.11, each on the common anchored 1-to-10 rubric (uncapped above 10 for documented beyond-normal work) specified in the Notation section, with a confidence interval. The common scale is required because the composite is a committee-weighted average across steps, which is coherent only when the steps share a scale. No canonical weighting across steps is specified. Evaluating committees select role-appropriate weights, so a theory-focused search weights question formulation and synthesis while a methods or core-facility hire weights design and execution, which converts the anti-homogenization safeguard (Section 6.7) into a structural feature of the meta-index rather than an external qualification. The composite reports the weighted average alongside the full step vector, never in place of it, and any above-10 step-scores are surfaced so a committee sees where a paper broke the field's normal range.

AI specification (shared across step-indices).

Each step-index is a scorer that keys on the structural evidence the step requires rather than on claim language, field-stratified because step norms differ across experimental, computational, and theoretical work, and outputting the common anchored 1-to-10 score (uncapped above 10) with a confidence interval. The scale is anchored to the reference population, not to expert opinion, so a 7 is the median computed feature value in the field stratum. A 10 is the 95th percentile, and above-10 is beyond the observed range. That score reports where a paper's computed features fall relative to the field's actual distribution of those same features rather than where an annotator judged it to sit. Surface features that are cheap to fabricate (claim phrases, section headers, boilerplate) are held out of every step-index feature set, so the cheapest path to a high score is to perform the operation rather than name it. The step-index specifications below give the step-specific evidence each scorer uses, the feature-to-score mapping, the rubric anchors, and the step-specific gaming defense.

Failure mode and mitigation (meta-index level).

The meta-index inherits the correlation among its steps, since design and analysis co-vary, as do gap identification and synthesis. The composite must therefore report per-step sub-scores and attribute shared signal transparently, rather than presenting correlated steps as independent evidence. Homogenization is addressed by mandatory per-step reporting and committee-selected weighting. Gaming is addressed at the step level, since each step-index requires its own structural evidence, and a paper cannot raise the composite without performing the steps it claims.

RFI category alignment.

Rigor and Reproducibility (primary), Foundational Scientific Exploration (secondary), cross-cutting content calibrator.

3.5.0 Shared Implementation and Calibration Protocol

Every step-index in Sections 3.5 and 3.5E uses the common protocol specified here. Each step section below states only its own features and rubric, and inherits the seven elements defined once here. The components and thresholds named below are starting values, chosen because they can be built today. They are calibratable rather than fixed, and each should be tuned during calibration against the acceptance criteria.

Element 1, unit of analysis and extraction.

The scored unit is the claim, meaning a single asserted empirical or reasoning statement. A paper is broken into its claims before scoring, rather than scored as one undifferentiated whole. A claim-extraction model segments the full text into candidate claims. It tags each claim by type, meaning observation, gap, hypothesis, prediction, design, execution, analysis, interpretation, synthesis, or reporting, and by the section it came from. That model is a fine-tuned long-context transformer, either GPT-4-class or a domain-adapted open model such as a biomedical Llama variant. Each step-index scores only the claims of its own type. The paper-level step-score is then the weighted aggregate of its claim scores, weighted by citation or salience, with the primary claim, meaning the abstract's central assertion, weighted highest. Before scoring proceeds, extraction is validated for boundary agreement, meaning where one claim ends and the next begins. That is a mechanical text-structure task rather than a quality judgment, and it must reach an F1 above 0.80.

Element 2, reference corpus and retrieval.

The reference corpus is OpenAlex, which holds roughly 250 million works under an open license, with abstracts, citation edges, and institution and author metadata. It is mirrored locally for the compute described in Element 7. PubMed supplies biomedical abstracts, and Semantic Scholar supplies citation-context sentences. Retrieval works as follows. Each claim is embedded with a scientific sentence-embedding model, either SPECTER2 or a comparable citation-informed encoder. The system retrieves the 50 nearest prior claims by cosine similarity, and counts as near neighbors those above a cosine threshold of 0.75, tuned per field during calibration. A natural-language-inference model then classifies each neighbor as entailing, contradicting, or neutral to the claim. That model is a fine-tuned DeBERTa-v3 or a biomedical equivalent. A claim counts as already present when any prior neighbor entails it with confidence above 0.85. A claim counts as linked to a gap when a prior neighbor leaves the claim's target open. Field distance is the shortest path between the claim's field and each connected field in the OpenAlex concept co-citation graph, normalized to the distribution of field-pair distances. An embedding-cosine distance is reported alongside it as a robustness check.

Grounding-evidence requirement (applies across all elements).

No step may assign a score by pattern-matching to summary features, surface templates, or presumed absence. Scoring the summary of the evidence, rather than the evidence, repeats the error this framework indicts in the h-index, substituting a countable proxy for the thing

measured. Each score must instead be grounded in the primary evidence its claim rests on. That evidence is emitted and checked before the score issues. Three hard gates follow:

22. A novelty score is invalid unless two things happened. The same-year-or-prior neighbor set of Element 2 was actually retrieved, and its grounding sentences were emitted. Novelty is therefore never asserted from presumed absence of prior work. That is the false-novelty failure the Section 6.8 check catches, prevented here at the scorer's own hand.
23. A fidelity score is invalid unless the scorer examined a sample of the primary data, meaning the traces, waveforms, or images the summary statistics come from, and judged it physically and physiologically plausible. A recording claim therefore earns no credit from yield, signal-to-noise, and amplitude-stability summaries alone, because the underlying waveforms may be non-physiological in shape or may lack refractory-period validation. Sometimes a paper supplies too little primary data to support the claim it makes. That insufficiency is scored as a deficiency rather than assigned to cannot-verify, because the burden of evidence sits with the paper, and an unsupported claim is a shortfall the paper owns. The size of the debit tracks the gap between how strong the claim is and how much evidence supports it. A modest claim on modest data is not penalized, while a categorical claim on the same data is. Cannot-verify is reserved for one narrow case, where the paper contains sufficient evidence but the scorer cannot reach it, for example inside a figure it cannot parse. Even then, cannot-verify caps the fidelity score below what a readable and sufficient artifact would earn. Data that is missing or unreadable can never score as high as data that is present and adequate.
24. A rigor score must first identify how the claim is warranted empirically, then apply the matching rigor criteria. An observationally-warranted claim is therefore scored by observational-rigor criteria. It is not marked deficient for lacking the inferential statistics an experimental claim would require.

These gates prevent one common failure, scoring the summary of the evidence in place of the evidence itself. Each gate names the primary artifact that must exist before its score issues, meaning a retrieved neighbor set, an examined data sample, or an identified warrant. The gates bind every AI-based scoring component in this framework. That includes the SRQI and EMI step-indices, which use this full protocol, and equally the standalone AI indicators that read content against priors, meaning AAI, MGCI, GI, FDI, CFII, VPEI, and the analytical-mode floor. Section 3.15 gives the full enumeration and marks the rule-based boundary. Every one of those components scores by comparing a claim either to prior work or to a rigor standard, so all of them face the same three failure modes, scoring from summary features, presentation-gaming, and prior-poverty in young fields. Each therefore inherits three things rather than reimplementing or omitting them. It inherits the grounding-evidence gates above. It inherits the held-out-feature discipline, which bars a claim's presentation strength from raising its score. And it inherits the always-on adjacent-field prior borrowing and per-claim prior-density weighting of Section 6.9. Defining these guards once and binding them by scope follows the same inherit-once design as the seven elements. It also prevents an indicator from silently omitting a guard the others include.

Gate-contract interface, the build-phase specification every AI metric must instantiate.

Whether a guard runs correctly is settled during the build, and the builder works from this specification. A metric specified only as applying another metric's machinery therefore gets built to whatever local standard the builder infers, which under time pressure is the minimum that returns a number. To stop a gate from being built as a checkbox, every AI-based metric must instantiate a gate contract with four required fields. None may be left blank. A metric with any blank field is not build-ready and is not deployed. The four fields follow:

25. Primary grounding artifact. This names the specific evidence the score must rest on and must emit before the score issues, stated concretely for this metric rather than inherited by reference. For a novelty-bearing metric it is the retrieved prior-neighbor set. For a fidelity-bearing metric it is the examined primary-data sample. For a rigor-bearing metric it is the identified warrant with its checked criteria.
26. Retrieval scope. This names the corpus the metric retrieves against. It must include the target subfield and the adjacent fields reached by the concept-graph path of Element 2. Adjacent-field borrowing is therefore an instruction the builder wires in, not a principle the builder may quietly drop.
27. Held-out features. This lists the presentation and surface signals barred from contributing to the score, given concretely for this metric's content. The rule against presentation-gaming is therefore code the builder writes, not a value the builder is trusted to honor.
28. Per-metric acceptance test. This names the known-outcome cases the built scorer must reproduce before deployment, stated for this metric's own failure mode. A scorer whose gate runs cosmetically then fails its own test rather than passing silently.

The emitted artifact must actually constrain the score, and ablation verifies that it does. A builder can otherwise satisfy the letter of the contract by emitting an artifact the score never uses. So removing or perturbing the artifact must change the score by a stated minimum. A metric whose score does not move when its grounding artifact is withheld has a cosmetic artifact and fails the contract. The acceptance test is the essential field, because it is what detects a contract filled in form but not in function. It has no default and no waiver.

Staged deployment depends on every AI-based metric passing its own acceptance test, not on an aggregate score. An aggregate would let a few well-built metrics offset several cosmetic ones while the composite still scores as acceptable. A metric that fails or omits its acceptance test is disabled rather than shipped. The framework therefore deploys the AI metrics that are real and withholds those that are not, which matches the staged and conditional posture the preface takes toward the framework as a whole. Section 3.15 gives the filled contract for every AI-based metric. Any metric missing from that section is named there as rule-based or constitutive, with its reason. Fixing the boundary between AI-scored and rule-computed metrics in the specification, rather than leaving it to a builder, blocks one evasion, reclassifying an AI metric as rule-based to avoid the gate.

Element 3, feature-to-score function.

Each step-index computes its own feature vector, defined in its section. It then maps that vector to the 1-to-10 score by where the vector sits in the reference-population distribution of the same computed features. The anchors at 3, 7, and 10 are therefore distributional cut points, roughly the lower range, the median, and the 95th percentile of the field stratum. They are not expert-

assigned grades. The mapping needs no human quality-labels, and that is deliberate. A scale learned from expert judgments of what counts as good would reintroduce the relationship-laundering this framework avoids, this time at the calibration step. A scale anchored to the field's own computed feature distribution has no per-paper human input to capture. Some steps have two dimensions, as SObs has fidelity and novelty. Each dimension is percentile-scored separately, and the step-score combines them at a stated weight. That weight starts equal and is adjustable only through transparent governance, never by per-paper tuning. Both dimension sub-scores are reported, so a committee sees the composition. Scores above 10 arise two ways, when a paper's computed features exceed the reference range, and when the incomputability detector each relevant step defines fires. The tail is therefore produced by a stated rule rather than by unbounded model output.

Element 4, calibration across NIH-funded science.

The reference population for normalization is every publication acknowledging NIH funding in the trailing ten years, identified through the NIH acknowledgment and RePORTER linkage. Papers are stratified into fields by primary study-section assignment where that is available, and otherwise by OpenAlex top-level concept, targeting at least 2000 papers per stratum. Within each stratum, the 1-to-10 rubric is anchored directly to the computed feature distribution. A 7 is the stratum median, and a 10 is approximately the 95th percentile of the same features across the reference population. The anchors are therefore read off the distribution rather than estimated from an annotated sample. The rate of above-10 scores is monitored per stratum, and the expected rate is below roughly 2% of claims. A stratum whose rate drifts well above that is re-examined, because drift indicates either miscalibration or a real shift in the field's normal range that the reference distribution must accommodate. Recalibration runs annually against the updated reference population to track changing field norms. Because the anchors are distributional, recalibration is a recomputation rather than a re-annotation.

Element 5, feature-validation protocol.

Feature validity means that a high computed feature value really does track the quality its step names. That validity is established by retrospective outcome-controls rather than by expert annotation. The scorer must reproduce the correct signal on held-out cases whose outcomes are already known. The positive controls are the eventually-vindicated work the preface names, and the negative controls are the volume-first patterns. Features are therefore validated against what the field's later evidence showed, not against what a panel believed at scoring time. This removes the human-label capture surface entirely, and with it the need for rater-bias correction, because there are no annotator judgments to debias. Human judgment enters at one point only, in the small and auditable governance decision of which retrospective cases count as positive and negative controls. That decision is made once and openly rather than per paper, and the control set is published so the validation can be reproduced independently. The acceptance gate is that the scorer recovers the outcome-controls at a stated rate within each field stratum before deployment.

Element 6, validation and acceptance criteria.

A step scorer is accepted only when it clears five gates. First, it must recover the retrospective outcome-controls within each field stratum at a stated rate. That replaces any correlation-with-expert-consensus criterion, since this framework uses no expert consensus scores. Second, it

must pass the graceful-failure gate of Section 7, meaning its errors must not correlate with suppression of historically-vindicated foundational work. That is tested on the retrospective positive controls, Karikó, Brock, and Maynard, and on the volume-first negative controls. Third, its above-10 detector must hold a false-positive rate below 5%. That is measured by open adjudication of flagged incomputability cases against the mundane alternatives, meaning niche subfield, poor indexing, or jargon mismatch, with the procedure and cases published so the check can be reproduced. Fourth, the field-distance signal must be robust under the alternative distance measure, with the two measures agreeing on the ordinal ranking of claims above a rank correlation of 0.8. Fifth, it must pass the grounding-evidence gate stated above, meaning the scorer emitted and checked the primary evidence behind each score, the neighbor set for novelty, the data sample for fidelity, and the identified warrant for rigor, rather than scoring from summary features. A scorer failing any gate is not deployed, per the staged recommendation.

Element 7, compute and data access.

The build needs three things. It needs a local mirror of OpenAlex, on the order of hundreds of gigabytes for metadata and citation edges. It needs precomputed claim embeddings for the reference corpus, meaning billions of claim vectors at the embedding model's dimension, on the order of tens of terabytes, amortized across all step-indices because retrieval is shared. And it needs per-claim inference, whose cost is dominated by the extraction, entailment, and scoring passes. Scoring the full NIH-funded corpus is a batch job over tens of millions of papers with tens of claims each, feasible on a modest GPU cluster over weeks. A pilot on one investigator or a small cohort runs on a single workstation. Author, institution, and venue identity are excluded from every scorer's features at inference, and no lasting per-author dossier is kept, per Section 6.

3.5.1 Observation Quality Index (SObs)

Definition.

The fidelity and novelty of the empirical observations a paper reports. Fidelity combines the precision and the completeness of the reported observation. Novelty measures how far that observation departs from the prior literature and how it connects to it. The index treats observation as the foundational and most differentiating scientific step.

Motivation.

Observation is where genuine discovery most often originates, and it is the step outcome metrics reward least. A citation count cannot tell a paper that reported what it observed from one that saw selectively. This index scores whether observations are reported with the precision, artifact controls, and completeness that let others trust them and build on them.

Formulation and AI specification.

SObs scores two dimensions, both instantiating the shared protocol of Section 3.5.0. The fidelity dimension reads four features from the claim's supporting results and methods. The first is measurement precision, a binary-plus-degree feature for stated units, error bars, and instrument resolution. The second is observational completeness, the ratio of reported to methods-implied measurements, which captures whether negative and off-target observations appear alongside

the positive finding. The third is artifact control, meaning the presence of explicit confound discrimination. The fourth is primary-data plausibility.

The primary-data-plausibility feature examines a sample of the primary data behind the summary statistics, meaning the raw traces, waveforms, or images, and judges it for physical and physiological plausibility. It is a hard gate on the fidelity score, per Section 3.5.0. A neural-recording claim earns credit only when the reported waveforms show physiological action-potential morphology and refractory-period structure, rather than the non-physiological shapes that signal motion artifact. Yield, signal-to-noise, and amplitude-stability summaries can all stay high while the underlying signal is artifact rather than spikes. Sometimes a paper supplies too little primary data to substantiate its recording claim, for example no inter-spike-interval or refractory structure for a single-unit claim. That insufficiency is scored as a fidelity deficiency rather than left neutral, under the burden-on-the-paper rule of Section 3.5.0. This dimension keys on the examined primary data and on the reported measurements with their uncertainty, not on adjectives like careful or rigorous.

The novelty dimension applies the Element 2 retrieval to the observation claim and computes three features. Presence is one minus the maximum entailment confidence over the 50 same-year-or-prior near neighbors, so an observation that any prior neighbor already entails scores near zero novelty regardless of framing. Linkage is the maximum gap-entailment confidence over that same neighbor set, which credits an absent observation addressing a gap or contradiction the prior set actually contains. Distance is the normalized concept-graph path length between the fields the observation connects. Both novelty axes must clear their thresholds. An observation that is absent but unlinked is irrelevant, and an observation that is linked but already present is restatement.

Every relational feature emits the specific prior paper and sentence that produced it, and that emission is a hard gate per Section 3.5.0. No novelty is scored from presumed absence of prior work unless the retrieval ran and its neighbors were checked, and omitting that discipline is what produces false-novelty credit. The presence feature is also down-weighted by the substance score of the nearest prior neighbors. A novelty that exists because the adjacent prior work was too weak to have found the result is therefore discounted.

A dormancy exception modifies the presence feature, so that resurfacing lost prior work is credited rather than penalized. The exception fires when the nearest prior neighbor is old and no longer in the field's working memory. That is measured three ways, by its age, by a dormant citation trajectory with few or no citations in the years before the current submission, and by low retrieval accessibility. The exception also requires that the current paper cites and builds on that neighbor rather than concealing it. The observation is then scored as recovery and extension rather than restatement, because re-establishing a finding the field had lost is a contribution the citation record itself failed to preserve. The discriminating signal separates two cases. A paper that acknowledges the dormant source and builds forward is recovering it. A paper that claims fresh novelty over an actively-cited recent source is running the false-novelty gaming route addressed in Section 6.8.

Implementation.

29. Extract observation claims from the paper using the Element 1 claim-extraction model, tagging each by section of origin and identifying the primary observation from the abstract.
30. Score the fidelity features per observation claim from the results and methods, computing precision, completeness, and artifact-control features.
31. For each observation claim, run the Element 2 retrieval against the OpenAlex same-year-or-prior corpus, using SPECTER2 embeddings with k equal to 50 and a cosine threshold of 0.75. Then run the entailment pass, a DeBERTa-v3-class model with an entailment threshold of 0.85, to compute presence and linkage. Compute the concept-graph path length for distance, and retain the grounding sentences.
32. Map the fidelity and novelty feature vectors to two 1-to-10 sub-scores through the Element 3 ordinal regressions. Combine those sub-scores into the SObs claim-score at the calibrated weighting. Fire the above-10 incomputability detector where the near-neighbor set is sparse or empty and a distant linkage is verified.
33. Aggregate claim-scores to the paper-level SObs by salience weighting, with the primary observation weighted highest. Report the fidelity and novelty sub-scores separately, with confidence intervals and the grounding citations. Normalize against the Element 4 NIH-funded reference stratum for the paper's field.

Failure mode and gaming defense.

An author could pad the results with precise-looking but irrelevant measurements. The defense conditions fidelity on relevance to the stated question, so precision earns credit only for observations bearing on the claim. Completeness is verified separately, by checking whether the methods imply measurements the results omit. A second attack targets the novelty dimension by coining terminology that makes retrieval look sparse. The defense embeds the observation's meaning and judges absence by that meaning and by the cited grounds rather than by surface terms. New words alone therefore do not create apparent novelty.

Observation alone cannot separate a scientist from a fiction writer. An invented off-map observation scores high on novelty and distance as a mathematical consequence. A high SObs is therefore credited as science only when the downstream steps of Sections 3.5.3 through 3.5.11 also hold. High observation score with low downstream signal is the computable signature of overclaiming, and the separately-scored step vector is what exposes it.

Rubric anchors and scale.

On the common rubric, a 3 reports observations without stated precision, uncertainty, or artifact controls, or reports an observation already present in prior work. A 7 reports precise, uncertainty-bounded observations relevant to the claim, with basic artifact control and genuine absence from near-prior work. A 10 reports complete observations with full uncertainty quantification and explicit confound discrimination at the field's best standard, absent from prior work and linked to a real prior gap across a measurable field distance.

A score above 10 is reached two ways. One is measured excellence beyond that range. The other is principled incomputability, meaning the same-year-or-prior near-neighbor set is sparse or empty because the observation opened ground the field had not been standing on. That second route is conditioned on a verified distant linkage, which establishes that the observation

is founding rather than disconnected. Incomputability against the prior corpus is therefore the most informative novelty signal rather than a scoring failure, because founding work is inherently hard to place on a map it created. The above-10 score is a claim about novelty and foundingness only. It says nothing about correctness, which is deferred to the trajectory metrics and eventual replication.

3.5.2 Gap-Identification Index (SGap)

Definition.

The quality and originality of the knowledge gap the paper identifies and targets, scored at the reasoning level within the paper, distinct from the network-level Gap Index (Section 3.7) which tracks gaps across the literature.

Motivation.

Identifying the right gap is the reasoning act that separates consequential questions from incremental ones, and it is invisible to citation metrics. This index scores whether the paper locates a real, explicitly-stated gap and motivates why closing it matters, rather than asserting novelty generically.

Formulation and AI specification.

SGap instantiates the shared protocol of Section 3.5.0. It reads the introduction and discussion for three things. First, an explicit statement of what is unknown. Second, evidence that the gap is real, meaning prior work is cited as leaving it open rather than as having closed it. Third, a stated consequence of closing it. The scorer keys on the logical structure of a gap argument, meaning an unknown, why it matters, and why it is not yet resolved. It does not key on the word novel. This overlaps deliberately with the network-level Gap Index, and Section 3.14 reconciles the two.

Implementation.

34. Extract the gap claim, the cited prior work, and the stated consequence using the Element 1 extraction model.
35. Compute a gap-reality feature by testing, via Element 2 retrieval and NLI, whether the cited prior work is classified as leaving the gap open rather than as having closed it, retaining the grounding sentences.
36. Compute a gap-independence feature by cross-checking the claimed gap against the network-level Gap Index (Section 3.7) for corroboration that others independently treat it as open.
37. Map the features to the 1-to-10 score via the Element 3 ordinal regression and normalize against the Element 4 reference stratum.

Failure mode and gaming defense.

Generic novelty claims are the gaming vector. The defense requires the gap to be tied to specific cited prior work shown to leave it open, so an unsupported claim of novelty scores low, and cross-checks the claimed gap against the network-level Gap Index for independent corroboration.

Rubric anchors and scale.

On the common rubric, a 3 asserts novelty generically with no cited prior work leaving the gap open. A 7 states a specific gap tied to cited work and a stated consequence of closing it. A 10 identifies a narrowly-bounded gap with strong evidence it is real and unaddressed and a compelling stated stake. A score above 10 identifies a gap the field had not recognized as a gap, reframing what is unknown.

3.5.3 Question-Formulation Index (SQ)

Definition.

The falsifiability, scope, and precision of the research question the paper states.

Motivation.

A well-formed question is a precondition for interpretable results, and a vague or unfalsifiable question produces work that cannot be wrong and therefore cannot be informative. The index scores whether the question is specific, answerable, and falsifiable.

Formulation and AI specification.

SQ instantiates the shared protocol of Section 3.5.0 and reads the stated question or aims for three properties. Falsifiability asks whether some outcome would answer the question no. Scope calibration asks whether the question is matched to what the methods can address. Precision asks whether the variables and conditions are specified. The scorer keys on the presence of a falsifiable structure rather than on question phrasing.

Implementation.

38. Extract the stated question or aims using the Element 1 extraction model.
39. Compute a falsifiability feature (is there a stated outcome that would answer the question no), a scope-calibration feature (does the question match what the methods can address, tested by entailment between question and methods), and a precision feature (are variables and conditions specified).
40. Where a preregistration exists, check question-to-registration consistency to detect questions reverse-fit to the results.
41. Map the features to the 1-to-10 score via the Element 3 ordinal regression and normalize against the Element 4 reference stratum.

Failure mode and gaming defense.

Post-hoc question fitting, stating a question the results happen to answer, is the gaming vector. The defense checks consistency between the stated question, the pre-registered aims where available, and the methods, flagging questions that appear reverse-fit to the results.

Rubric anchors and scale.

On the common rubric, a 3 states a vague or unfalsifiable question. A 7 states a specific, falsifiable question matched to the methods. A 10 states a narrowly-scoped falsifiable question whose answer either way is informative. A score above 10 poses a question that opens a line of inquiry rather than closing one.

3.5.4 Hypothesis-Generation Index (SHyp)

Definition.

The clarity of predicted outcomes, the presence of a stated null, and the use of multiple working hypotheses where the domain supports coexisting mechanisms rather than a single strong-inference dichotomy.

Motivation.

Neural interface biology, and much of biomedicine, involves mechanisms that coexist rather than compete. A Chamberlin multiple-working-hypotheses structure therefore often fits better than a Platt strong-inference dichotomy. The index credits explicit, falsifiable predictions. Where appropriate it also credits a structured set of alternatives, rather than a single hypothesis defended against a strawman.

Formulation and AI specification.

SHyp instantiates the shared protocol and scores three features. Novelty applies the SObs Element 2 machinery to the hypothesis claim, testing it against the same-year-or-prior corpus of hypotheses. A hypothesis already stated in prior work therefore scores low novelty regardless of framing, and one that is absent but linked to a real prior gap scores high. An internal-logic feature tests two things, whether the hypothesis follows from the stated gap, and whether each stated alternative includes a distinct predicted observation. Undifferentiated alternatives are not genuine multiple working hypotheses.

The third feature is hypothesis-to-claim alignment, and it is the most sensitive overclaiming detector in this step. It tests whether the hypotheses the paper actually tested are the ones its conclusions are about. A common reasoning failure tests one hypothesis and then concludes about a different claim. An entailment check computes this across the stated hypotheses, the reported tests, and the abstract's conclusions, and it flags a paper whose conclusions are not entailed by the hypotheses it tested. The scorer keys on the logical relations among gap, hypothesis, test, and conclusion, not on the word hypothesis.

Implementation.

42. Extract the hypothesis claims, the stated gap, the reported tests, and the conclusion claims using the Element 1 extraction model.
43. Compute the novelty feature via SObs Element 2 retrieval and entailment against the prior-hypothesis corpus, retaining grounding sentences.
44. Compute the internal-logic feature by testing gap-to-hypothesis entailment and checking that alternatives include distinct predicted observations.
45. Compute the alignment feature by testing whether the conclusion claims are entailed by the tested hypotheses, flagging tested-H-but-concluded-C mismatches.
46. Map the three features to the 1-to-10 score via the Element 3 ordinal regression and normalize against the Element 4 reference stratum.

Failure mode and gaming defense.

Two gaming vectors apply here. Listing token alternative hypotheses that are never distinguished is the first, and the defense requires each stated hypothesis to state a distinct

predicted observation. Stating hypotheses that match the eventual claims while testing something narrower is the subtler one. The alignment feature defends against it by scoring the entailment from tested hypotheses to stated conclusions directly, so a paper cannot raise SHyp by claiming a broad hypothesis while testing a narrow one.

Rubric anchors and scale.

On the common rubric, a 3 states no explicit hypothesis or an unfalsifiable one or draws conclusions unrelated to the hypotheses tested. A 7 states an explicit falsifiable hypothesis with a null whose conclusions align with what was tested. A 10 states a structured set of distinguishable working hypotheses each with a distinct predicted observation and full hypothesis-to-conclusion alignment. A score above 10 formulates a hypothesis, novel against the prior corpus, that reconciles previously conflicting observations.

3.5.5 Prediction-Specificity Index (SPred)

Definition.

The specificity and testability of the concrete predicted observations the paper derives from its hypotheses, treated as a distinct step because a hypothesis without a stated prediction is not falsifiable.

Motivation.

Prediction is the step that converts a hypothesis into something an experiment can confront, and its absence is what lets a hypothesis be defended against any outcome after the fact. The index scores whether the paper states, before the results, what specific observations each hypothesis predicts and what observations would falsify it.

Formulation and AI specification.

SPred instantiates the shared protocol and scores two features. Specificity reads for explicit predicted observations tied to each hypothesis. Those predictions should be quantified where the domain allows, meaning a predicted direction, magnitude, or threshold rather than a vague expectation, and stated before the results rather than inferred from them. The feature scores whether the prediction names a measurable outcome together with a value or a direction.

The second feature is hypothesis linkage, paired with the SHyp alignment feature. It tests whether each prediction follows from a stated hypothesis by logic, computed as an entailment check from hypothesis to prediction. A prediction unconnected to the hypothesis it claims to derive from is not a test of that hypothesis. The scorer keys on the presence of a specific, falsifiable predicted observation logically linked to a hypothesis, not on the word predict.

Implementation.

47. Extract prediction claims and their asserted parent hypotheses using the Element 1 extraction model.
48. Compute the specificity feature by testing whether each prediction names a measurable outcome with a direction, magnitude, or threshold.
49. Compute the hypothesis-linkage feature by the hypothesis-to-prediction entailment, flagging predictions not logically derived from a stated hypothesis.

50. Where a preregistration exists, check prediction-to-registration consistency to detect predictions stated after the results (HARKing).
51. Map the features to the 1-to-10 score via the Element 3 ordinal regression and normalize against the Element 4 reference stratum.

Failure mode and gaming defense.

Stating predictions after seeing the results (HARKing, hypothesizing after results are known) is the gaming vector. The defense compares predictions against any pre-registration. It flags a prediction that matches the results with suspicious specificity when no registered protocol exists. A hypothesis-linkage feature blocks the related move, attaching a well-formed prediction to a hypothesis it does not test.

Rubric anchors and scale.

On the common rubric, a 3 states no specific predicted observation or a prediction not derived from any stated hypothesis. A 7 states a falsifiable predicted observation logically tied to the hypothesis. A 10 states quantified predictions (direction, magnitude, or threshold) distinguishing among alternatives before the results. A score above 10 makes a risky prediction that the field expected to fail and that the data confirm.

3.5.6 Experimental-Design Index (SDes)

Definition.

The adequacy of controls, sample-size justification, blinding, and randomization against the subfield standard.

Motivation.

Design quality determines whether results can support their claims, and it is the step most often asserted rather than demonstrated. The index scores the design against what the question requires.

Formulation and AI specification.

SDes instantiates the shared protocol of Section 3.5.0 and rises only when the scorer locates an actual control condition described in methods and referenced in analysis, an actual power calculation or stated effect-size assumption behind the sample size, and actual randomization or blinding procedures, not the phrases appropriate controls or adequately powered. The comparative standard against subfield norm is a learned per-subfield baseline of each design feature.

Implementation.

52. Extract the described design elements (controls, sample-size basis, blinding, randomization) using the Element 1 extraction model.
53. Score each design feature by presence as structure, so a control counts only when described as a condition and referenced in the analysis, and a power calculation counts only when the computation is present, against the per-subfield baseline of Element 4.
54. Compute the discrimination feature by testing whether the design can distinguish the stated hypotheses (entailment between design and the SHyp hypothesis set).

55. Map the features to the 1-to-10 score via the Element 3 ordinal regression and normalize against the reference stratum.

Failure mode and gaming defense.

Claiming controls without describing them is the gaming vector. The defense requires the claimed control to appear as a described condition that is actually analyzed, and a claimed power calculation to contain the computation, so claim language without structure does not score.

Rubric anchors and scale.

On the common rubric, a 3 lacks described controls or sample-size justification. A 7 has adequate controls, a stated power basis, and appropriate blinding or randomization for the subfield. A 10 has a design that discriminates cleanly among the stated hypotheses at the field's best standard. A score above 10 introduces a design or control that becomes a methodological template others adopt.

3.5.7 Execution-Fidelity Index (SExec)

Definition.

The reproducibility and technique fidelity of the experimental execution as evidenced in the methods, the step that is distributable to trainees or technicians while the work remains science, and the one least directly visible in a written paper.

Motivation.

Execution is step six of the scientific method and the only step that can be delegated without ceasing to do science, but a written paper shows execution only through the traces it leaves in the methods, so the index scores those traces rather than the bench work itself. Poor execution invisible in a paper is a reproducibility failure, and the traces that predict it are underspecified protocols and missing technique detail.

Formulation and AI specification.

SExec instantiates the shared protocol of Section 3.5.0 and scores two features. The essential one is reproducibility-gap detection. For each method described, a checklist model extracts the procedural parameters an independent laboratory would need, and compares them against a reproducibility checklist for that class of method. For a reagent, those parameters are its source, catalog identifier, and concentration. For a protocol step, they are duration, temperature, and equipment setting. For a computational method, they are version, hyperparameters, and random seed.

The reproducibility-gap feature is the fraction of required parameters present. A method that names a reagent without its concentration, or an antibody without its catalog number, therefore scores the gap explicitly. This targets the specific omissions that cause cross-laboratory non-reproducibility, where a reader is left guessing at a value the original laboratory knew but did not report. It does not penalize brevity by itself. The feature applies the same silence-counts-as-deficiency default as the analysis-rigor index of Section 3.5.8, scoring an absent required parameter as not done rather than as unknown, because assuming an unstated step was

performed correctly rewards incomplete reporting. Its required-parameter set is field-calibrated, so a field's genuine reporting convention is not mistaken for a lapse.

The second feature is methods-results consistency, an entailment check. It tests whether the reported data are of the kind and scale the described procedure could actually produce, and it flags results the stated methods could not have generated. The checklist for each method class is the point of calibration. It is populated from existing reporting standards, for example the ARRIVE guidelines for animal work and MIQE for quantitative PCR, together with equivalent community checklists, rather than invented here.

Implementation.

56. Extract each described method and classify it by method class using the Element 1 extraction model, mapping each to its reproducibility checklist.
57. For each method, extract the checklist parameters actually present in the text and compute the reproducibility-gap feature as the present-to-required ratio, retaining the list of missing parameters as the human-checkable output.
58. Run the methods-results consistency entailment (Element 2 NLI machinery) to flag results inconsistent with the described procedure.
59. Map the two features to the 1-to-10 score via the Element 3 ordinal regression, and normalize against the Element 4 reference stratum for the method's field, since reproducibility norms differ across wet-lab, computational, and clinical work.

Failure mode and gaming defense.

Padding the methods with boilerplate detail that does not enable reproduction is the gaming vector. The defense scores presence of the specific checklist parameters that matter for reproduction rather than methods-section length, so verbose but parameter-incomplete methods score low, and the methods-results consistency check catches procedures described but not actually followed. Because the missing-parameter list is emitted, a reviewer can verify each claimed gap against the manuscript rather than trusting the score.

Rubric anchors and scale.

On the common rubric, a 3 provides a protocol too underspecified to reproduce. A 7 provides a protocol reproducible by a skilled peer with the error-prone steps described. A 10 provides a fully reproducible protocol with deposited step-level detail meeting the subfield's best reproducibility standard. A score above 10 establishes a technique or protocol refinement that raises the field's reproducibility baseline.

3.5.8 Analysis-Rigor Index (SAna)

Definition.

Statistical validity is checked as logical consistency and disclosure. Four questions apply. Does the test match the data type? Does the analysis account for the dependence structure of the data? Do the conclusions match what the test measured? Are the reporting elements the field requires present? A required element that is not stated is scored as not done.

Motivation.

Analytic rigor is where many false positives enter, and current metrics reward the positive result rather than the correctness of the analysis that produced it. Any significance claim requires its p value and the correction appropriate to the test count.

Formulation and AI specification.

SAna instantiates the shared protocol of Section 3.5.0 and runs a set of logical-consistency and disclosure checks. Their ground truth is the fixed rules of statistics and the paper's own internal logic, not a human-labeled quality corpus, so the metric detects patterns computationally rather than reproducing expert taste. Four checks determine the score.

A test-to-data-type check reads two things from the methods, the data structure and the named test. The data structure may be continuous, count, categorical, ordinal, time-series, or nested. A statistical rule then verifies that the test matches the data type, and a test applied to a data type it does not fit is flagged as a mathematically-defined error rather than an opinion. A dependence-structure check catches pseudoreplication. It compares the described sampling, for example repeated measures on the same units or multiple cells from the same animals, against how the analysis treated the sample size, and it flags an analysis that treats non-independent observations as independent. A claim-alignment check, run by entailment, tests whether the conclusions are about the quantity the test actually measured, since a common error runs a test on one quantity and states results about another. A disclosure check audits whether the field-required reporting elements are present in the text. Those elements are assumption checks such as normality and variance homogeneity, multiple-comparison correction when the test count requires it, effect sizes, and exact p values.

The disclosure check applies one strict default. A required element that is not stated is scored as not done rather than as unknown. Assuming an unstated safeguard was performed rewards incomplete reporting, and the alternative would force the model to guess at hidden practice it cannot observe. Silence therefore scores as the deficiency. The required-element set is field-calibrated to what that field's rigorous papers actually disclose, so a field's genuine reporting convention is not mistaken for a lapse. Three essential validity elements are never waived for silence, because the validity of the result turns on them, and they are dependence-structure accounting, multiple-comparison correction, and the test-to-data match. Correctly stated caveats about a test's assumptions and their limits raise the score, since a paper that names its assumptions and whether they were met demonstrates the rigor this metric rewards.

Implementation.

60. Extract the data structure, sampling design, named tests, corrections, effect sizes, and p values from the methods and results using the Element 1 extraction model.
61. Run the test-to-data-type check by statistical rule, flagging any test not valid for its data type.
62. Run the dependence-structure check, comparing described sampling against the treated sample size to detect pseudoreplication and unaccounted repeated measures.
63. Run the claim-alignment check by entailment, flagging conclusions asserted about a quantity the reported test did not measure.

64. Run the disclosure audit against the field-calibrated required-element set, scoring each absent required element as a deficiency rather than as unknown, with the essential validity elements never waived for silence.
65. Map the check results to the 1-to-10 score via the Element 3 ordinal regression and normalize against the reference stratum, since reporting norms differ across fields.

Failure mode and gaming defense.

Reporting a correction in the methods but never applying it in the results is one gaming vector. The defense verifies application against the results tables rather than the methods statement, and flags a mismatch between the claimed and the applied correction. Omitting disclosure to dodge the audit is the second vector, and the silence-counts-as-deficiency default blocks it, since saying less scores as the error rather than as unknown. Full disclosure is therefore the dominant strategy.

One honest boundary remains. The checks catch errors visible in the text, meaning a wrong test for the data type, a disclosed nesting treated as independent, and conclusions that do not match the test. They cannot catch an undisclosed confound, or a nesting the methods never mention. The statistical checks are therefore strong where methods are complete and they fail where methods omit the structure. That failure is itself scored, by the disclosure audit here and by the reproducibility-gap feature of the execution index.

Rubric anchors and scale.

On the common rubric, a 3 uses a test mismatched to the data type, treats non-independent observations as independent, or omits required disclosures. A 7 uses appropriate tests with dependence correctly handled, conclusions aligned to what was tested, and the field-required elements disclosed. A 10 additionally reports assumption checks with their outcomes and full sensitivity treatment at the field's best standard. A score above 10 introduces or correctly applies an analytic approach ahead of the field's norm. At every level, unstated required elements are scored as not done.

3.5.9 Interpretation-Calibration Index (SInt)

Definition.

The proportionality of the paper's conclusions to its evidence and the presence of stated alternative interpretations.

Motivation.

Overclaiming beyond the evidence is a common and costly failure that citation counts do not penalize and can even reward. The index scores whether conclusions are calibrated to what the data support and whether alternatives are acknowledged.

Formulation and AI specification.

SInt instantiates the shared protocol of Section 3.5.0. It compares the strength of the stated conclusions against the strength of the evidence the results provide, and it reads the discussion for explicitly stated alternative explanations. The scorer keys on the calibration between claim strength and evidence strength. Honest disclosure of limitations is credited as part of that calibration. A paper that names the specific constraints on its evidence, and qualifies its

essential claims to match, therefore scores higher than one asserting the same claims without the qualification.

The credit attaches to substance rather than to a heading. The mere presence of a limitations section is a held-out surface feature. Only a disclosed limitation that corresponds to a real constraint, and that actually narrows the affected claim, earns the credit, which keeps a boilerplate limitations paragraph from collecting the reward. Calibration is scored for overclaiming and underclaiming alike, so a claim hedged below what its evidence supports is also a miscalibration, though a less harmful one than overclaiming. That keeps the credit from rewarding reflexive hedging rather than accurate self-assessment. The insufficiency debit of the fidelity gate and this calibration credit are the same principle from two sides, the burden of evidence resting on the paper. A categorical claim on thin data is debited by the gate and uncredited here. A modest claim on the same data with honest limitations is spared the debit and earns the credit.

Implementation.

66. Extract the conclusion claims and the evidence each rests on using the Element 1 extraction model.
67. Compute a calibration feature as the entailment gap between conclusion strength and the strength the results actually support, flagging conclusions stronger than their evidence.
68. Compute an alternatives feature crediting explicitly stated alternative explanations tied to distinguishing evidence, not boilerplate hedging.
69. Map the features to the 1-to-10 score via the Element 3 ordinal regression and normalize against the reference stratum.

Failure mode and gaming defense.

Ritual limitations paragraphs that acknowledge alternatives without engaging them are the gaming vector. The defense requires stated alternatives to be tied to specific evidence that would distinguish them, so boilerplate hedging does not score.

Rubric anchors and scale.

On the common rubric, a 3 draws conclusions beyond what the evidence supports. A 7 draws conclusions proportional to the evidence and states alternatives. A 10 calibrates every claim to its evidence and engages alternatives with distinguishing evidence. A score above 10 correctly identifies the broader significance of a result that the narrow data alone would not imply.

3.5.10 Synthesis-and-Integration Index (SSyn)

Definition.

The density of genuine cross-scale and cross-field connection in the paper's reasoning, with verified methodological dependency, scored at the reasoning level and overlapping deliberately with the Cross-Field Integration Index (Section 3.9) and Mechanism-Graph Centrality Index (Section 3.2).

Motivation.

Cross-domain synthesis density is the signature of high scientific reasoning, integrating materials, glial and vascular biology, electrophysiology, neurocomputation, and clinical translation within a single framework, and it is the operation the flat prior version of this index underweighted. This step-index restores it to full status. It scores how many distant domains the paper connects and how densely, not how many it name-checks.

Formulation and AI specification.

SSyn instantiates the shared protocol of Section 3.5.0 and measures the number and distance of domains connected in the paper's own argument and requires methodological dependency (a connected domain's method or concept is actually used in the analysis), reusing the CFII field-distance machinery and the MGCI mechanism graph. The scorer keys on structural dependency between domains, not on citations spanning fields.

Implementation.

70. Extract the domains the paper connects and the methods or concepts it borrows using the Element 1 extraction model.
71. Apply the SObs Element 2 field-distance machinery to measure the distance of each connected domain, and require methodological dependency by a semantic match above 0.6 between the borrowed method and the paper analysis, so listed-but-unused domains do not count.
72. Compute connection density as the count of distance-weighted genuine dependencies, retaining the grounding sentences.
73. Map the feature to the 1-to-10 score via the Element 3 ordinal regression and normalize against the reference stratum, applying the incomputability above-10 rule where a connection has no prior precedent.

Failure mode and gaming defense.

Tokenistic cross-field citation is the gaming vector. The defense requires each connected domain to appear in the paper's analysis with methodological dependency above a 0.6 semantic-match threshold, so listed-but-unused domains do not raise the score. Overlap with CFII and MGCI is reconciled in Section 3.14.

Rubric anchors and scale.

On the common rubric, a 3 connects no domains beyond its own or name-checks them without use. A 7 integrates one adjacent domain's method or concept into the analysis. A 10 integrates multiple distant domains with verified methodological dependency at high connection density. A score above 10 forges a cross-domain connection that opens a new integrative direction, the signature of the highest synthetic reasoning.

3.5.11 Reporting-Transparency Index (SRep)

Definition.

The explicitness and traceability of the causal chain from evidence to conclusion, including data, code, and method availability.

Motivation.

Transparency is what makes a result reusable and checkable, and it is rewarded by no current metric. The index scores whether a reader can trace each conclusion to the evidence and reproduce the path.

Formulation and AI specification.

SRep instantiates the shared protocol of Section 3.5.0 and reads for an explicit traceable chain from evidence to each conclusion, and for data, code, and protocol availability statements that resolve to actual deposited resources. The scorer keys on the presence of a followable chain and on availability statements that verify against real repositories.

Implementation.

74. Extract the evidence-to-conclusion chain and the data, code, and protocol availability statements using the Element 1 extraction model.
75. Compute a traceability feature testing whether each conclusion links to specific reported evidence, and a resource-resolution feature testing whether availability statements resolve to actual accessible deposits (a live check against the named repository).
76. Flag availability statements that do not resolve to a real deposit.
77. Map the features to the 1-to-10 score via the Element 3 ordinal regression and normalize against the reference stratum.

Failure mode and gaming defense.

Data-availability statements that point to nothing are the gaming vector. The defense verifies that availability statements resolve to actual accessible resources, so a statement without a real deposit does not score.

Rubric anchors and scale.

On the common rubric, a 3 leaves conclusions untraceable to evidence and provides no working data or code availability. A 7 provides a traceable evidence-to-conclusion chain and availability statements that resolve to real resources. A 10 provides fully transparent reasoning with complete deposited data, code, and protocols at the field's best standard. A score above 10 sets a transparency or open-resource practice that others adopt.

3.5E Engineering-Method Index (EMI), a meta-index over engineering-method steps

Definition.

A meta-index parallel to SRQI that composes six named engineering-method step-indices into a profile of the engineering reasoning a paper demonstrates. EMI applies to design-and-build work, where the governing process is not observe-hypothesize-test but define, design, prototype, characterize, validate, and iterate, and it is reported as the vector of its six step-scores.

Motivation.

A large fraction of NIH-funded work, including most of neural engineering, is engineering rather than discovery. Scoring an engineering paper on scientific-method steps misjudges it. A device

paper has no hypothesis in the Popperian sense, but it does have a requirement specification, a design-space exploration, and a validation against that specification. SRQI and EMI are separate meta-indices because the scientific and engineering methods are distinct processes with distinct step structures, and conflating them reintroduces the field-mismatch error this framework warns against. Many neural-engineering papers are scored on both, because they do both, characterizing a biological mechanism, which is science, while building and validating a device, which is engineering.

Formulation.

$$EMI(p) = (EReq, EDes, EProto, EChar, EVal, EIter)$$

reported as the vector of the six step-indices defined in Sections 3.5E.1 through 3.5E.6, each on the same common anchored 1-to-10 rubric (uncapped above 10) as SRQI, with a confidence interval, no canonical weighting, and committee-selected role-appropriate weights, so the EMI composite is a weighted average commensurable across its steps.

AI specification and mode selection.

A mode classifier first determines whether a paper is discovery-mode, engineering-mode, analytical-mode, or a combination, reading its aims and methods. Three structural markers drive that choice. The first is generation of new empirical data, whether by experiment or by systematic observation. The second is the presence of a requirement and a validation. The third is the absence of new data in favor of an argument built by integrating prior literature.

Discovery-mode covers two kinds of work. Experimentally-warranted work tests a hypothesis by manipulation and inferential statistics. Observationally-warranted work establishes claims by systematic observation and inference to the best explanation. These are coequal modes of empirical warrant rather than a norm and a lesser exception, so the rigor steps are polymorphic by warrant. SExec, SAna, and SHyp apply experimental-rigor criteria to experimentally-warranted claims, meaning statistical validity, control conditions, and hypothesis-test structure. They apply observational-rigor criteria to observationally-warranted claims, meaning sampling adequacy, observer and instrument bias controls, marker or measurement specificity, reproducibility of the observation, and whether the inference actually ruled out the alternatives.

An observational paper is therefore scored by the standard its own warrant sets, and it is not marked deficient for lacking inferential statistics. Brock 1969, one of this framework's positive controls, is observational natural history with no controlled experiment and no statistics. A discovery mode presuming the hypothesis-test template would mishandle the framework's own founding case. Analytical-mode covers theory, review, and perspective papers, whose contribution is identifying a gap, a contradiction, or a wrong shared assumption by linking literature rather than generating new data. Section 3.5A specifies it, and it is distinct from observational discovery because it generates no new data at all. Each paper is scored on SRQI, EMI, the analytical-mode step subset, or a combination as appropriate.

Each EMI step-index instantiates the shared protocol of Section 3.5.0 and scores two axes. An internal-rigor axis specific to the engineering step asks three questions. Was the requirement quantified? Was the design space actually explored? Did validation run against every requirement? A corpus-novelty axis then applies the SObs Element 2 machinery to the prior engineering literature, so a requirement, design, or validation approach is credited as more

innovative when it is absent from and distant to prior engineering work. The incomputability above-10 rule applies where an approach has no prior precedent.

Each EMI step reads structural engineering evidence rather than claim language. Its score is anchored to how the same computed features are distributed across the reference population, with no expert quality-labels, matching the SRQI protocol. The same held-out surface features apply, so naming an operation does not substitute for performing it. Papers scored on multiple modes report each vector separately, so a bidirectional-translation paper is credited for its science and its engineering separately rather than averaged into one misfitting number.

Failure mode and mitigation (meta-index level).

The dominant risk is mode-misclassification, scoring an engineering paper as science, an analytical paper as discovery, or any other mismatch, mitigated by multiple scoring where the mode classifier is uncertain and by reporting the mode decision with its confidence. A paper mis-routed to discovery-mode when it is analytical would be scored on execution and analysis steps it has no basis to satisfy, which the analytical-mode branch of Section 3.5A prevents by not instantiating those steps. Gaming is addressed at the step level.

RFI category alignment.

Entrepreneurship and Translation and Foundational Scientific Exploration where the engineering is foundational infrastructure, cross-cutting content calibrator for design-and-build work.

3.5E.1 Requirement-Definition Index (EReq)

Definition.

The clarity, completeness, and justification of the specification the engineering work targets, the engineering analogue of question formulation.

Motivation and specification.

A design is only as good as the requirements it is judged against, and unstated or post-hoc requirements make a device impossible to evaluate. EReq reads for explicit, quantified performance requirements stated before the design (target values with units and tolerances), justification tying each requirement to the application need, and constraints (power, size, biocompatibility, cost). The scorer keys on the presence of quantified, justified specifications, not on the word requirement.

Implementation.

78. Extract the stated performance requirements, constraints, and their justifications using the Element 1 extraction model.
79. Score the rigor axis by whether requirements are quantified (target values with units and tolerances) and justified against the application need, checking for reverse-fit by comparing requirements to any preregistration or prior protocol.
80. Score the novelty axis by applying SObs Element 2 retrieval to the requirement against the prior engineering corpus.

81. Map both axes to the 1-to-10 score via the Element 3 ordinal regression and normalize against the Element 4 engineering reference stratum.

Failure mode and gaming defense.

Reverse-fitting requirements to the achieved performance is the gaming vector. The defense checks consistency between stated requirements and any pre-registration or prior protocol, and flags requirements that match achieved results as suspect.

Rubric anchors and scale.

On the common rubric, a 3 states no quantified requirements or reverse-fits them to results. A 7 states quantified, justified performance requirements with constraints before the design. A 10 states a complete, well-justified specification tied rigorously to the application need. A score above 10 defines a requirement set that reframes what the device class should achieve.

3.5E.2 Design-Space-Exploration Index (EDes)

Definition.

The breadth and rigor of the design alternatives considered and the justification for the chosen design, the engineering analogue of multiple working hypotheses.

Motivation and specification.

A design chosen without considering alternatives is an assertion, not an engineering decision. EDes reads for stated design alternatives, explicit trade-off analysis among them, and a justified selection tied to the requirements. The scorer keys on the presence of a real trade-off analysis, not on a single design presented as inevitable.

Implementation.

82. Extract the stated design alternatives and the trade-off analysis using the Element 1 extraction model.
83. Score the rigor axis by whether real alternatives are considered with genuine trade-offs on stated axes and a justified selection, rejecting strawman alternatives that present no real trade-off.
84. Score the novelty axis by SObs Element 2 retrieval of the chosen design approach against the prior engineering corpus.
85. Map both axes to the 1-to-10 score and normalize against the engineering reference stratum.

Failure mode and gaming defense.

Strawman alternatives introduced only to be dismissed are the gaming vector. The defense requires each alternative to present a real trade-off against the chosen design on a stated axis, so token alternatives do not score.

Rubric anchors and scale.

On the common rubric, a 3 presents one design with no alternatives considered. A 7 considers real alternatives with a trade-off analysis and justified selection. A 10 explores the design space

rigorously with quantified trade-offs across the relevant axes. A score above 10 identifies a design approach the field had not considered.

3.5E.3 Prototype-Realization Index (EProto)

Definition.

The completeness and reproducibility of the build, the engineering analogue of experimental execution.

Motivation and specification.

A described design that was not actually built and characterized is a proposal, not a result. EProto reads for evidence the artifact was realized (fabrication methods, materials, and process parameters sufficient to reproduce it) and for the fidelity of the realized artifact to the design. The scorer keys on the presence of reproducible build detail.

Implementation.

86. Extract the fabrication methods, materials, and process parameters using the Element 1 extraction model.
87. Score the rigor axis by whether the build is specified in enough detail for independent reproduction (the engineering analogue of the SExec reproducibility-gap feature) and whether realized-to-design fidelity is demonstrated.
88. Score the novelty axis by retrieval of the fabrication approach against the prior engineering corpus.
89. Map both axes to the 1-to-10 score and normalize against the engineering reference stratum.

Failure mode and gaming defense.

Claiming a build without reproducible detail is the gaming vector. The defense requires fabrication and process parameters sufficient for independent reproduction, so an underspecified build does not score.

Rubric anchors and scale.

On the common rubric, a 3 describes a build too underspecified to reproduce. A 7 provides fabrication and process detail sufficient for independent reproduction. A 10 provides complete reproducible realization with verified design fidelity at the field's best standard. A score above 10 achieves a fabrication result the field considered infeasible.

3.5E.4 Characterization-and-Measurement Index (EChar)

Definition.

The rigor and completeness of the measurements characterizing the realized artifact's performance, the engineering analogue of analysis rigor.

Motivation and specification.

Characterization is where a device's real performance is established, and selective or under-powered characterization overstates it. EChar checks that performance is measured across the

operating range, with appropriate sample sizes across fabricated units, statistical treatment of unit-to-unit variability (any significance claim requires its p value), and reporting of failure cases and off-nominal behavior. The scorer keys on the completeness and statistical treatment of the measurements.

Implementation.

90. Extract the performance measurements and their sampling using the Element 1 extraction model.
91. Score the rigor axis by coverage across the operating range, adequate unit-to-unit sampling with variability statistics (any significance claim requires its p value), and failure-case reporting, flagging best-unit-only reporting.
92. Score the novelty axis by retrieval of the characterization method against the prior engineering corpus.
93. Map both axes to the 1-to-10 score and normalize against the engineering reference stratum.

Failure mode and gaming defense.

Best-unit reporting, characterizing only the device that worked, is the gaming vector. The defense requires reporting across fabricated units with unit-to-unit variability, so single-best-unit results are flagged.

Rubric anchors and scale.

On the common rubric, a 3 reports only best-unit or partial-range performance. A 7 characterizes across the operating range with adequate unit-to-unit sampling and failure reporting. A 10 characterizes comprehensively with full statistical treatment of variability at the field's best standard. A score above 10 establishes a characterization method others adopt.

3.5E.5 Validation-Against-Specification Index (EVal)

Definition.

Whether the characterized performance is tested against the stated requirements, the engineering analogue of interpretation calibration and the step with no scientific-method equivalent.

Motivation and specification.

Validation closes the engineering loop by testing the built artifact against the requirements defined at the start, and its absence lets a paper claim success without a standard. EVal reads for an explicit comparison of characterized performance against each stated requirement, an honest account of which requirements were met and which were not, and validation under application-relevant conditions rather than only benchtop-ideal ones. The scorer keys on the presence of a requirement-by-requirement comparison.

Implementation.

94. Extract the stated requirements from EReq and the validation results using the Element 1 extraction model.

95. Score the rigor axis by requirement-by-requirement comparison of characterized performance against every stated requirement under application-relevant conditions, flagging silently dropped unmet requirements as a mismatch between the requirement set and the validation set.
96. Score the novelty axis by retrieval of the validation approach against the prior engineering corpus.
97. Map both axes to the 1-to-10 score and normalize against the engineering reference stratum.

Failure mode and gaming defense.

Validating only against the requirements that were met is the gaming vector. The defense requires validation against every stated requirement from EReq, so silently dropping an unmet requirement is flagged as a mismatch between the requirement set and the validation set.

Rubric anchors and scale.

On the common rubric, a 3 validates against no stated requirement or only the met ones. A 7 compares performance against every stated requirement under realistic conditions with honest met-and-unmet accounting. A 10 validates comprehensively under application-relevant conditions at the field's best standard. A score above 10 validates against a more demanding standard than the field currently requires.

3.5E.6 Iteration-and-Optimization Index (EIter)

Definition.

Evidence of design iteration informed by characterization and validation, the engineering analogue of the self-correcting cycle of science.

Motivation and specification.

Engineering advances through iteration, and a paper that reports a single design pass without learning from measured shortfalls demonstrates weaker engineering reasoning than one that closes the loop. EIter reads for evidence that characterization or validation results drove a subsequent design change, with the change justified by the measured shortfall. The scorer keys on the traceable link from a measured result to a design revision.

Implementation.

98. Extract claimed design iterations and their driving measurements using the Element 1 extraction model.
99. Score the rigor axis by whether each claimed iteration links a specific measured shortfall to a specific design change with before-and-after data, rejecting unsupported iteration narratives.
100. Score the novelty axis by retrieval of the iteration or optimization approach against the prior engineering corpus.
101. Map both axes to the 1-to-10 score and normalize against the engineering reference stratum.

Failure mode and gaming defense.

Narrating iteration that did not occur is the gaming vector. The defense requires each claimed iteration to link a specific measured shortfall to a specific design change with before-and-after data, so an unsupported iteration narrative does not score.

Rubric anchors and scale.

On the common rubric, a 3 reports a single design pass with no learning from shortfalls. A 7 links a measured shortfall to a justified design change with before-and-after data. A 10 demonstrates a rigorous closed iteration loop driving measured performance gains. A score above 10 achieves through iteration a performance level the field considered a hard limit.

3.5A Analytical-Mode Assessment for Theory, Review, and Perspective Papers

Why a third mode.

Theory, review, and perspective papers make no primary observations, run no experiments, and build no devices. The execution and statistical-rigor steps of SRQI, and the realization and validation steps of EMI, therefore have nothing to read on them. Their contribution is identifying a gap, a contradiction, or a wrong shared assumption, by linking literature that had not been linked. That contribution is among the hardest to count and among the most consequential when it redirects a field. Routing such a paper to discovery-mode would score it on experimental steps it has no basis to satisfy. The silence-counts-as-deficiency default would then convert the structural absence of an experiment into a cascade of false deficiencies, penalizing the paper for not being an experiment. Analytical-mode exists so that a synthesis contribution is scored on what it does rather than on what it cannot do.

What is scored, and what is inert.

Analytical-mode scores only the reasoning and integration steps of the scientific method. Those are the observation step in its corpus-relational sense (SObs), gap identification (SGap), hypothesis or thesis generation (SHyp), interpretation (SInt), and synthesis (SSyn). It also scores the standalone indicators that already read integration and origination rather than experiment, meaning the assumption-addressing indicator (AAI), mechanism-graph centrality (MGCI), cross-field integration (CFII), the field-define index (FDI), and the corpus-relational novelty machinery. Three components are not instantiated at all in analytical-mode, rather than scored as deficient. Those are the execution-fidelity step (SExec), the analysis-rigor step (SAna), and the primary-data-ratio component of the paper-substance score (PSS). A paper with no experimental execution has no execution to check, no experiment-derived statistics to validate, and no primary data whose absence is a fault. Making these steps inert rather than gating them to a lenient setting is the essential decision. A gated but active experimental checklist would still record every experimental element as absent, and the silence-counts-as-deficiency default would still fire.

The deficiency floor, stated with its uncertainty.

A mode that rewarded only positive signals would credit the appearance of synthesis. It would have no way to score down a fluent repackaging that links nothing new and argues nothing sound, which would make analytical-mode a gaming surface and an unusable signal for an evaluator. Analytical-mode therefore has its own deficiency floor, the analogue of the experimental required-element checklist, built from three corpus-relational checks.

The linkage-novelty check tests whether the connections the paper claims to draw are actually absent from the prior literature, using the same retrieval machinery as SObs novelty detection. A synthesis that restates known connections therefore scores low on the contribution it claims. The dormancy exception applies here too. A synthesis that resurfaces connections lost in old, uncited, hard-to-retrieve work and builds on them is credited as recovery rather than penalized as restatement, which follows the framework's founding concern for neglected work, since reviving a forgotten linkage is the synthesis analogue of producing a neglected finding. A source-characterization accuracy check tests whether the cited sources are characterized consistently with what they actually report. It applies the entailment machinery of the SAna claim-alignment check to the citing sentences rather than to a statistical claim, so a paper that mischaracterizes what it cites is flagged. An assumption-exposure check tests whether the paper states the assumptions its argument depends on rather than smuggling them in. That is detectable as the presence or absence of explicit conditional and scope statements in the argument structure.

These three checks sit in the corpus-relational tier of Section 2 rather than the rule-based tier, and this is stated rather than hidden. Whether a linkage is new, and whether a source is faithfully characterized, are questions about the structure and content of the existing literature, which retrieval and entailment can address. They rest on softer ground than a statistical-validity check whose ground truth is mathematics. Analytical-mode substance is therefore assessed with more uncertainty than experimental rigor. The checks return a finding with an explicit confidence rather than a determinate pass or fail, and that confidence is reported alongside the score. A program officer can then weight the analytical assessment for what it is, a corpus-relational judgment of a contribution that resists exact measurement rather than a hard adjudication of whether an argument is correct. The framework does not computationally decide whether an idea is good. It detects whether a synthesis introduces a new and faithfully argued linkage, and it reports the confidence of that detection.

Relation to the framework's general treatment of hard-to-count contribution.

Analytical-mode is one instance of a general principle rather than a special case. A synthesis paper and a patent share a property. Their value lies not in a countable artifact but in what they introduce and what the field or the market subsequently takes up. Both are therefore assessed by the novelty and the uptake of what they introduce rather than by artifact count. That is the same move the field-define index makes for originated concepts, and the same move the incentive arm makes when it credits constitutive behaviors rather than proxies.

Where a contribution is hard to count, the framework's answer stays consistent. Measure the novelty of what was introduced against the prior corpus and the independent uptake that followed, and report the measurement with its uncertainty, rather than defaulting to the artifact tally that undercounts the work this framework credits.

RFI category alignment.

Foundational Scientific Exploration, since a field-redirecting synthesis is a foundational contribution, with a secondary tie to any category where review and perspective work shapes subsequent research direction.

3.6 Verifiable Public-Engagement Index (VPEI)

Definition.

A score quantifying an author's substantive public-communication output through verified audience reach, content-quality scoring, and downstream policy or educational citation.

Motivation.

Current metrics ignore public-engagement contributions entirely, and many institutional cultures actively penalize them. Carl Sagan published approximately 500 peer-reviewed papers across 39 years while also producing the most influential public-science communication of the late 20th century. The National Academy of Sciences blocked his 1992 nomination explicitly because of the public work. Later bibliometric analyses have shown that scientifically active science communicators outperform non-communicators on traditional research metrics, so the Sagan Effect, the perception that scientific popularity is inversely related to scientific quality, is empirically false. This indicator treats substantive public communication as an output class, with verified reach, content-quality scoring, and downstream policy or educational impact.

This indicator applies beyond a single impact category. A foundational-science funder in a democracy has a structural vulnerability. Its payoff horizon is longer than its funding horizon, so the work is perpetually exposed to appropriations cycles that resolve faster than the science does. The only durable protection is a public that understands why slow, risky, non-appropriable work deserves patience.

Public communication of foundational science is therefore the meta-incentive that protects the survival of every other category, since it determines whether the appropriations funding all the evaluation decisions endure across political cycles. The indicator must be scoped tightly to communication that builds durable public understanding of why foundational science matters. It is weighted by the quality and durability of understanding produced rather than by volume of engagement or self-promotion, or it degrades into the volume-first failure mode. A researcher who makes taxpayers understand why slow foundational work deserves support is protecting the entire enterprise, and that contribution should receive credit.

Formulation.

$$VPEI(a) = \Sigma \text{ over } e \in E(a) \text{ of } reach(e) \cdot quality(e) \cdot downstream_impact(e)$$

where *E(a)* is the set of engagement events for author *a* (books, op-eds, broadcasts, lectures, podcasts, verified science journalism), *reach(e)* is the verified audience size from publisher data, broadcast Nielsen ratings, or attendance records (not platform-reported only), *quality(e)* is the LLM-classified science content quality benchmarked against established science-communication exemplars, and *downstream_impact(e)* is the count of citations of *e* in policy documents, regulatory filings, educational curricula, and textbook adoption.

AI specification.

A multi-source aggregation system combining NLP classification of engagement content quality, external reach databases for audience verification, and cross-reference matching against policy-document and curriculum corpora for downstream impact tracking. Quality benchmarks calibrated against historical exemplars (Sagan's Cosmos, Hawking's A Brief History of Time,

Carson's Silent Spring) at the upper range and against demonstrated misinformation at the lower range.

Parameters and thresholds.

Each engagement event has three scored factors on a common 0 to 1 scale. Reach is the log of verified audience size, meaning log base 10 of publisher print runs, broadcast Nielsen figures, or verified attendance, normalized against the field distribution. Platform-reported figures are excluded, and only independently-verified reach is admitted. Quality is a classifier score benchmarked so the named exemplars anchor near 1.0 and demonstrated misinformation anchors near 0, with a hard gate zeroing the event if adversarial-content detection flags it, so reach cannot compensate for false content.

Downstream impact counts verified citations of the event in policy documents, regulatory filings, curricula, and textbooks, entered as the log of one plus the citation count. It is weighted most heavily, at a suggested 0.5 of the event score, because it is the least gameable and tracks durable understanding most closely. The event score is the product of the three factors. Per the failure-mode analysis, this indicator's total weight in any evaluation is capped below the constitutive indicators, since it remains a proxy rather than a constitutive metric.

Implementation.

102. Build engagement-event corpus through ORCID extension supporting author self-report of verified outputs.
103. Verify reach against publisher and broadcast data sources, with platform-reported metrics treated as unverified.
104. Score the verifiable-count dimensions computationally (independent reach verification, policy-document and curriculum citation of the content), and treat the content-substance dimension as the disclosed exception below, since substance is the one dimension the framework does not claim to compute objectively.
105. Track downstream policy and educational citation through automated matching against policy and curriculum corpora.
106. Aggregate per-author with field-normalized weighting.

Failure mode and mitigation.

This is the least Goodhart-resistant indicator in the framework, and that is stated so evaluators can weight it appropriately. Unlike the constitutive incentive-arm metrics, it is a proxy. Reach and engagement stand in for the thing actually wanted, meaning durable public understanding of foundational science, rather than being that thing directly. Reach can be gamed through paid distribution, bot amplification, or strategic platform selection, and high-reach low-substance content is the influencer failure mode. Three partial mitigations apply, independent reach verification rather than platform-reported figures, a downstream-impact term that weights policy and educational citation over raw audience, and open methodology. These reduce the gaming risk without eliminating it.

The recommendation is therefore to cap this indicator's evaluative weight below the constitutive indicators and to treat it as advisory rather than determinative. It still promotes a wanted behavior, since the cheapest path to a durable-impact score runs through producing

communication that policy documents and curricula actually cite, which is the public-understanding contribution the enterprise depends on. This is also the one place the framework does not claim objective computation of a quality dimension. Content substance resists the corpus-relational and rule-based methods the other indicators use. Rather than smuggle in the expert quality-labels the reasoning indices deliberately avoid, the framework specifies that substance judgment here is human, advisory, and the specific reason the indicator is capped.

RFI category alignment.

Public Impact.

3.7 Gap Index (GI)

Definition.

Credit accumulated by papers that fill knowledge gaps documented in the prior literature, weighted by independent identification of those gaps across distinct author groups.

Motivation.

Current metrics reward citation accumulation without distinguishing incremental work from gap-filling work that addresses documented unresolved questions. That creates a structural preference for safe incremental research. Incremental work cites established work that cites established work in self-reinforcing loops, while genuine gap-filling work that diverges from established narratives receives less initial citation.

This index tracks documented field-wide gaps, extracted from the future-directions, limitations, and open-questions sections of the literature, and it credits papers that address them with mechanism-level engagement.

Formulation.

$$GI(p) = \Sigma \text{ over } g \in G_corpus \text{ of } sim(p, g) \cdot ind(g, t_p) \cdot addr(p, g)$$

where *G_corpus* is the corpus of identified gaps extracted from "future directions," "limitations," and "open questions" sections across the literature, *sim(p, g)* is the semantic similarity between paper *p* and gap statement *g*, *ind(g, t_p)* is the independent-identification count of *g* across distinct sources prior to publication date *t_p*, and *addr(p, g)* is whether *p* actually addresses *g* mechanistically rather than asserting it.

AI specification.

A three-stage pipeline. First, a gap-extraction LLM identifies unresolved-question statements across large volumes of review articles, position papers, and limitations sections. Second, gap-clustering identifies independent restatements as the same underlying gap using semantic clustering with independence verification across author groups. Third, a semantic-match LLM scores the filling-extent for new papers, requiring mechanism-level engagement with the gap rather than rhetorical claim.

Parameters and thresholds.

Gap extraction reads sentences in the Future Directions, Limitations, Discussion, and Conclusion sections to locate stated gaps. It favors precision over recall, since a missed gap

costs less than a spurious one, and the corpus-relational logic below determines the score rather than any judgment of gap quality. Gap-clustering uses embedding cosine similarity with a merge threshold near 0.85, set per field because gap resolution varies, so same-gap statements merge and distinct-gap statements do not.

A gap qualifies for the corpus only if at least three distinct author groups stated it independently before the candidate paper's submission date, with no shared authors and no shared institution. That sets the independent-identification floor and defeats self-defined gaps. Filling-extent is a semantic-match score from 0 to 1, thresholded at 0.6 for partial credit and 0.8 for full credit, with a mechanism-engagement gate requiring the paper's Results or Methods to reference the gap, and not its Introduction alone. The independent-identification count enters as a log-weighted multiplier, the log of one plus the number of independent groups, so a gap noted by 50 groups scores higher than one noted by three, with diminishing returns.

Implementation.

107. Extract gap corpus from review articles, position papers, and Discussion sections across the literature.
108. Cluster gaps semantically with independence checks (distinct authors, distinct institutions, distinct time periods).
109. Timestamp gaps with earliest published identification.
110. Apply to new papers as filling-extent calculation with substance scoring.
111. Weight by independent-identification count and substance score.

Failure mode and mitigation.

Authors can frame work to claim gap-filling regardless of substantive contribution. Mitigation requires that the gap be documented before paper submission date (verifiable timestamp) and that the paper actually address the mechanism, scored by substance index. Independence of gap identification across distinct author groups prevents self-citation cliques from defining the gap distribution.

RFI category alignment.

Foundational Scientific Exploration and Rigor and Reproducibility.

3.8 Field-Define Index (FDI)

Definition.

Credit accumulated by papers that originate concepts, terminology, methods, or frameworks that subsequent work adopts as foundational, distinguishing originators from elaborators and co-opters.

Motivation.

Current metrics credit citation accumulation but not concept origination. A paper that introduces a foundational concept receives one citation per downstream paper using that concept, the same as a paper that uses the concept later, so originators are not distinguished from adopters. Richard Normann's Utah intracortical microelectrode array, developed at the University of Utah

through the late 1980s, defined the standard architecture for cortical recording that BrainGate and most subsequent clinical brain-computer interface work has used. The existing metrics give no extra weight to the originating papers. The Field-Define Index tracks concept genealogies through the literature, identifies originators, and weights their credit by the breadth, longevity, and fidelity of downstream adoption.

Formulation.

$$FDI(p) = \Sigma \text{ over } c \in Concepts(p) \text{ of } origin(p, c) \cdot adopt(c) \cdot stab(c) \cdot fid(c)$$

where *Concepts(p)* is the set of concepts contributed by paper *p*, *origin(p, c)* is 1 if *p* is the first substantive treatment of concept *c* and 0 otherwise, *adopt(c)* is the count of downstream papers using *c*, weighted toward adopters with no co-authorship relationship to the originator so reciprocal within-clique adoption does not inflate it, *stab(c)* is whether *c* persists in the literature 10 or more years later, and *fid(c)* is the fraction of downstream uses consistent with the originator's definition (semantic fidelity).

AI specification.

A concept-genealogy system using LLM-based concept extraction, citation-network tracking, and semantic-consistency checking to identify originators and downstream users. The system distinguishes passing mention from substantive treatment using paragraph-level context analysis.

Parameters and thresholds.

Concept extraction runs against field-stratified dictionaries seeded from UMLS, MeSH, and Gene Ontology, plus emergent-term detection for concepts not yet in ontologies. Substantive treatment, which sets the origination term, requires the concept to appear in at least two distinct sections of the paper with a definitional or empirical passage of at least three sentences. That structural rule distinguishes origination from a one-line mention, rather than matching an annotated set.

Adoption counts downstream papers using the concept, weighted by independence. Adopters sharing no co-authorship with the originator count at full weight, and within-network adopters are down-weighted by a factor near 0.3, which defeats reciprocal-citation cartels without zeroing legitimate collaborator use. Stability is a binary gate, set to 1 only if the concept still appears in the literature 10 or more years after origination, and applied only to concepts old enough to test. Semantic fidelity is the fraction of downstream uses whose local context matches the originator's definition above a 0.7 embedding-similarity threshold, which penalizes term co-option. The four factors combine multiplicatively, so a coined term with high adoption but low fidelity or no stability scores low.

Implementation.

112. Build concept-extraction pipeline over full-text corpus with field-stratified concept dictionaries.
113. For each concept, identify earliest substantive treatment using paragraph-level context (not passing mention).

114. Track downstream papers using the concept with attribution and semantic-fidelity scoring.

115. Verify persistence at 10-year horizon for concepts old enough to test stability.

116. Aggregate to paper-level scores with confidence intervals.

117. Calibrate originator and adoption thresholds against the NIH-funded reference population per Section 3.5.0 Element 4, using the Element 2 corpus and concept graph for genealogy tracking.

Failure mode and mitigation.

Two gaming modes require mitigation. First, the index can reward coining over substance, where papers introduce flashy terminology without real conceptual work. Joint conditions with mechanism-graph centrality and the substance score mitigate that, since the new concept must connect to broader knowledge structure. Second, a coordinated group can manufacture adoption by reciprocally using each other's coined terms, inflating the adoption count without genuine field uptake.

An independence check on adoption mitigates the second mode, analogous to the gap index, crediting adoption weighted toward downstream users with no co-authorship relationship to the originator, so reciprocal within-clique adoption does not drive the score. Semantic-fidelity scoring further requires downstream adoption to use the concept the originator's way rather than co-opting the term differently. The independence-weighted adoption count is the primary defense against organized citation cartels, and where residual coordinated behavior evades it, expert review of claimed originations at the career-terminal horizon provides a backstop.

RFI category alignment.

Foundational Scientific Exploration.

3.9 Cross-Field Integration Index (CFII)

Definition.

Score measuring whether a paper synthesizes methods, findings, or concepts from separate fields through genuine methodological dependency rather than tokenistic cross-field citation.

Motivation.

Current metrics fail to distinguish genuine cross-field integration from tokenistic cross-field citation. A paper can cite work from multiple fields without using those fields' methods or concepts in its own analysis, accumulating apparent breadth without genuine integration.

Wuchty, Jones, and Uzzi in Science 2007 established that team-produced cross-field work has higher disruption scores than within-field work. Current metrics still provide no way to tell which papers actually integrate fields from which papers cite them without integration. This index combines reference-list field span with semantic verification of methodological dependency on the cited work.

Formulation.

$$CFII(p) = fs(refs(p)) \cdot md(p) \cdot ds(p)$$

where *fs(refs(p))* is the average pairwise field-distance among citations in *p*'s reference list (computed from the field-distance matrix), *md(p)* is the LLM-verified methodological dependency (cited methods are used in the analysis, not just listed), and *ds(p)* is the downstream field-span of papers citing *p* (genuine integration produces downstream integration).

AI specification.

A field-distance matrix computed from topic modeling and citation networks across the corpus, plus LLM verification of methodological usage in text. The field-distance matrix is the foundational object and must be field-stratified and updated quarterly as fields evolve.

Methodological dependency is the decision that separates genuine integration from tokenistic breadth. It is scored by locating each cross-field cited method in the paper's own analysis, requiring a semantic match above 0.6 between the citation and the surrounding methods or results text. A reference cited in the introduction but never used in the analysis contributes to apparent breadth and is excluded from this term, so the score rises only when a foreign-field method is actually executed.

The field-distance component is normalized against the field-specific distribution of reference breadth, so fields with naturally broad literatures are not scored as integrative by default. Downstream field-span is computed only for papers at least five years old, and it confirms that genuine integration produced integration in the work that cited it. The 0.6 methodological-dependency threshold is calibrated against the field-specific distribution rather than against a hand-labeled exemplar set, so the separation of integrative from tokenistic breadth rests on measured method-execution in the text rather than on annotated judgments.

Implementation.

118. Build field-distance matrix from topic models on full corpus with confidence intervals.
119. Compute reference-list field-span per paper using pairwise distances.
120. LLM-verify methodological dependency by checking that cited methods appear in analysis sections.
121. Track downstream citation field-span for papers older than 5 years.
122. Aggregate as multiplicative score with field-normalized output.
123. Draw the field-distance matrix from the Section 3.5.0 Element 2 concept graph and calibrate integration thresholds against the field-stratified reference population per Element 4.

Failure mode and mitigation.

Tokenistic cross-field citation without methodological dependency can inflate the index. Mitigation requires that cited methods appear in the paper's analysis sections (semantic match between citation and surrounding methodological text), not just in the reference list. Downstream span verification at 5-year horizon further filters tokenistic claims.

RFI category alignment.

Collaboration, with secondary alignment to Foundational Scientific Exploration.

3.10 Co-Author Composition Index (CACI)

Definition.

A team-configuration score along innovation-correlated structural axes including team size (number of authors), geographic heterogeneity (international and inter-regional collaboration), institutional heterogeneity (R1, R2, government laboratory, industry, non-US institution), career-stage range, and disciplinary breadth, with optional demographic variables included only where consented through ORCID self-report.

Motivation.

Current metrics treat all author lists equally. They ignore the team-configuration variables that bibliometric research has associated with innovation outcomes across more than two decades of empirical work. Wuchty, Jones, and Uzzi (Science 2007) established that team-produced work has dominated knowledge production for over 50 years, and that team-produced papers receive substantially higher citation rates than solo-authored work. Team size also shows a positive relationship to disruption scores, up to a field-specific optimum. Later bibliometric work documented three further axes that each independently predict disruption outcomes, geographic heterogeneity (Adams 2013 Nature), institutional heterogeneity (Jones, Wuchty, and Uzzi 2008 Science), and disciplinary breadth. The Co-Author Composition Index scores team configuration along these empirically validated axes. It supplies decision-relevant signal about likely innovation output at the moment of evaluation, rather than waiting for impact to accumulate.

Formulation.

$$CACI(authors) = \Sigma \text{ over } d \in D \text{ of } w_d \cdot het_d(authors)$$

Here D is the set of structural axes, meaning team size, geography, institution, career stage, and discipline, with optional opt-in demographic variables for gender, ethnicity, and age. The term het_d(authors) is the team-configuration score along axis d, computed as a raw count for team size and as Shannon entropy or a normalized Gini coefficient for categorical axes. The weight w_d is tuned to the empirical correlation with disruption outcomes for that axis. Field-stratified weighting is required, because optimal team configurations differ across disciplines, as they do between theoretical physics and clinical translational research.

Empirical justification.

Several bibliometric studies report that varied team configurations predict innovation. They include Wuchty Jones and Uzzi 2007 Science (team size and disruption), Jones Wuchty and Uzzi 2008 Science (institutional heterogeneity and impact), Adams 2013 Nature (geographic heterogeneity), AlShebli Rahwan and Woon 2018 Nature Communications (ethnic heterogeneity and citation impact), and Yang et al. 2022 PNAS (gender heterogeneity and novelty). The empirical case rests on innovation outcomes (citation impact, disruption scores, novel-idea generation), not on normative claims about representation. CACI is justified on these empirical grounds.

AI specification.

A multi-source aggregation system extracting team-configuration from ORCID, institutional records, and (where consented) demographic self-report. Algorithmic inference of demographic

variables from names or photographs is prohibited as both privacy-violating and known to be biased. Structural variables (team size, geography, institution type, career stage, disciplinary background) are inferable from public records without requiring consent.

Each structural axis is computed by a defined rule. Career stage is binned field-relative rather than by fixed year cutoffs. Years since first publication are normalized to the field's time-to-independence distribution, meaning the field's median interval from first publication to first independent grant or last-author paper. Early, mid, and senior are therefore defined against how careers in a given field actually progress, rather than by a uniform boundary that would misjudge fields with slower or faster paths to independence. That follows the field-stratified calibration of Section 3.5.0 Element 4. Institution type is classified from affiliation against Carnegie-style categories, meaning R1, R2, government laboratory, industry, and foreign. Disciplinary background is assigned from each author's prior-publication topic distribution. Heterogeneity per axis is Shannon entropy normalized to its maximum. Team size enters as a raw count weighted against the field-specific optimum rather than on a more-is-better scale.

The gaming-relevant decision is that each co-author's axis contribution is gated by verified participation. An author counts toward heterogeneity only when CRediT contribution roles are recorded and the author has a prior-publication track record in the declared specialty. A name added purely to widen a heterogeneity axis, without a contribution role or a track record, does not raise the score. CRediT roles postdate many older papers, so retrospective application to pre-CRediT work runs the participation gate on weaker evidence. It falls back to author position and prior-topic track record without the role declaration. Retrospective CACI scores therefore have more attribution uncertainty than contemporary ones and should be reported as weaker evidence.

Implementation.

124. Extract author count and affiliation from ORCID and Web of Science author records.
125. Identify career stage from publication history (years since first publication, total publications, position type when available).
126. Identify institutional type (R1, R2, government laboratory, industry, foreign institution) and geographic location from affiliations.
127. Identify disciplinary background from prior publication topic modeling.
128. Apply demographic variables only via consented ORCID self-report (opt-in only, with explicit downstream use disclosure).
129. Compute heterogeneity scores per axis using entropy or Gini normalization, with team size as a raw count weighted against field-specific optimum.
130. Combine into composite score with empirically tuned per-axis weights, field-stratified.

Failure mode and mitigation.

Two gaming modes require naming. First, tokenistic co-authorship can inflate the index without substantive contribution from the added authors. Weighting by CRediT contribution roles partly mitigates this, as does verifying that co-authors have downstream citation linkage to prior work

in their declared field. CRediT roles are self-reported and themselves gameable, so this mitigation is partial.

Second, and harder, a well-resourced laboratory can assemble credentialed collaborators across varied geographies and institutions to raise the heterogeneity score more easily than an independent investigator can. That inverts the intent, rewarding resource access rather than the collaborative behavior itself. Measurement cannot fully close this mode, since the added collaborators are real and contributing. The framework names it for two reasons. Evaluators should weight CACI as a team-context signal rather than an individual-merit score, and CACI should never be used to disadvantage independent or single-institution investigators. Even in its gameable form the indicator promotes a wanted behavior, since assembling heterogeneous teams that collaborate is the behavior the bibliometric evidence associates with higher innovation output. CACI should be reported descriptively alongside team size rather than as a rank.

RFI category alignment.

Collaboration.

3.11 Citing Breadth Index (CBI)

Definition.

Score quantifying the breadth, balance, and timing of a paper's reference list, inverting the citation asymmetry that drives Matthew Effect concentration in current metrics. Every existing metric rewards being cited, while CBI rewards how widely and well a paper cites and characterizes the literature.

Motivation.

Current citation metrics suffer from the Matthew Effect described by Merton in 1968, where established work accumulates citations automatically while comparably valuable work outside dominant citation cliques is overlooked. Every existing metric measures who cites you. None measures whom you cite. The metric system therefore rewards being read by established networks and gives no incentive for reading widely outside them. The Citing Breadth Index inverts that asymmetry. It scores whether a paper engages original sources or only reviews, whether it spans field boundaries, whether it balances cited authors across career stages, and whether it covers geographic breadth. It also addresses an operational limitation of citation-based metrics, that they lag, by producing a signal at publication when the reference list is immediately available.

Formulation.

$$CBI(p) = \alpha \cdot OSR(p) + \beta \cdot FSpan(p) + \gamma \cdot CSB(p) + \delta \cdot GBr(p)$$

where *OSR(p)* is the original-source ratio (citations to primary work versus review papers), *FSpan(p)* is the field span of references across the field-distance matrix, *CSB(p)* is the career-stage balance of cited authors (entropy across career-stage bins), *GBr(p)* is the geographic and institutional breadth of cited authors, and weights *α, β, γ, δ* are tuned empirically against downstream impact correlation.

AI specification.

A reference-list analyzer that classifies each citation by source type (primary versus review), field, cited-author career stage, and institution, then computes balance measures across each axis. Citation context analysis ensures cited works are used in analysis (semantic match between citation and surrounding text), rather than listed.

Parameters and thresholds.

Each of the four components is normalized to a 0-to-1 range. The original-source ratio is the fraction of references that are primary sources rather than reviews, classified against journal and article-type metadata. Field span is the mean pairwise field distance across references, using the field-distance matrix of Section 3.9, normalized against the field-specific reference-breadth distribution. Career-stage balance is the Shannon entropy of cited-author career stages, binned into early, mid, and senior, normalized to its maximum. Geographic breadth is the Shannon entropy of cited-author countries and institution types, normalized the same way.

The four weights are tuned by regression against downstream impact on a held-out corpus rather than set by assertion, starting from an equal-weight prior of 0.25 each. A padding gate requires each counted citation to have a semantic match above 0.6 between the citation context and the cited abstract, so padding a reference list with unused citations does not inflate breadth. The index is computable at submission, which supplies the immediate signal the motivation names.

Implementation.

131. Parse reference lists with structured citation extraction tools.
132. Classify each citation by source type, field, cited-author career stage, and geography.
133. Compute balance measures along each axis using entropy or Gini normalization.
134. Aggregate as weighted sum with empirically tuned weights.
135. Validate weights against downstream impact correlation in retrospective data.

Failure mode and mitigation.

Reference-list padding with tangentially relevant work can inflate breadth scores without substance. Mitigation requires substance scoring on citations themselves, the cited work must be used in the paper's analysis (semantic match between citation and surrounding text), not just listed.

RFI category alignment.

Cross-cutting calibrator with primary alignment to Foundational Scientific Exploration and Collaboration.

3.12 Data-Driven Innovation Composite (DDIC)

Definition.

A learned meta-indicator that weights the preceding twelve indicators against historical disruption labels, providing a single calibrated probability that a paper will produce disruptive contribution while preserving per-component sub-scores for inspection.

Motivation.

Current metrics provide no calibrated probability of future scientific impact, reporting aggregate counts and rankings without any explicit model of what predicts innovation outcomes. Reviewers must infer likely impact from h-index numbers and citation counts using intuition that varies across panels and institutions, producing inconsistent evaluation across comparable applications. The Data-Driven Innovation Composite fills this gap by training a supervised ensemble on historical disruption labels and the twelve upstream indicators, providing a single probability output that summarizes content-level scoring across the framework while preserving per-component inspection for cases where the composite and peer judgment diverge.

Formulation.

$$DDIC(p) = f_ML(\ SBTI,\ MGCI,\ AAI,\ PSS,\ SRQI,\ EMI,\ VPEI,\ GI,\ FDI,\ CFII,\ CACI,\ CBI\ )$$

where *f_ML* is a supervised ensemble model (gradient boosting plus neural ensemble) trained on retrospective disruption scores from the Funk and Owen-Smith CD index, Wu et al. disruption score, and verifiable long-term impact labels (Nobel Prize awards, NAS membership elections, and foundational-paper recognition in field histories).

AI specification.

A supervised ensemble model with explicit bias-correction terms in the loss function, for known demographic and institutional biases in citation networks. The model is required to be inspectable, with feature importance reported per inference and per-component sub-scores always available.

The upstream indicators are not statistically independent. SBTI already incorporates mechanism-graph centrality and gap-filling status among its features, and several indicators share the substance-score calibrator. The ensemble handles that correlation directly, since tree and neural ensembles do not assume feature independence. Feature-importance reporting must still attribute shared signal transparently rather than obscuring that correlated inputs contribute jointly, so evaluators are not misled into treating correlated components as independent evidence.

Scope.

This composite is computed at two explicit scopes, to keep its inputs coherent. The paper-level composite uses the twelve inputs in the formulation above. Ten of those are paper-level, meaning SBTI, MGCI, AAI, PSS, SRQI, EMI, GI, FDI, CFII, and CBI. Two are author-level or team-level, meaning VPEI and CACI, and they enter only through their paper-attributable components rather than as raw author or team scores. SRQI and EMI enter as their composite meta-index scores rather than as their individual step-indices.

An author-level composite additionally incorporates the full author-level and team-level indicators, including the Independent Recognition Indicator of Section 3.13. That indicator is

excluded from the paper-level composite because it is an author-level measure. Stating the two scopes explicitly prevents a category error, feeding an author-level score into a per-paper prediction. It is also why the paper-level formulation lists twelve inputs while the author-level composite additionally draws on IRI, and why neither draws on the SRQI and EMI step-level sub-indices directly, since those enter only through their meta-index composite scores.

Implementation.

136. Compute per-indicator scores for the training corpus across the past 30 years of literature.
137. Label with disruption scores at 20-year horizon plus categorical labels for Nobel-recognized work and historical foundational papers.
138. Train ensemble with bias correction for gender, ethnicity, institutional tier, and geographic-region biases verified in citation networks.
139. Validate against held-out historical disruption cases including Karikó, Brock, Normann, and Sagan papers.
140. Publish model weights, feature importance, and training data composition openly.

Failure mode and mitigation.

The composite concentrates evaluation authority in a single model that can become opaque if governance is weak. Mitigation requires open-source model with public audit trail, use as advisory input to peer review rather than autonomous decision-maker, mandatory per-component sub-score inspection in any application, and external audit every 18 months by an independent bibliometric panel.

RFI category alignment.

Meta-indicator applicable across all RFI categories.

Consequence and calibration are separate reported axes, and calibration cannot raise the consequence read.

Two axes are reported separately and are never netted into one number, a consequence axis and a calibration axis. The consequence axis estimates how much a contribution changes what a field asks, does, or believes, and it aggregates the forward-looking signals, the disruption probability of this composite, the Field-Define Index of Section 3.8, the Sleeping-Beauty Trajectory Index of Section 3.1, and the mechanism-graph centrality shift of Section 3.2. A calibration axis estimates the honesty and rigor with which a paper supports its stated claims, and it aggregates the SRQI and EMI reasoning scores, the claim-alignment check, and the honest-limitations credit. Each paper receives an independent score on each axis, and both scores appear side by side rather than combined, so an evaluator reads a consequence estimate and a calibration estimate rather than one value that obscures which estimate produced it.

The separation enforces one rule. Calibration cannot raise the consequence read. A rigorous, well-calibrated paper reporting a small but correct increment receives a high calibration score with a low or uncertain consequence score rather than a strong-contribution score, because honest reporting of a minor result does not make that result major. Substitution of calibration for

magnitude is the most likely failure. Calibration components number eleven reasoning steps, six engineering steps, and the claim-alignment and limitations checks, against four indicators on the consequence side, so an evaluator reading the full vector can treat the larger volume of calibration signal as evidence of magnitude. Separate axes and a prohibition on netting remove that substitution, so an honest minor paper and a field-redirecting paper separate on the axis that discriminates them rather than combining on the axis that does not.

The consequence axis has the widest uncertainty in the framework, because magnitude at publication time is a forward prediction rather than a measurement, so it is reported with an explicit confidence interval rather than a point estimate. Low predicted consequence is not scored as a deficiency, because the rule that incomputability scores high of the founding-novelty case applies here. A contribution too far ahead of its field for measurable downstream signal is flagged as high-uncertainty and possibly dormant rather than as minor, which preserves the protection the Karikó case in the preface establishes. Three states are therefore distinguished rather than two, high consequence, low consequence from a small increment, and unresolved consequence for a contribution a field has not yet incorporated. Downstream signal accrues over time, so an early high-uncertainty score is revisited rather than fixed, and the retrospective validation of Section 7.5 tests whether the axis separates a small increment from a dormant one.

3.13 Independent Recognition Indicator (IRI)

Definition.

A distribution-robust measure of how much of a researcher's recognition comes from authors with no co-authorship relationship to them. It is computed per paper across the entire publication portfolio, then aggregated by a median, which prevents one or two high-citation papers from dominating the result. It therefore distinguishes recognition earned from independent groups from recognition inherited through collaboration networks or borrowed from senior co-authors.

Motivation.

Current citation counts conflate independent recognition with network-internal citation. A researcher heavily cited by their own collaborators, trainees, and co-authors therefore appears identical to one cited by strangers who found the work on its merits. That favors researchers embedded in large collaboration networks and disadvantages independent investigators whose work is recognized outside any citation clique, which reproduces the Matthew Effect at the author level. This indicator uses citation author-position analysis to separate two kinds of citation, those from non-collaborators who have never co-authored with the researcher, and those from inside the collaboration network. It therefore measures recognition that network membership cannot explain. This is the one salvageable mechanism from perception-based reputation scores, extracted as an objective network measurement rather than a subjective impression.

Formulation.

Computing independence over pooled citations across a researcher's whole portfolio lets one or two high-citation papers dominate the statistic. Citation distributions run from lognormal to

power-law, so a small number of papers typically hold most citations. A researcher with one widely-cited method paper would then have their independence score determined almost entirely by that paper rather than by their body of work.

The indicator avoids this by computing independence within each paper first, then aggregating with a distribution-robust central tendency that treats each paper as one observation rather than weighting by citation count.

$$r(p) = |\{\ q \in Cit(p) : coauth(a,\ author(q)) = 0\ \}| \ /\ |Cit(p)|$$

computes the per-paper independence ratio, where *Cit(p)* is the set of papers citing paper *p* and *coauth(a, author(q)) = 0* holds when no author of citing paper *q* has ever co-authored with researcher *a*. Each ratio is then field-normalized against same-field, same-era papers, and the researcher-level indicator is the median across the portfolio:

$$IRI(a) = median\ over\ p \in Papers(a),\ |Cit(p)| \geq c_min\ \ of\ \ znorm(\ r(p)\ )$$

where *znorm(...)* field-normalizes the per-paper ratio to a z-score against the distribution of independence ratios for same-field, same-era papers (correcting for smaller fields having fewer possible non-collaborators), and *c_min* is a minimum citation floor excluding near-zero-citation papers that would add noise. The median across papers is robust to the one or two outlier papers distortion as a mathematical consequence, since a single dominant paper is one observation among many rather than the bulk of a pooled sum, which prevents the skewed-portfolio artifact that a citation-weighted pooled fraction produces.

AI specification.

A citation-network analysis over a co-authorship graph, computed per paper and then aggregated. The co-authorship graph is built from the full publication corpus. For each citation to each paper, the system checks whether any author of the citing paper has ever co-authored with the researcher. Per-paper independence ratios are field-normalized against same-field same-era distributions before the median aggregation. This requires accurate author disambiguation, which ORCID and modern name-disambiguation models provide with high but imperfect accuracy.

Parameters and thresholds.

Co-authorship is defined over all prior years with no decay, so any shared authorship at any time marks a citing author as network-internal. The citation floor excludes papers below the field-median citation count for papers of the same age, which removes low-signal papers without discarding recent or early-career work disproportionately, since the floor is relative to field and age rather than an absolute count. Hyperauthorship papers above 50 authors are excluded from the co-authorship-relationship computation, since consortium membership does not indicate a genuine collaboration tie.

Aggregation uses the median across qualifying papers as the primary statistic. The interquartile range is reported alongside it, which exposes a bimodal portfolio where some papers are highly independent and others network-internal. The full per-paper distribution is shown rather than collapsed, so a researcher is characterized rather than ranked. Where disambiguation

confidence falls below a per-author threshold, the researcher is flagged for manual review rather than given a possibly-erroneous automated score.

Implementation.

141. Construct co-authorship graph from full corpus with ORCID-based author disambiguation.
142. For each paper by the researcher, compute the per-paper independence ratio (non-collaborator citations over total citations to that paper).
143. Field-normalize each per-paper ratio against same-field, same-era papers to a z-score or percentile.
144. Aggregate across the researcher's papers using the median (or trimmed mean) over papers above the citation floor, so no single high-citation paper dominates.
145. Report alongside raw citation count and the per-paper distribution so evaluators see recognition quality across the portfolio, not just a pooled fraction driven by outliers.

Failure mode and mitigation.

Three failure modes require mitigation. Skewed-portfolio distortion, where one or two high-citation papers would dominate a pooled fraction, is addressed by the per paper, then median aggregation above. Author disambiguation errors misclassify co-authorship relationships, mitigated by ORCID-verified disambiguation. Large collaborations (consortium papers with hundreds of authors) distort the co-authorship graph, mitigated by excluding hyperauthorship papers above a co-author threshold from the relationship computation. The indicator is reported as a descriptive pattern with its per-paper distribution rather than a single rank, so network-embedded researchers are characterized, not penalized.

RFI category alignment.

Collaboration, as a measure of recognition independent of collaboration-network membership.

3.14 Reconciliation of Overlapping Indices

Note on deliberate overlap.

Building the SRQI step-indices fresh produced intended overlaps with three existing indicators, and the framework resolves each by scope rather than by deletion.

Gap identification appears twice. The reasoning-level Gap-Identification Index of Section 3.5.2 scores how well a single paper identifies and argues its gap from its own text. The network-level Gap Index of Section 3.7 tracks gaps across the literature and credits filling a gap that others independently identified. The first is a within-paper reasoning measure and the second is a cross-literature contribution measure. Synthesis also appears twice. The reasoning-level Synthesis-and-Integration Index of Section 3.5.10 scores connection density in the paper's own argument, while the Cross-Field Integration Index of Section 3.9 scores reference-list field span together with downstream integration, so the first reads the reasoning and the second reads the citation structure. Mechanism density appears as the mechanism dependency inside SSyn and as the Mechanism-Graph Centrality Index of Section 3.2, which scores a paper's concepts against the field-wide mechanism graph, so the first is internal and the second is network-level.

The pattern is the same in all three cases. The step-index is the within-paper reasoning signal, and the standalone indicator is the network or literature signal. They are reported separately, and the Data-Driven Innovation Composite treats them as correlated inputs per its scope note rather than as independent evidence.

RFI category alignment.

Cross-cutting, clarifying the relationship among Rigor, Foundational Exploration, and Collaboration indicators.

3.15 Gate-Contract Instantiation for Every AI-Based Component

The table fills the gate contract of Section 3.5.0 for each AI-based metric, so the building phase inherits a populated specification rather than a principle it must remember to apply. Each row is a required deliverable, a blank cell blocks the metric from build-readiness, and the acceptance-test column is the field that detects a build populated in form but not in function. Metrics with no content gate are named after the table as rule-based or constitutive, with the reason each is exempt, so the boundary is fixed here rather than left for a builder to draw.

Metric	Grounding artifact (emitted, ablation-verified)	Retrieval or evidence scope	Held-out features	Acceptance test (known-outcome)
SObs (3.5.1)	Retrieved k=50 prior-neighbor set with entailing sentences (novelty), plus an examined primary-data sample, waveform, trace, or image (fidelity)	Target subfield plus adjacent fields by geodesic	Categorical-language strength, venue prestige, abstract salience, narrative framing	Credits a corroborated novel observation, flags a re-presented known finding, and flags physiologically implausible primary data on held-out cases
SGap, SHyp, SSyn (3.5.2, 3.5.4, 3.5.10)	Retrieved prior claims that leave the gap open, that the hypothesis departs from, or that the synthesized link connects, each with source sentences	Target plus adjacent and connected fields by geodesic	The paper's own claim of a gap, of originality, or of interdisciplinarity	Credits a gap or link absent from the prior set, flags one the prior set already contains
SExec (3.5.7)	Extracted procedural parameters checked against the per-method-class reproducibility checklist (ARRIVE, MIQE, or a community equivalent)	Method-class checklist, not a corpus	Method-section length and the words rigorous or validated	Flags a named reagent without concentration or an antibody without catalog number, and flags results the stated method could not produce
SAna (3.5.8)	The paper's reported statistics and extracted data structure, checked against fixed statistical rules	Fixed rules of statistics and the paper's internal logic, not a corpus	The word significant without a p value, and conclusion-strength language	Flags test-to-data-type mismatch, pseudoreplication, and a conclusion about a quantity the test did not measure
MGCI (3.2)	The extracted mechanism-graph edges with the source sentence asserting each edge, emitted for human check	Biomedical relation-extraction corpus plus adjacent fields by geodesic	Claim language asserting a concept is central or important	Extracted edges match human-adjudicated edges above a stated F1, and centrality reproduces known hub concepts on

				held-out graphs
AAI (3.3)	Retrieved prior papers that state or rely on the assumption the paper claims to overturn, correct, or build past, with the entailing sentences	Target plus adjacent fields, so an assumption settled next door is caught	The paper's framing that it is the first to, and venue	Credits a genuine correction, does not credit a paper overturning an assumption no prior neighbor held, and recognizes a correction already made in an adjacent field
GI (3.7)	Retrieved prior papers documenting the gap across distinct author groups, with the documenting sentences	Target plus adjacent fields	The paper's self-declaration that a gap exists	Credits a gap independently documented by at least a stated number of distinct groups, does not credit a self-declared gap no prior work states
FDI (3.8)	Retrieved later adopting papers plus the first-use origin evidence, with citation-context sentences separating originator from elaborator and co-opter	Target plus adjacent fields plus downstream adopters	The paper's self-claim of origination, and venue prestige	Attributes origination to the true originator on held-out concept-lineage cases, does not credit a popularizer over the originator, and catches prior origination in an adjacent field
CFII (3.9)	Retrieved prior work in each connected field plus the sentences showing the borrowed method or concept is used in the analysis, rather than cited	Each connected field plus adjacents by geodesic	Cross-field citations that are tokenistic, and interdisciplinarity framing	Credits genuine methodological dependency and flags tokenistic cross-citation on held-out cases
VPEI (3.6)	The examined communication artifacts, verified reach records, and downstream policy or educational citation records	Content-quality scoring against a domain-accurate reference	Reach inflated by virality or controversy, and follower counts	Credits substantive accurate communication, does not credit high-reach low-substance output, verified against known policy-citation outcomes
CBI (3.11), content part	The reference set with its field, recency, and group distribution (structural), plus the entailment that each	The paper's reference list and the cited works	Raw reference count and self-citation padding	Rewards broad and apt referencing, does not reward a padded or off-target reference list

	sampled citation supports its citing sentence (content)			
SBTI (3.1), content part	Early-trajectory features (structural), plus the prior-neighbor set showing the paper is ahead of its contemporaries (content)	Target plus adjacent fields for the content part	Early citation count and venue	Content score is high at publication on known sleeping-beauty cases, does not fire on mere obscurity or low quality
DDIC (3.12), meta-learned	The per-component sub-scores it weights, each already gated and inspectable	The twelve upstream indicators, no direct corpus retrieval of its own	Raw citation counts, so it cannot learn to proxy citation	Predicts held-out disruption above a citation-only baseline, and preserves per-component transparency
EMI steps (3.5E)	The quantified requirement, the explored design points, and the validation results against each requirement, plus the prior-engineering neighbor set for the novelty axis	Prior engineering literature plus adjacent fields	Naming an operation in place of performing it, and device-marketing language	Credits a quantified and validated design, flags an unquantified requirement or a validation that skips a requirement, on held-out engineering papers
Analytical -mode floor (3.5A)	The retrieved prior set for the linkage-novelty check, and the citing sentences for the source-characterization check	Target plus adjacent fields	Fluent synthesis language and the number of citations spanned	Flags a synthesis that re-states known links, one that mischaracterizes a cited source, and one that smuggles an unstated assumption

Four metrics ground their scores in a retrieved set of prior claims rather than in primary data, meaning AAI, GI, FDI, and CFII. Their contract is met only when that retrieved set is the object the score is computed from, rather than a set displayed beside a score computed some other way. The ablation applies with full force here, since a relational metric is the easiest place to emit a nominal artifact. A builder who retrieves the neighbor set but scores from the paper's own framing has produced the presentation-gamed score the contract forbids. Each of these acceptance tests therefore includes a claim whose framing asserts a contribution the retrieved set does not support, and the metric passes only if it scores that claim on the set rather than on the framing.

The Data-Driven Innovation Composite is a learned meta-weighting rather than a content scorer, so its grounding artifacts are the per-component sub-scores it combines, each with its own gate, and its distinctive risk is learning to reproduce citation rank because citation is the easiest signal correlated with its disruption labels. Its held-out feature is therefore the raw citation count, excluded from its inputs so it cannot proxy citation, and its acceptance test requires out-of-sample disruption prediction that beats a citation-only baseline. A composite that

reproduces the citation ordering has added nothing and has reintroduced the incumbent it would replace.

Three metrics are part structural and part AI, and the contract binds only their AI part. SBTI computes early-trajectory features by rule and scores content novelty by the SObs machinery, so it inherits the SObs row for its content half. PSS computes data novelty and the primary-data ratio by provenance matching, and it reuses the SAna checks for statistical rigor, so its AI exposure is the SAna row rather than a separate contract. CBI computes breadth, balance, and timing from structure and scores citation aptness by entailment, so only the aptness component includes the content gate. Naming which half is AI keeps the structural half from being treated as gated, and the AI half from being treated as rule-based.

Two indicators have no content gate, because they compute from structure rather than from reading claims against priors. The Co-Author Composition Index derives team configuration from authorship and consented ORCID fields. The Independent Recognition Indicator computes over the co-authorship graph. Both are therefore rule-based, and their correctness requires correct counting rather than correct judgment. The constitutive metrics of the incentive arm are also rule-based. Stating this boundary explicitly matters, because the failure this whole section guards against is a builder assigning an AI metric to the rule-based side to escape the gate. The ablation is the enforcement. A metric that claims to be rule-based, but whose score moves when a retrieved neighbor set is withheld, is an AI metric operating without its contract.

Each row's acceptance test joins the Element 6 acceptance gates and the thin-prior stress test of Section 7.4 as a per-metric deployment condition, so a metric ships only when it passes the test written for its own failure mode. The framework that reaches deployment is therefore the subset of AI metrics that are real, rather than the full set with some built cosmetically and offset by an aggregate.

4. Diagnostic Arm: Indicator-to-RFI Category Mapping

The following table summarizes which of the fourteen top-level diagnostic-arm indicators address each of the seven RFI categories, where the Scientific-Reasoning Quality Index and Engineering-Method Index are meta-indices each composing named step-level sub-indices (Sections 3.5 and 3.5E). Indicators with primary alignment are noted as primary, and indicators with secondary alignment are noted as secondary. The incentive arm (Section 5) maps separately.

RFI Category	Indicators (primary and secondary)
Rigor and Reproducibility	AAI (primary), GI (primary), SRQI (primary), PSS (secondary)
Data, Software, and Model Sharing	Addressed by RGI (Reuse Generation Indicator) in the incentive arm, Section 5.3.1
Training and Mentorship	Addressed by TTI (Trainee Trajectory Indicator) in the incentive arm, Section 5.3.3
Collaboration	CFII (primary), CACI (primary), IRI (primary), CBI (secondary)
Entrepreneurship and	EMI (primary, for design-and-build work). Existing metrics

Translation	overweight translation as an outcome, so the framework adds no outcome-based translation indicator, but the Engineering-Method Index scores the quality of the engineering reasoning itself, which existing metrics ignore
Foundational Scientific Exploration	SBTI (primary), MGCI (primary), FDI (primary), GI (secondary), SRQI (secondary), CBI (secondary)
Public Impact	VPEI (primary)
Cross-cutting	PSS (substance calibrator), SRQI (reasoning-quality calibrator), CBI (real-time signal), DDIC (meta-composite)

Foundational Scientific Exploration is intentionally over-represented in the framework, with three primary indicators plus two secondary, because the existing metrics most damage this category. Entrepreneurship and Translation is intentionally not addressed with new diagnostic indicators because existing metrics already overweight translation, and the structural argument of this submission is that translation as the primary metric is itself the problem driving the false-positive failure mode.

5. Incentive Arm: Metrics Whose Construction Rewards Wanted Behaviors

This arm is the primary contribution of the submission. Most of its indicators require no artificial intelligence and can be enacted through human judgment and verified counts, which makes them deployable now and independent of whether any diagnostic-arm scorer is ever built. Where an incentive indicator does use AI components, it reuses the diagnostic-arm infrastructure and inherits the same gating preconditions. The arm leads with foundational and exploration incentives because those behaviors have the longest payoff horizon and the weakest existing constituency.

5.1 Design Principle: Constitutive Rather Than Proxy Metrics

The diagnostic indicators in Sections 3 and 4 detect quality the existing metrics miss, and they share the vulnerability of all proxy metrics. Goodhart's Law holds that when a measure becomes a target it stops being a good measure, because people optimize the proxy rather than the quality it stands for. That applies to any metric that stands in for something else.

It applies far less to a constitutive metric, where the measured behavior is itself the thing wanted. Consider a system that credits verified independent reuse of a shared dataset. The only way to score high is to produce a dataset other groups actually reuse, so gaming the metric is the wanted behavior. The incentive arm is built entirely from constitutive metrics. Each one is designed so that the dominant strategy for a researcher optimizing it is to perform more of the intended behavior.

The incentive arm rewards behaviors at their cost point, meaning the place where a researcher currently pays a career penalty for doing the wanted thing. Exploration costs failure risk, so the arm credits the failed high-risk attempt rather than only the success. Replication costs novelty, so it credits the replication. Data and tool sharing costs competitive advantage, so it credits verified downstream reuse. Corrective work costs the appearance of productivity, so it credits the dead end retired. Each indicator neutralizes the specific disincentive that currently suppresses a behavior the public-funding mission depends on.

The incentive-arm indicators vary in how fully constitutive they are, and the framework specifies this rather than claiming uniform gaming-resistance. The reuse, replication, corrective, and trainee indicators are fully constitutive. The only way to score high on them is to produce reusable artifacts, run replications, publish genuine corrections, or train independent scientists, so optimizing the metric is the wanted behavior.

Two indicators are only partly constitutive. The research risk portfolio is gameable by cheap-risk accumulation until its initiation-assessment mechanism exists, and the foundational investment indicator is gameable by relabeling incremental work. Even these tend to promote productive behavior, because the cheapest path to a high score still runs through attempting more ambitious work, or through foregrounding mechanism over translational packaging, and both are intended behaviors even in their gamed form. Where a residual gaming mode cannot be closed, the framework names it, so evaluators can weight the indicator appropriately and later work can address it, rather than presenting the metric as fully robust.

The incentive arm should receive heavier evaluative weight than the diagnostic arm, and the foundational and career-terminal metrics within it should receive the most. That weighting is structural rather than technical. Study sections reward near-term clinical translation because it is fundable within a grant cycle. Public and congressional sentiment reward the same, because visible cures arrive during the appropriations window.

Both forces already push toward the short horizon. An evaluation system that weights all work equally therefore under-invests in the long-horizon foundational research that has the greatest cumulative payoff and the weakest existing constituency. Heavier weight on long-horizon incentives corrects that standing imbalance, and only a public funder with a multi-decade mission has the mandate to make the correction. The incentive metrics are largely Goodhart-resistant, which is what makes the heavier weight safe to apply. The two partly-constitutive indicators are weighted more cautiously until their residual gaming modes are closed.

5.2 Foundational and Exploration Incentives

5.2.1 Research Risk Portfolio Indicator (RRPI)

Definition.

A career-spanning score of the high-variance, high-potential-value work a researcher has attempted, crediting the attempt itself including failed attempts, conditioned on the potential value of each attempt if successful.

Motivation.

Current metrics reward outcomes. A researcher who attempts ten high-risk projects and succeeds at one therefore scores worse than one who completes ten safe incremental projects, even though the first produced the more valuable result and advanced the field's state of knowledge. That penalizes the exploratory behavior foundational science requires, and only public funding tolerates it, because private capital cannot absorb a portfolio with a high failure rate. The Research Risk Portfolio Indicator credits the attempt at high-value risk rather than only the success. A researcher optimizing this indicator therefore attempts more ambitious work, which counters the chilling effect that suppresses exploratory volume.

Formulation.

$$RRPI(a) = \Sigma \text{ over } j \in Projects(a) \text{ of } risk(j) \cdot value_if_success(j)$$

where *risk(j)* is the prior probability of failure at project initiation (higher risk scores higher, and the score does not depend on whether the project succeeded), and *value_if_success(j)* is the potential value of the project conditional on success, assessed at initiation. Conditioning credit on potential value prevents the cheap-risk-accumulation failure mode, in which a researcher accumulates risk credit through many low-value attempts unlikely to succeed and unlikely to matter. Only high-stakes risk earns credit, not mere failure.

Implementation.

Robust specification requires a registered-report-style initiation mechanism. At project start, before outcomes are known, peer assessment records two things, the project's risk, meaning its probability of failure, and its potential value if it succeeds. This front-loaded assessment is

reliable because it does not require estimating the counterfactual payoff of already-failed work, which post-hoc scoring cannot do accurately.

A lighter interim specification uses AI scoring at publication or preprint time, measuring novelty distance from the researcher's prior work and departure from field consensus. That version is buildable immediately, and it is weaker, because it cannot assess potential value as dependably and is more exploitable by cheap-risk accumulation. The registered-report mechanism requires NIH to build initiation-time assessment infrastructure. That is a program-design investment beyond pure measurement, and it sits squarely within the incentives the RFI solicits.

Failure mode and mitigation.

Cheap-risk accumulation, meaning credit gathered through low-value long-shot attempts, is the dominant failure mode. Conditioning all risk credit on potential value assessed at initiation mitigates it. That mitigation depends on the registered-report initiation mechanism, which does not yet exist. The lighter interim AI-only specification therefore has no working defense against cheap-risk accumulation, and it should receive zero evaluative weight in funding or promotion decisions until the initiation-assessment infrastructure is built. The interim version is useful descriptively, for surfacing exploratory-work patterns, not for ranking.

Even in its unguarded interim form the indicator promotes a wanted behavior. The cheapest path to a high score still runs through more ambitious departures from consensus, which raises exploratory volume even without the potential-value conditioning. Peer-assessment capture at initiation is a secondary risk in the full version. Blinded multi-reviewer assessment mitigates it, together with periodic audit of initiation scores against realized trajectories.

RFI category alignment.

Foundational Scientific Exploration (primary).

5.2.2 Foundational Investment Indicator (FII)

Definition.

A score crediting sustained investment in basic, mechanism-level, and pre-competitive research that has no near-term translational application at the time of the work, measured by the depth and continuity of foundational contribution rather than by translational output.

Motivation.

Current metrics and study-section incentives reward translational framing. Researchers doing foundational mechanism work must therefore dress it in translational language to stay funded, which distorts the work and defunds the pure foundational inquiry with the longest eventual payoff. Basic science that industry will not touch, because its returns are decades away and no single actor can capture them, is the public-funding mission, yet no metric rewards it. The Foundational Investment Indicator credits sustained foundational contribution directly. Doing basic science is then rewarded rather than penalized, and researchers need not disguise mechanism work as near-term translation.

Formulation.

$$FII(a) = \Sigma \text{ over } p \in Papers(a) \text{ of } found(p) \cdot depth(p) \cdot (1 + delayed_realization(p))$$

where *found(p)* measures how far paper *p* is foundational rather than translational (scored from content, high for mechanism and theory work with no near-term application claim), *depth(p)* reuses the mechanism-graph centrality and paper-substance scores, and *delayed_realization(p)* is a bonus that accrues if the foundational work later enables translation, rewarding the researcher retroactively for foundational bets that matured.

Implementation.

Foundational-versus-translational scoring reuses the diagnostic-arm content classifiers, inverting the translation-hook detection so that absence of a near-term application claim combined with high mechanism centrality scores high. The delayed-realization bonus links foundational papers to later translational work that cites them as enabling, computed at long horizons (10 or more years), which requires the same longitudinal citation infrastructure as the sleeping-beauty indicator.

Failure mode and mitigation.

Researchers could relabel incremental work as foundational to capture the credit, mitigated by joint conditioning on mechanism-graph centrality and substance scores, so that foundational credit requires demonstrated mechanistic depth rather than a mere absence of translational claims. Field-stratified calibration is required because the foundational-translational balance differs across disciplines.

RFI category alignment.

Foundational Scientific Exploration (primary).

5.3 Behavior Incentives

5.3.1 Reuse Generation Indicator (RGI)

Definition.

Credit for verified independent reuse of datasets, code, reagents, protocols, and other research artifacts a researcher has shared, weighted by the independence and productivity of the reusing work.

This indicator addresses the Data, Software, and Model Sharing category the RFI names and the diagnostic arm defers. It is constitutive, since scoring high requires producing artifacts that other groups reuse, so gaming it is the wanted behavior. Verified reuse is tracked three ways, through data and code citation, through repository fork and download provenance, and through reagent-request records. Each is weighted by whether the reusing group is independent of the originator. Sharing costs competitive advantage, and this indicator credits at that cost point. Foundational and infrastructure artifacts are the highest-value case, meaning shared datasets, reference methods, and model organisms, because they are pure public goods the private sector will not create. RFI category alignment: Data, Software, and Model Sharing (primary), and infrastructure commons.

Formulation and data availability.

RGI(r) sums, over the artifacts a researcher has shared, the independence-weighted count of downstream reuses, where each reuse is weighted by whether the reusing group shares no co-

authorship tie (the IRI independence machinery) and by the reusing work's own substance, so a dataset reused by many independent productive groups scores far above one reused only by the originating lab. Reuse is detected from data-and-code citation (DataCite and software-citation records), repository fork and download telemetry where available, and formal reagent-and-materials citation.

Data requirement, partially available. Citation of data and code through DataCite and Software Heritage, together with repository telemetry, is queryable today, so the dataset-and-code portion is buildable now. Systematic reagent-and-materials reuse tracking is not. No queryable resource records which cell line, plasmid, antibody, or mouse line a downstream paper obtained from whom.

Capturing the reagent-and-materials contribution the definition names as highest-value therefore requires a materials-and-data-reuse registry that does not currently exist. NIH is uniquely positioned to build it, maintain it, and drive its adoption, since such a registry is itself non-appropriable public-good infrastructure the private sector will not create, and building it is a program-design investment squarely within the incentives the RFI solicits. Adoption would plausibly run through linkage to existing repositories such as Addgene, ATCC, and model-organism stock centers, and through a data-management-and-sharing-plan reporting requirement. Until that registry exists and reuse reporting is normalized, this indicator measures data-and-code reuse well and reagent reuse poorly, and the reagent gap should be stated wherever it is reported.

5.3.2 Corrective work (folded into the Assumption-Addressing Indicator)

Corrective and negative work earns near-zero annual reward but has large cumulative value. A field spared decades of chasing a wrong model is an enormous saved cost, and only non-commercial science performs this work, because there is no product in proving something does not work.

Earlier versions scored this with a separate corrective-contribution indicator. What that indicator measured at publication time, meaning how well a paper engages and overturns a wrong assumption, is the same act the Assumption-Addressing Indicator of Section 3.3 now scores. Its lagging question, meaning whether the correction was right and whether the field came to agree, is now the career-terminal Assumption-Correctness Indicator of Section 5.4.2. The two-metric split is deliberate. The quality of a corrective engagement is observable at publication, while its correctness is observable only over years, and binding them into one score would force the correctness question to be answered prematurely. Corrective work is therefore credited by AAI at the paper level and by the career-terminal correctness metric across a career, rather than by a separate incentive-arm indicator here.

5.3.3 Trainee Trajectory Indicator (TTI)

Definition.

Credit for the positive career outcomes of a researcher's trainees, broadly construed to include both independent scientific success and successful careers outside research, weighted toward outcomes that reflect the trainee's own trajectory rather than the mentor's. A trainee who launched an independent research direction distinct from the mentor's and a trainee guided into a well-fitting non-research career both count as successes, and not the former alone.

A career's largest long-term multiplier is often the people it launched, and current metrics credit none of this terminally. This indicator credits a trainee's independent scientific success, meaning their own funding, their own foundational contributions, and their own trainees. It credits verified successful non-research careers equally. Both are weighted toward divergence from the mentor's exact topic, so the metric rewards training for independence rather than cloning.

One design principle is essential here. The metric must not reward retention for its own sake, meaning the pattern of keeping a poorly-suited trainee in the laboratory for years of low-productivity work before an eventual washout. Recognizing early that a trainee is not suited to research, and guiding them into a career where they will thrive, is a good mentorship act that saves the trainee years of lost productivity. This indicator therefore counts that redirected trainee as a success rather than a zero, and it never scores a departure as a negative. It is constitutive, since scoring high requires producing people who succeeded, in or out of research, rather than retaining bodies. This addresses the Training and Mentorship category the RFI names and the diagnostic arm defers. RFI category alignment: Training and Mentorship (primary).

Formulation and data availability.

This indicator sums each verified trainee's positive outcome score across a researcher's trainees. A research-track trainee contributes their independent scientific success, meaning independent funding, last-author output, and their own trainee production, weighted toward topical distance from the mentor by the SObs field-distance machinery. A non-research-track trainee contributes a verified successful career outcome, meaning a stable position in a science-adjacent field, industry, policy, education, or another profession. Departures are never scored negatively, so a mentor is not penalized for a trainee who left. A trainee redirected early into a well-fitting career counts as a positive outcome, which removes any incentive to retain a poorly-suited trainee for the mentor's short-term labor. The sum is normalized by trainee count and career stage, so large laboratories do not dominate by volume and recent mentors are scored on their trainees' trajectories so far with appropriate uncertainty.

Data requirement, partially available and partly manual, with one component difficult. Mentor-trainee links are recoverable from co-authorship followed by independence patterns, and from dissertation and ORCID records for some trainees. A per-laboratory research-track score is therefore assemblable by hand today. Population-scale automated scoring requires the training-lineage registry listed in Section 6.6, which NIH would build and which does not currently exist. The non-research career outcome component is harder still. No systematic source tracks where trainees went after leaving academia, so that component would depend on the same registry incorporating voluntary career-outcome reporting, plausibly through institutional alumni tracking or a trainee-outcome reporting requirement tied to training grants.

One limitation is intrinsic and stated here. Honest early redirection of a poorly-suited trainee, and neglectful churn of a mismentored trainee, can produce similar external traces, since both show a trainee leaving. Distinguishing them requires knowing whether the departure served the trainee, which the records do not hold. The rule against scoring departures negatively avoids penalizing honest redirection, and it also means the metric cannot by itself reward good redirection over bad churn. Fully separating the two would require a trainee-side signal, meaning an alumni assessment of whether the training served their eventual career, which no

current mechanism collects and which would raise its own gaming and privacy considerations. Until such a signal exists, this indicator credits verified positive outcomes and stays silent on departures, rather than pretending to judge the quality of a departure it cannot observe.

5.3.4 Replication Contribution Indicator (RCI)

Definition.

Credit both for performing replication studies and for producing original work that independent groups successfully replicate, where the replication relationship between two papers is detected by comparing their experimental structure across the corpus rather than requiring either paper to declare itself a replication.

Replication scores near zero on current career metrics despite being foundational to a reliable evidence base. This indicator credits two things, performing replications, which cost novelty, and producing replicable work, verified when an independent group re-runs the design and confirms the result. Detection compares experimental structure across the corpus, so a replication need not declare itself. The same comparison distinguishes three relationships a later paper can bear to an earlier one, and only replication is credited here. It is constitutive on both sides, since scoring high requires either running real replications or writing methods rigorous enough that independent replication succeeds. RFI category alignment: Rigor and Reproducibility (primary).

Formulation and data availability.

This indicator detects the relationship between a later paper and an earlier one by comparing experimental structure together with the temporal and citation pattern between them. The compared structure is the experimental groups, control groups, model system, and measured endpoints, extracted by Element 1 of Section 3.5.0 and matched by Element 2. The relationship is then classified into one of three, and only replication earns credit.

Pure replication is a large publication-time gap, plus a back-reference to the earlier paper, plus a re-run of substantially the same design. It earns credit for the replicator and, on the other side, for the original paper whose work proved replicable when the results concord. Independent convergence is a small time gap, within the field-calibrated concurrent-submission window, with no back-reference, which indicates two groups had the same idea at once. That is corroboration rather than replication, so it is labeled but not scored here. Extension is a shared experimental base with added analysis, characterization, or data the original lacked. That is a distinct valuable act, scored by the Extension and Deepening Indicator of Section 5.3.5 rather than credited here.

Both replication directions are weighted by the independence of the replicating or replicated group, through the IRI co-authorship machinery. Concordance, meaning whether the replication confirmed or contradicted the original, is detected by comparing the reported results through the Element 2 entailment machinery.

Data availability, computable now with qualifications. The three-way detection reads the existing corpus through the Element 1 and Element 2 machinery, so this indicator no longer depends on a replication-linkage registry as a precondition and is buildable now. The registry listed in Section 6.6 would improve precision and record outcomes the corpus does not state explicitly.

Separating independent convergence from replication turns on a concurrent-submission window, field-calibrated from each field's publication-lag distribution rather than fixed. That window also accounts for preprint priority dates, since a preprint posted a year before journal publication moves the priority date and can turn an apparent replication into independent-but-earlier work, or the reverse. Three limitations remain. Design extraction and cross-paper matching have a nontrivial error rate that varies by field reporting convention. Concordance detection, meaning whether the replication confirmed the original, is a further entailment step with its own error rate. And the convergence window is noisy at its boundary, where publication lags overlap.

5.3.5 Extension and Deepening Indicator (EDI)

Definition.

Credit for taking a prior experimental base and adding the analysis, characterization, or data collection the original lacked, the research goes deeper act that matures a field and that current metrics undervalue because it is neither a celebrated novel breakthrough nor a pure replication.

Motivation.

Most work that reuses a prior experimental design does not replicate it. It extends it, running substantially the same conditions while collecting data the original missed, characterizing a mechanism the original left unexamined, or pushing the analysis deeper. This is how fields actually mature, yet it falls in the gap between novelty and replication that citation counts and venue prestige both undervalue. This indicator credits the deepening act directly. It credits only the extender, since the original base is credited by its own metrics, and it is constitutive because scoring high requires actually going deeper on a real prior base rather than repackaging it. RFI category alignment: Rigor and Reproducibility, and Foundational Scientific Exploration.

Formulation and data availability.

This indicator shares the replication indicator's corpus-comparison detection. It identifies an extension relationship as a shared experimental base, meaning high design overlap by the Element 1 and Element 2 machinery, combined with added analysis, characterization, or data absent from the original. Credit requires clearing two guards that together block trivial add-ons.

A depth guard scores the substance of what was added relative to the original, using the SObs and SAna machinery on the added material. A single confirmatory figure or a cosmetic reanalysis therefore does not qualify, and only a substantive deepening does. A gap-filling guard then requires the addition to answer a question the original raised or left open, tested by entailment between the original's stated limitations or open questions and the extension's contribution. Adding arbitrary data unrelated to what the original needed therefore does not score. Only work clearing both guards is credited, and only to the extender. This is computable now from the corpus through the shared detection machinery, with the same design-extraction error-rate qualification as the replication indicator, and the registry of Section 6.6 would improve precision without being a precondition.

Failure mode and gaming defense.

Salami extension is the dominant gaming vector here, meaning credit gathered through many minor me-too papers with one added figure. The depth guard blocks it by scoring the substance

of an addition rather than its existence, and the gap-filling guard blocks it by requiring the addition to address a question the original actually left open. A second vector is claiming a shared base the paper did not truly build on, blocked because design overlap is measured from the extracted experimental structure rather than from the paper's framing. Both guards emit their grounding, meaning the original's stated open question and the specific added analysis, so a reviewer can verify each extension credit against the two papers rather than trusting the score.

5.4 Career-Terminal Incentives

5.4.1 Lifetime Foundational Impact Indicator (LFII)

Definition.

A metric computed once, late in or after a career, scoring the whole arc of a researcher's foundational contribution across dimensions no annual metric can observe, ordered by what most advances science over the longest horizon.

Motivation.

The Nobel Prize and National Academy membership are the field's existing career-terminal metrics. They are sparse, slow, politically mediated, and they recognize a few researchers per generation. Every other metric in the evaluation system operates on annual or grant-cycle horizons that cannot register origination, delayed payoff, or generativity, because those resolve over decades.

The absence of a computable career-terminal metric means the vast majority of researchers have nothing to strive for that rewards the long-horizon foundational bets the annual metrics actively punish. This indicator scores, once and late, the accumulated foundational contribution of a career. It gives researchers a long-term target that rewards the risks, discoveries, and field-building the near-term system suppresses. It is almost inherently Goodhart-resistant, since a metric computed once over decades of work cannot be gamed by writing to a scorer within any tractable optimization horizon.

Formulation.

$$LFII(a) = w_1 \cdot Orig(a) + w_2 \cdot Delayed(a) + w_3 \cdot Risk(a) + w_4 \cdot Gen(a) + w_5 \cdot Corr(a) + w_6 \cdot Infra(a)$$

The six components are ordered by how much each advances science over the longest horizon, and equivalently by how strongly each represents work only public funding will support, since the two orderings coincide (the most generative and longest-maturing acts are also the least appropriable by private capital):

146. Origination (Orig): fields, methods, concepts, or tools that did not exist and now do, and that others built on. First because origination is the rarest and most generative scientific act, enabling thousands of downstream careers.

147. Delayed impact realized (Delayed): work that scored low for years or decades and then became foundational, the sleeping-beauty payoff. Second because it is the direct empirical case for patient public funding, visible only to a metric that waited.

148. Risk portfolio (Risk): the lifetime record of high-variance, high-value attempts, crediting the failed bets alongside the successes. Third because it rewards the exploratory behavior itself.

149. Generativity (Gen): the independent scientific lineage launched, trainees who became independent scientists and diverged from the mentor's topic. Fourth because a career's largest long-term multiplier is often the people it produced.

150. Corrective contribution (Corr): work that overturned an error, retired a dead end, or resolved a controversy, sparing the field wasted decades. Fifth because corrective value is large cumulatively and rarely rewarded elsewhere.

151. Infrastructure and commons (Infra): datasets, reagents, tools, cohorts, and methods that became shared infrastructure the whole field depends on. Sixth, and the most direct public-goods case, since shared infrastructure is non-appropriable by its nature.

Implementation.

Each component reuses infrastructure from the diagnostic and incentive arms computed at career-terminal horizons. Origination reuses the field-define indicator. Delayed impact reuses the sleeping-beauty trajectory realized at long horizon. Risk reuses the research risk portfolio. Generativity reuses the trainee trajectory indicator. Correction reuses the assumption-addressing indicator at the paper level and the assumption-correctness indicator (Section 5.4.2) at career horizon. Infrastructure reuses the reuse generation indicator aggregated over a career. The weights are set by the evaluating body (an academy, an award committee, a lifetime-achievement mechanism) rather than canonically, preserving the anti-homogenization principle.

Failure mode and mitigation.

The dominant risk is not gaming but measurement lag and attribution, since career-terminal scoring depends on decades of downstream data and on correctly attributing origination and delayed impact across long citation chains. Mitigation requires the long-horizon citation and provenance infrastructure specified for the sleeping-beauty and field-define indicators, and explicit uncertainty reporting on each component given the attribution difficulty over long time spans.

RFI category alignment.

Foundational Scientific Exploration (primary), spanning Public Impact, Training and Mentorship, and Rigor through its component structure.

5.4.2 Assumption-Correctness Indicator (ACI)

Definition.

A career-terminal measure of whether a researcher's assumption-challenges turned out to be right. It is scored once and late, and it states the time lag openly. It supplies the correctness layer that the Assumption-Addressing Indicator of Section 3.3 deliberately leaves out at publication time.

Motivation.

The Assumption-Addressing Indicator scores how well a paper engages a wrong assumption at publication. It cannot know whether the challenge was correct, since correctness is revealed only by what the field's later evidence shows. Separating the two questions along the horizon axis is the point. The quality of an assumption-challenge is observable now, while its correctness is observable only over years, so one metric measures the act at publication and a second measures the outcome across a career. Placing the second among the career-terminal incentives matches how an academy, an award committee, or a retrospective assessment already reasons, asking not what a researcher published but what of it held up.

Formulation and data availability.

ACI sums, over the assumption-challenges a researcher mounted (identified by their AAI-scored papers), whether each was subsequently vindicated by the field's accumulated evidence, measured as eventual field-concordance, how far later independent work came to agree with the challenge, read from the downstream corpus at long horizons. It is inherently a lagging measure and is scored only at career or post-career timescales, with explicit per-challenge uncertainty given the attribution and timing difficulty. The measure is buildable from the downstream corpus without new infrastructure, though it accrues over the decades its horizon requires.

Failure mode and mitigation.

The honest limit is that this indicator measures eventual field-concordance as the best available proxy for correctness, rather than correctness directly. A field can be slow, can be wrong for a long time, or can vindicate a challenge only partially, and these usually converge with truth but not always and not quickly. The indicator states that proxy relationship rather than claiming to measure correctness itself, and it reports each challenge's outcome with its horizon and uncertainty rather than as a settled verdict. Attribution across long citation chains is the secondary difficulty, mitigated by the same long-horizon provenance infrastructure the sleeping-beauty and field-define indicators require.

RFI category alignment.

Rigor and Reproducibility and Foundational Scientific Exploration, at career-terminal horizon.

6. Implementation Principles

6.1 Auditability and the Open-Reproducibility Precondition

Every model in the framework must be inspectable, and open reproducibility is a hard precondition for deployment rather than an aspiration. No scorer should be deployed whose training data, calibration, and computation cannot be independently reproduced and audited by parties outside the body that built it. This is the specific defense against training-capture, meaning the risk that incumbents calibrate the scorer to ratify the existing hierarchy while appearing objective. Black-box outputs are not acceptable for funding decisions or career evaluation. Implementation requires four things, open-source code, published training-data composition, feature-importance transparency at the per-inference level, and external audit by independent bibliometric experts every 18 months with results published openly. A scorer that cannot meet the reproducibility precondition should not be used, and the incentive-structure reforms, which require no scorer, should be pursued independently.

6.2 Network-Blindness

The content scorers deliberately exclude the social graph that current committee judgment is captured by. They score what is in the paper and, for citation-based indicators, the pattern of who cites whom. They do not take institutional prestige, training lineage, or collaboration-network membership as inputs to quality scoring. That exclusion is how this framework escapes relationship-laundered evaluation, replacing errors structured by who you know with errors that are random with respect to social position. The Independent Recognition Indicator applies the same principle to citation analysis, separating recognition from non-collaborators from network-internal citation. Network-blindness is a design requirement rather than an incidental property, and any modification that reintroduces prestige or lineage as a scoring input would reintroduce the bias this framework corrects.

6.3 Bias Testing

Every indicator must be tested for bias across demographic and structural subgroups including gender, race, ethnicity, institutional tier, geographic region, and career stage. Bias-correction terms in loss functions are required during training. Continuous monitoring after deployment is required with quarterly bias reports published openly. The DDIC meta-composite has the highest bias risk because it aggregates across all components and can amplify component-level biases.

6.4 Temporal Stability and Adversarial Retraining

Adversarial retraining cycles every 12 to 18 months are required, to prevent gaming of the diagnostic-arm indicators. Researchers will learn the indicators' features over time, so those indicators must evolve to keep their validity. The PSS, AAI, SRQI, and VPEI components are most vulnerable, because they score content-level features authors can engineer. Incentive-arm indicators need far less adversarial retraining, because they are constitutive rather than proxy, so gaming them produces the wanted behavior. The career-terminal metric is ungameable within any tractable optimization horizon.

6.5 Complementarity with Peer Review and Weighting Guidance

Diagnostic-arm indicators are advisory inputs to peer review and not autonomous replacements for expert judgment. The incentive-arm indicators, being largely Goodhart-resistant, can receive heavier evaluative weight, and the foundational and career-terminal metrics in particular receive the heaviest weight for the reason given in Section 5.1. Cases where indicator scores and peer judgment diverge are the most informative for both indicator refinement and review-process improvement, and any deployment must preserve peer-review primacy for the diagnostic arm.

Because indicators differ in gaming-resistance, the framework recommends tiered weighting rather than uniform application. The fully constitutive incentive indicators can bear the heaviest weight, meaning reuse generation, replication, corrective contribution, and trainee trajectory, together with the career-terminal metric. Gaming those produces the wanted behavior, and the terminal metric is ungameable within any tractable horizon. Partly constitutive incentive indicators bear moderate weight, meaning the research risk portfolio and foundational investment, with the interim risk-portfolio version receiving zero weight until its initiation mechanism exists. Diagnostic indicators bear advisory weight. VPEI, the one proxy metric near the incentive arm, is capped below the constitutive indicators.

Three residual gaming modes have been named and cannot be closed by measurement. They are the resource-access advantage in co-author composition, reach-gaming in public engagement, and cheap-risk accumulation in the interim risk portfolio. Naming them is deliberate, for three reasons. Evaluators can then discount appropriately, each indicator is used descriptively rather than as a rank, and later methodological work can target the open problem. Even these partly-gameable indicators are retained, because their gamed forms still route through intended behaviors.

6.6 Governance

Indicator development, audit, and update must occur through transparent multi-stakeholder governance including NIH Office of Science Policy, university research offices, scientific societies, bibliometric experts, and patient and public advocacy representatives. Single-contractor opaque development is unacceptable. Funding for indicator development should follow open-science principles with code and training data available to the research community. The registered-report initiation infrastructure required by the Research Risk Portfolio Indicator is a program-design investment that NIH would build and govern directly.

Several indicators depend on data infrastructure that does not yet exist, and NIH is uniquely positioned to build it, maintain it, and drive its adoption. Each piece is non-appropriable public-good infrastructure the private sector will not create, and each raises the measurability of behaviors the RFI identifies for reward.

152. A materials-and-data-reuse registry linking downstream work to the datasets, code, and physical reagents it reused, required by the Reuse Generation Indicator and plausibly built on existing repositories and on data-management-and-sharing-plan reporting.

153. A citation-intent layer, classifying whether a citation corrects, supersedes, or builds on the cited work. This would improve the precision of the Assumption-

Correctness Indicator and related citation-derived signals. Those signals are computable by citing-sentence entailment without the layer, at a higher noise level.

154. A replication-linkage registry, connecting each replication attempt to its target study and recording the outcome. This would improve the precision of the Replication Contribution Indicator. That indicator is now computable from corpus comparison without the registry.

155. A training-lineage registry recording verified mentor-trainee relationships and, ideally, trainee career outcomes in and out of research, required by the Trainee Trajectory Indicator.

156. The registered-report initiation-assessment mechanism required by the Research Risk Portfolio Indicator.

Building these is a program-design investment. The metrics that depend on them reward sharing, correction, replication, mentorship, and risk, which are the specific behaviors currently under-rewarded. The infrastructure investment and the incentive it enables are therefore the same act.

6.7 Anti-Homogenization

Content-scoring indicators create a convergence risk that citation counts do not have. If a scorer rewards a particular reasoning structure, methodological template, or synthesis style, the field will converge on that structure and penalize different cognitive approaches, which is the opposite of what foundational exploration requires. The framework addresses this risk with three provisions.

First, multi-component indicators must report their components separately rather than as a collapsed score, meaning SRQI, PSS, CBI, and LFII, so no single composite ranking forces convergence. Second, evaluating committees select role-appropriate component weights rather than applying a canonical weighting, so different roles reward different profiles. Third, scorers that assess reasoning or synthesis quality, meaning SRQI component eight, MGCI, and CFII, must be calibrated to reward variety in reasoning structure across a field rather than proximity to any single high-scoring exemplar. That calibration is measured by monitoring the diversity of high-scoring reasoning structures over time, and convergence is flagged as an indicator-failure signal requiring recalibration. An indicator that produces increasing homogeneity of high-scoring work across successive cohorts is malfunctioning regardless of its predictive accuracy on historical cases.

6.8 Gaming Defenses and Their Coverage

Gaming defenses here are judged against the metric they would replace, not against perfection. The applicable baseline for the automated layer is h-index and high-impact-venue counting. A defense that reduces a gaming route the incumbent leaves open, or actively rewards, is an improvement even when it does not close that route completely.

The defenses follow an escalation principle. Exhaust the corpus-anchored checks first, then the ecosystem-level checks, and invoke human review last. That ordering follows from this document's own premise. The founding argument is that human judgment is laundered through relationships and biased by prestige, which makes human review the most biased tier rather

than a clean backstop above the automation. Human review is never eliminated, because it remains the honesty and transparency check no automated tier can fully replace. It is pushed as late as possible and bounded when invoked, applied to the residue the corpus and ecosystem checks cannot reach rather than treated as the court of first appeal. The tiers below appear in escalation order, corpus checks, then ecosystem checks, then the bounded human check. Each tier's own bias is named, so the ordering reduces total bias rather than exchanging a visible bias for a hidden one.

The principal defense is the temporal-corpus false-novelty check, which anchors gaming detection outside the paper's own text. Each novelty or gap claim is tested against the corpus published before the submission or preprint date, using the same-year-or-prior retrieval of the SObs novelty dimension and the SGap gap-reality feature. The test asks one question. Was the connection the paper claims as novel already present at threshold density in the prior literature? A claim of novelty for a connection the field was already making is flagged as false novelty. The prior corpus is fixed at submission time, so an author cannot manipulate it. That moves the defense from a property of the paper, which a writer controls, to a fact about the corpus, which a writer does not. The check credits recovery of dormant work through the dormancy exception of the SObs novelty dimension, so resurfacing lost prior art is rewarded rather than penalized. The h-index has no concept of novelty at all and cannot tell a new finding from a restatement of common knowledge, so this check adds a discrimination the incumbent entirely lacks.

Two limits belong here. First, the check verifies that a claimed-novel connection was absent, but it cannot judge whether an absent connection is worth drawing. Novelty that is real but trivial therefore passes at submission, and only the lagging uptake metrics catch it later. Second, the check inherits the coverage of its retrieval substrate. A novelty claim against a thinly-indexed, non-English, or adjacent-field literature the retrieval does not reach can pass falsely, and an informed gamer could aim at that. Multiple retrieval methods and cross-lingual embeddings reduce this without closing it. The check also reports its corpus-coverage confidence for each claim, so novelty confirmed only against a thin corpus is flagged as lower-confidence rather than passed as clean.

Cross-metric consistency is the second defense, and it catches a broad class of gaming, because an isolated gamed signal produces a profile its neighbors contradict. A paper that games novelty but has thin execution shows high claimed novelty against low SExec and SAna. That is a recognizable adversarial signature, and the ensemble scores it as a mismatch. A citation clique that inflates one signal also finds no corroboration from the independence-weighted others. This defense is real, and it closes most single-route gaming. Its limit is structural. It catches only gaming that produces inconsistency. A sophisticated actor who builds coherent mediocrity, where every signal is individually plausible and all of them agree at a modestly inflated level, is not caught, because nothing contradicts. Summing consistent but inflated signals conceals the gaming instead of revealing it, which repeats the false confidence in one aggregate that this framework opposes elsewhere. The claim here is therefore narrow. The ensemble closes gaming that produces inconsistency, not gaming in general.

Two routes survive the per-paper defenses. Both are addressed by widening the scope from the single paper to the temporal ecosystem around it, which is an AI computation over a body of work and its distribution rather than a human judgment call.

The first route is salami-slicing, meaning one study divided into several papers, each individually honest and each novel against the prior corpus. It is invisible at the per-paper level, and invisible to any sum of per-paper scores, because the gaming lives in the division rather than in any one paper. AI author-ecosystem analysis can detect it. That analysis reads the cluster of papers one group published in a time window and looks for the signature of an artificially divided study, meaning shared methods and cohorts, overlapping data, and telling citation timing. This is a corpus computation over an author's output rather than a per-paper score. It is also less biased than a human auditor, who might wave salami through from a prestigious laboratory and flag it from an unknown one. The ecosystem check belongs to the governance and audit function of Section 6.6, because it operates on a person's output, but it runs before any human audit. The h-index actively rewards salami-slicing, since more papers accruing citations raise the index. Even before the ecosystem check, then, per-paper contribution scoring is less vulnerable to salami-slicing than the incumbent that encourages the division.

The second surviving route is coherent multi-metric inflation, the coordinated-mediocrity attack described above, and it is the harder one. Adversarial retraining in Section 6.4 addresses part of it, and ecosystem-covariance detection addresses another part. Real quality co-varies across metrics in specific, spiky ways. Engineered coherence instead sits at a too-uniform level of modest inflation, which is anomalous against the way genuine papers' signals actually co-vary. Building coherent inflation across many independent dimensions costs far more than the volume-and-prestige gaming the h-index rewards. The framework therefore raises the cost of gaming even where it does not eliminate gaming, which is the achievable win against an incumbent that is cheaper to game.

Some fabrication produces a plausible artifact. Invented but internally consistent data, a real but cherry-picked finding, or an experiment repeated until a clean result appears all yield a paper whose signals look honest, because the dishonesty sits in the process that generated the data rather than in the artifact.

A large fraction of this is still reachable by AI ecosystem analysis that no single reader can match. That analysis includes statistical fingerprinting across a paper's reported values, meaning digit-distribution and variance-structure anomalies, cross-paper image-reuse detection, impossible-consistency detection, and comparison of the paper's reported effect sizes against the distribution of real effect sizes in the field. These checks catch more fabrication than citation counting, which never examines the artifact, and more than a human reading one paper alone. They are therefore exhausted before human review rather than after.

Two things remain. One is fabrication that survives ecosystem fingerprinting because the invented data were built to match the field's real distributions. The other is the consequence question the corpus cannot answer, whether a technically novel connection is worth anything. For that residue, and only that residue, the framework invokes human review as an honesty and transparency check. That review is bounded by the same controls applied elsewhere, meaning blinding where feasible, diverse panels, and published reasoning, so the human tier's own relationship and prestige bias is constrained rather than trusted. Human review is never eliminated, because transparency and honesty require a human accountable for the judgment. It is invoked last and on the smallest possible residue, which inverts the usual arrangement, where biased human judgment is the court of first appeal.

The escalation reduces total bias but does not eliminate it. The ecosystem tier has its own bias, and naming it matters, or the reordering exchanges a visible bias for a hidden one.

Widening from the single paper to the ecosystem does not make the judgment objective. The ecosystem itself, meaning the citation network, the retrieval corpus, and the distribution of real effect sizes, is a product of the same prestige and access dynamics this framework criticizes. If the ecosystem skews toward well-funded and well-indexed work, then anomalous against the ecosystem can quietly come to mean unlike what advantaged laboratories produce. That would penalize the neglected and unconventional work this framework credits. It is also the recursive form of the problem of metrics becoming targets, applied to the ecosystem itself.

The ecosystem tier is therefore less relationship-biased than human review, but it has coverage and representation bias along field, language, and resource lines. Two provisions bound it. First, ecosystem anomaly flags are confidence-discounted for under-represented fields and thin-coverage literatures, so a paper flagged as anomalous only against a sparse ecosystem is treated as low-confidence rather than low-quality. Second, an anomaly against the ecosystem is never a quality penalty on its own, because founding and neglected work is inherently anomalous against the distribution it departs from. An ecosystem flag therefore routes to review rather than to a score reduction. That preserves the framework's core commitment, that departure from the field's normal range is a novelty signal and not a deficiency.

6.9 Always-On Grounding Guards, Adjacent-Field Borrowing, and Per-Claim Prior Weighting

Every AI-based scorer in this framework inherits the novelty, fidelity, and rigor gates of Section 3.5.0, and each gate judges a claim against priors. Three failure modes attend that judgment, and they occur in every field rather than only in young ones. Presentation-gaming scores a claim at the strength its framing asserts, meaning bold categorical language, a high-visibility venue, and a clean narrative, whenever the priors needed to check it are not brought to bear. False novelty credits a finding as new whenever retrieval does not reach the prior work that already holds it. Summary-scoring credits a measurement whenever its statistics are accepted without examining the primary data beneath them. These are pathologies of the scoring act rather than of a field's age. They surface wherever the prior for a specific claim is thin, and that happens at a mature field's newest questions and interdisciplinary boundaries as readily as across a young field's whole corpus. Two documented patterns fit this. A categorical tissue-response claim was credited on recording summaries whose underlying waveforms were physiologically implausible. An observation was credited as novel that an adjacent older literature had already established. Both occurred in a field with every surface marker of maturity, and each passed the field's citations and an initial trace for the same reason.

The guards below therefore run unconditionally, on every claim in every field. A gate that fired only when a field-level maturity detector flagged the field would skip the mature-looking fields where those two patterns occurred. What varies is not whether a guard runs, but how much confidence attaches to its output, and that variation is measured per claim rather than per field. A local-prior-density signal estimates how well-settled the relevant priors are for the specific claim. It draws on three things, the replication and citation-linkage density around the claim's own topic, the distance to the nearest prior work retrieval surfaced, and the terminological stability of the claim's central terms. This per-claim measure replaces the field's aggregate

maturity, which averages over the local prior-poverty that lets a specific claim slip through. A claim with thin local priors, or one scored against distant borrowed priors, has its guard outputs confidence-discounted and reaches human scrutiny earlier. Demoting the maturity signal from an activation gate to a per-claim confidence weight removes a single point of failure. A miscalibrated weight degrades a score's confidence gradually, whereas a gate that fails to fire skips the check entirely and silently. The local-prior-density estimate is imperfect and partly gameable, as the field-level version was, but its failure now costs a confidence miscalibration rather than a skipped guard.

Adjacent-field prior borrowing is the first guard, and it is correct retrieval, run always rather than as a rescue for young fields. Each gate draws its priors from the target subfield together with the adjacent fields identified by the Element 2 concept-graph path, never from the target subfield's citation graph alone. A finding is prior art, and a plausibility standard is binding, whether it sits in the subfield or one field over. Three instantiations follow:

157. The novelty gate always retrieves against the adjacent fields as well as the target subfield. A finding already present in an older or adjacent literature is therefore recognized as prior art, even when the subfield's own citation graph never linked to it. This is the check that catches a re-presented observation the subfield believes is new.

158. The fidelity gate always checks a claim's primary data against the nearest settled model of what physiological or physical data of that kind looks like, wherever that model resides. A subfield's failure to consolidate a prior therefore does not let implausible data through. This is the check that flags a non-physiological waveform against mature electrophysiology's settled action-potential morphology.

159. The rigor gate always evaluates a claim against the settled reporting and analysis standards of its parent discipline as well as the subfield. An emerging area is therefore held to the established standard of the discipline it grew from, rather than to its own norms that have not yet settled.

Adjacency is chosen by concept-graph path, and the borrowed-from fields are named in the output, so a human can reject an adjacency that does not apply. An area departing from its parent field may make the parent's priors wrong for a given claim. Borrowed priors are therefore applied as flags that route to scrutiny, never as automatic score reductions. The scorer also reports how much of each score rests on borrowed priors rather than native ones. A score resting heavily on distant borrowed priors receives lower confidence under the per-claim weighting, and it is read as a prompt for human judgment rather than a settled result. This preserves the framework's commitment that genuine novelty, including a real departure from a parent field, is a signal and not a deficiency. It still catches the false novelty and the implausible primary data that a within-subfield check would miss.

The anti-presentation-gaming rule is the second guard, and it is unconditional as a property of the rule, because it is a held-out-feature discipline rather than an escalation. A claim's presentation features are held out and cannot contribute to any score. Those features are the strength and boldness of its categorical language, its abstract salience, its venue prestige, and its narrative framing. This matches the held-out surface features already required for the reasoning steps, so that naming an operation does not substitute for performing it. Each score is set only by the grounding evidence Section 3.5.0 requires, meaning the retrieved neighbor set,

the examined primary-data sample, and the warrant-appropriate rigor criteria. A claim asserted more strongly than its grounding supports is flagged by the claim-alignment check rather than credited for the force of its assertion. This is the computational form of a discipline that a reader with mature priors applies without effort, judging a claim by its evidence and not its framing. Installing it as a hard held-out-feature rule is what keeps the scorer from being gamed the way a credulous citation record is gamed.

Escalation is modulated per claim rather than switched by a field-level flag. When a claim's local priors are thin, or when its score leans on distant borrowed priors, the automated output is confidence-discounted and the human-reviewer check moves earlier. In the mature-field ordering of Section 6.8 that check is the last-resort backstop. It is least trustworthy where the prior for a claim is thinnest, wherever in a field's maturity that claim sits. The honest ceiling where priors are thin is narrower than where they are settled. The framework's advantage over the incumbent citation count narrows to two things. It brings adjacent-field priors the local corpus has not consolidated, and it makes its grounding inspectable, so a human with better priors can see and overrule what the scorer accepted. That is narrower than a claim to independently correct a judgment the field itself has not settled. A thin-prior score is therefore an advisory flag for human scrutiny rather than a signal to act on, and the framework specifies this rather than implying a correction it cannot deliver where priors are not yet formed.

7. Validation Approach

7.1 Retrospective Validation

Retrospective validation against historical cases is the primary acceptance test, and it must demonstrate discrimination for positive and negative controls alike. Three positive controls apply. The framework must score Katalin Karikó's 2005 paper, Karikó and Weissman in Immunity 23:165-175, above the median of contemporaneous papers in immunology and molecular biology. It must score Brock and Freeze 1969, in J Bacteriol 98:289-297, above the median of contemporaneous microbiology papers. And it must score Maynard and colleagues 1997, in Electroencephalogr Clin Neurophysiol 102:228-239, above the median of contemporaneous neural-engineering papers.

As negative controls, the framework must score volume-first archetypes below field median on the paper-substance and scientific-reasoning-quality indicators. Those archetypes are papers running single-test analyses on reused public datasets with no primary data collection, which accumulate citation counts through publication volume rather than substantive contribution. A framework that scores foundational work high but fails to score volume-first work low is not discriminating. It is tracking eventual citation, and it fails the acceptance test.

The graceful-failure gate is a hard deployment condition rather than an aspiration. An indicator does not deploy unless its errors are demonstrably uncorrelated with systematic suppression of foundational, slow, corrective, or exploratory work, which is the specific failure mode of the incumbent metrics it would replace.

The bar is comparative. The indicator must predict genuine contribution better than the h-index does, and its residual errors must be closer to random with respect to research type than the incumbent's structured errors are. An indicator that is more accurate on average, but that concentrates its errors on the same foundational work the h-index penalizes, has not cleared the gate, because it reproduces the harm under a new name. Deployment is conditional on passing both the discrimination test above and this graceful-failure test, evaluated during the forward-validation pilot before any use in funding or promotion decisions.

One limitation is intrinsic to retrospective validation and applies to every historical control. The positive controls all predate CRediT and much of the machine-readable metadata that contemporary application relies on. Any co-author-composition score computed on them therefore runs on the weaker pre-CRediT participation evidence described in Section 3.10, and the older cases have more attribution uncertainty than a present-day paper would. This does not defeat the retrospective test, since the content-scoring and trajectory indicators read the paper text and citation structure that do exist for historical work. The retrospective controls exercise the team-composition indicator less dependably than the reasoning and impact indicators, so retrospective validation results should be read with that unevenness stated rather than treating every indicator as equally testable on decades-old papers.

7.2 Forward Validation

Forward validation requires a pilot study at three to five institutions across diverse fields, scoring in parallel against traditional metrics for 24 months before any use in funding decisions. Pilot sites should include at least one institution outside the United States, to test geographic

generalization. The incentive-arm indicators need a longer validation horizon than the diagnostic arm, because their behavioral effects appear only over multiple grant cycles, meaning whether researchers actually attempt more exploratory work.

7.3 External Audit

External audit by independent bibliometric experts is required every 18 months, with results published openly. The audit panel should include representatives from DORA, the San Francisco Declaration on Research Assessment, and from the Center for Open Science. It should also include at least two non-US institutions, to test international generalizability.

7.4 Thin-Prior Stress Test

The guards of Section 6.9 require their own validation, because the regime where the framework degrades is thin local priors, and those arise at a mature field's newest questions and interdisciplinary boundaries as well as across a young field. The controls of Section 7.1 are drawn from mature or historically-settled cases that do not probe that regime. This stress test therefore assembles, for several fields spanning a maturity gradient, a set that includes each field's most-cited early paper whose central claim was later qualified or overturned. It then measures whether the framework, with adjacent-field borrowing, flags at publication time the overclaim that the field's citation record missed.

Two design conditions are essential. The trace must run unprimed, meaning the evaluator is not told which papers are the suspected negative controls, since a primed evaluator who knows where to look overstates the framework's real detection, and that specific overstatement has already been observed. And success is measured against time-to-correction, meaning whether the framework reports the flag years before the field's own slower self-correction did, rather than against a single right-or-wrong score. The framework's honest value in a young field is shortening the correction lag and making the grounding inspectable, not independently settling a judgment the field itself had not yet settled.

7.5 Consequence-Calibration Separation Test

The consequence and calibration axes of Section 3.12 require their own validation, because the failure they address is substitution of calibration for magnitude, and the retrospective controls of Section 7.1 do not isolate it. This test assembles two sets matched on calibration, one of rigorous, well-calibrated papers reporting small increments and one of rigorous, well-calibrated papers that redirected their field, both drawn from settled history where magnitude is known. Acceptance requires similar scores on the calibration axis across the two sets and separated scores on the consequence axis, since equal calibration with divergent consequence is the discrimination the two-axis split is specified to produce. This test must run unprimed, for the same reason the thin-prior test must, because a calibration-weighted evaluator who knows which papers are minor conflates the axes and overstates the separation, a conflation observed during this framework's own development when honest minor papers were scored as strong contributions.

8. Discussion and Closing

The framework proposes two arms, answering the two jobs the RFI names. The diagnostic arm has fourteen top-level indicators that detect quality the existing metrics miss. They score content rather than metadata. They identify the shape of a trajectory rather than an aggregate count. They score scientific and engineering reasoning directly from paper text, through the SRQI and EMI meta-indices and their step-level sub-indices, rather than inferring quality from venue prestige. And they invert the citation asymmetry by rewarding reading breadth as well as being read. The incentive arm rewards wanted behaviors constitutively, so that gaming a metric produces the intended behavior. It leads with foundational and exploration incentives, because those behaviors have the longest payoff horizon and the weakest existing constituency.

The central argument is that evaluation should push back against the forces already acting on the research enterprise, rather than adding to them. Study sections and public sentiment both reward near-term translation. Weighting all work equally therefore under-invests in the long-horizon foundational research that only public funding will support, and so the incentive arm is weighted toward foundational and career-terminal metrics. The career-terminal Lifetime Foundational Impact Indicator gives researchers a long-horizon target. It rewards origination, delayed payoff, risk-taking, generativity, correction, and infrastructure, the acts that advance science most and that no annual metric measures. Public understanding of why foundational work matters is what protects its funding across political cycles. The public-engagement indicator therefore works as a meta-incentive, protecting the survival of everything else.

Each indicator has documented failure modes and required mitigations. The diagnostic arm will be valid only if adversarial retraining, bias testing, auditability, and anti-homogenization provisions are implemented. Without them it will fail as the h-index failed, by becoming a proxy that displaces the thing it was built to measure. The incentive arm is more robust, because its metrics are constitutive rather than proxy. It still has its own risks, meaning cheap-risk accumulation and attribution difficulty at long horizons, and its specifications address both directly. This submission does not argue that AI-enabled indicators solve the metric problem. It argues that they can be specified to do two things, to detect the failure modes existing metrics produce, and to reward constitutively the long-horizon foundational work that the public-funding mission supports.

What this framework does not yet establish should be stated as fully as what it does, and measured against the same baseline the framework competes with. Three limits are real. First, the gaming defenses of Section 6.8 close false novelty and inconsistency-producing gaming, but they leave salami-slicing to author-level audit and coherent multi-metric inflation to adversarial retraining and cost, and they defer fabrication of a plausible artifact to replication and human review. Second, validation to date shows that the procedures resolve and discriminate on rigorous work across contribution types, including negative results. The discrimination claim against low-quality and gamed work is a different matter. The retrospective negative controls of Section 7.1 test that claim rather than settle it, and the deficiency floors and the aggregate's neutrality toward rigorous nulls still need negative-control calibration in the validation plan. Third, the corpus-relational machinery inherits the coverage of its retrieval substrate. Fields and literatures that are thinly indexed, non-English, or pre-digital are therefore scored with lower confidence, and the framework reports that rather than hiding it.

A fourth limit is the most consequential. The novelty, fidelity, and rigor gates judge a claim against field priors. A scorer calibrated on a forming field's thin corpus therefore inherits what that field cannot yet see, and it can reproduce the field's credulity. That is the failure that lets a presentation-gamed claim pass the gates just as it passed the citation record. Section 6.9 addresses this three ways, with always-on adjacent-field prior borrowing, a held-out-feature rule against presentation-gaming, and a per-claim prior-density weighting that discounts and escalates scores where the local prior for a specific claim is thin. Those guards run always, rather than switching on at a field-level maturity flag. The honest ceiling in a forming field is narrow. The framework can borrow priors the field has not formed, and it can make its grounding inspectable. It cannot independently correct the field's judgment. Young-field scores are therefore advisory flags for human scrutiny rather than signals to act on.

Measured against the applicable baseline, each of these limits is a reduction of a problem the incumbent handles worse. The h-index has no concept of novelty, so it cannot detect false novelty at all. It actively rewards salami-slicing through paper counts. It rises identically for fabricated and for genuine citations. And it inherits the coverage biases of the citation graph, which are more severe than those of the retrieval corpora. The honest claim is not that this framework is ungameable or complete. On each named limit it moves the failure in the right direction relative to the metric it would replace, exhausting the corpus and ecosystem checks first and reserving human review for the irreducible residue. Human review is never eliminated, because honesty and transparency require an accountable human judgment. It is invoked last and bounded, which follows from this framework's own premise that human judgment is the most relationship-biased tier rather than the clean backstop above the automation.

A fifth limit is the one a domain expert raises first, and it runs against the applicable baseline more directly than the others. Novelty is measured as the presence of an increment against field priors. Calibration and rigor are measured densely. Contribution magnitude, meaning whether an increment changes what a field asks, does, or believes, is measured through only four forward-looking indicators, and only as a prediction with wide uncertainty. A paper can therefore pass every novelty check and score high on calibration while reporting a small increment. Without the consequence-axis separation of Section 3.12, the dense calibration signal is then treated as magnitude, and honest minor work receives a strong-contribution score.

One comparison is unfavorable and belongs in the record. The h-index is itself a lagging magnitude proxy, since citations accrue to consequential work over time. This framework therefore exchanges a lagging but real magnitude signal for calibration and honesty, plus a forward magnitude prediction that no validation at publication time can confirm. That exchange depends on the use case rather than being uniformly favorable. It is most favorable for forward-looking decisions on early-career or new-direction work, where the lagging proxy has not resolved and honest reporting supplies more information. It is least favorable for retrospective assessment of a long senior record, where accrued citations are a competitive magnitude signal. Measured against the baseline, this framework does not outperform citation counts at ranking realized magnitude. Its advantage lies in separating magnitude from honesty and in withholding credit for network position. The consequence axis is therefore where that advantage is narrowest and its uncertainty is widest, and the two-axis reporting with the Section 7.5 test makes that visible rather than hiding it.

Author disclosure. The author is an established, well-cited, network-embedded senior investigator. That is the profile relationship-biased evaluation advantages, and it is the profile this framework would constrain. The recommendation therefore runs against the author's own positional interest, since network-blind content scoring and independent-recognition weighting reduce the advantages a well-connected senior investigator currently holds. This critique is offered from inside a position that benefits from the current system, and it proposes to dismantle the arrangements that produce that benefit, which is the opposite of a self-interested recommendation.

References

Adams, J. (2013). Collaborations: The fourth age of research. Nature, 497(7451), 557-560.

AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. Nature Communications, 9(1), 5163.

Brock, T. D., & Freeze, H. (1969). Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile. Journal of Bacteriology, 98(1), 289-297.

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. Management Science, 63(3), 791-817.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 102(46), 16569-16572.

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. Science, 322(5905), 1259-1262.

Karikó, K., Buckstein, M., Ni, H., & Weissman, D. (2005). Suppression of RNA recognition by Toll-like receptors. Immunity, 23(2), 165-175.

Maynard, E. M., Nordhausen, C. T., & Normann, R. A. (1997). The Utah intracortical electrode array: A recording structure for potential brain-computer interfaces. Electroencephalography and Clinical Neurophysiology, 102(3), 228-239. https://doi.org/10.1016/S0013-4694(96)95176-0

Merton, R. K. (1968). The Matthew effect in science. Science, 159(3810), 56-63.

National Institutes of Health (2026). Request for Information on Measuring and Rewarding Scientific Impact. NIH Notice NOT-OD-26-087.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. Nature, 566(7744), 378-382.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316(5827), 1036-1039.

Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. Proceedings of the National Academy of Sciences, 119(36), e2200841119. https://doi.org/10.1073/pnas.2200841119

266. University of North Carolina at Chapel Hill, Renaissance Computing Institute, HEAL Data Stewardship Group

I am responding: On behalf of an organization

Name: Julie Horvath

Name of Organization: University of North Carolina at Chapel Hill, Renaissance Computing Institute, HEAL Data Stewardship Group

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Core recommendation: Recognize rigor and reproducibility not only through the quality of study design and analysis, but also through the conditions that allow others to understand, evaluate, reproduce, validate, and extend research findings. Our review of barriers to data sharing and reuse highlighted a broader lesson about scientific impact: rigorous science does not end when a study is completed or a manuscript is published. Reusing data requires understanding the explicit documentation supplied with a dataset and the implicit assumptions embedded in the original study design, which are rarely transferred with data. Long-term scientific value increasingly depends on whether findings can be independently interpreted and tested by others.

Potential indicators: Potential indicators should focus on whether the research itself was designed, conducted, analyzed, and reported in ways that support credible interpretation and independent verification of findings. Indicators could include use of appropriate study design, clearly specified research questions or hypotheses, transparent sampling and recruitment methods, valid and reliable measures, prespecified analytic plans where appropriate, robust statistical methods, sensitivity or validation analyses, clear reporting of limitations, and sufficient methodological transparency to allow others to assess whether the conclusions are supported by the evidence. Where data or code are relevant to reproducing findings, NIH could consider whether the analytic workflow, assumptions, and key methodological decisions are documented well enough to support independent evaluation or replication.

Implementation approaches and incentives: NIH could incentivize reproducibility-enabling work by recognizing documentation, curation, harmonization, stewardship, and preservation of contextual knowledge as legitimate scientific contributions. These activities could be reflected in progress reporting, biosketch contributions, funding opportunity language, review criteria, supplements for replication or validation work, and recognition of research products beyond traditional publications.

Risks and feasibility considerations: A dataset, protocol, or analytic workflow may be technically available but still insufficiently documented to support meaningful assessment or reproduction. Metrics should account for variation across career stages, disciplines, data types, and institutional resources, and should

avoid penalizing investigators working with sensitive data, emerging methods, smaller teams, or resource-constrained institutions.

2. Data, Software, and Model Sharing: Core recommendation: Distinguish among availability, accessibility, usability, and demonstrated reuse when assessing the impact of shared data, software, models, and related research products. Impact is determined not simply by whether research products are shared, but by whether they can be found, understood, governed appropriately, and used by others to advance new questions.

Potential indicators: Potential indicators should focus on the quality, accessibility, governance, and sustainability of shared research products after they are made available. For data, software, and models, indicators could include use of persistent identifiers, appropriate repository placement, clear access conditions, licensing or use terms, versioning, machine-readable metadata, data dictionaries or codebooks, governance information, common data elements or interoperable formats, and documentation that enables a user to find, understand, request, run, or apply the product appropriately. For software and models, indicators could also include version control, example workflows, test data, model cards or equivalent documentation, clarity about limitations and intended uses, and concrete maintenance and sustainability plans that describe how dependencies, documentation, updates, preservation, and user support will be managed over time. NIH could also consider evidence of uptake or reuse by investigators outside the original team, including citations, secondary analyses, integration into new studies, or use in policy, clinical, public health, or methodological work.

Implementation approaches and incentives: NIH could incentivize high-quality sharing by rewarding reusable research products as scientific outputs in applications, progress reports, and review processes. NIH could also support shared infrastructure, repository coordination, metadata standards, common data elements, governance templates, and resources for curation and long-term stewardship. Incentives should emphasize the quality and utility of shared products rather than volume alone.

Risks and feasibility considerations: NIH should avoid incentives that reward posting large numbers of products without assessing whether they are usable, interpretable, or responsibly governed. Care should be taken so that metrics do not encourage premature or inappropriate sharing, especially where privacy, consent, tribal sovereignty, community expectations, intellectual property, or legal restrictions require tailored governance. It is important to consider infrastructure disparities across institutions and how data stewardship expectations might disproportionately burden early-career investigators, smaller teams, or less-resourced organizations.

3. Training and Mentorship: Core recommendation: Recognize training and mentorship that prepare researchers to contribute to rigorous, collaborative, data-intensive, and reusable science. Future scientific impact will depend not only on disciplinary expertise, but also on researchers' ability to operate within complex research ecosystems that require documentation, stewardship, governance, interdisciplinary communication, and responsible knowledge sharing. Literature on data reuse barriers identify training gaps across disciplines around data reuse, including uncertainty about which expertise should be represented on reuse teams and how mentors should prepare trainees for rapidly changing tools such as AI-enabled analysis.

Potential indicators: Potential indicators include mentorship and training activities that build capacity in FAIR data practices, reproducible workflows, transparent documentation, interdisciplinary collaboration, research governance, ethical secondary use of data, and community-engaged stewardship. NIH could also recognize trainee contributions to shared data resources, software, standards, documentation, governance processes, or other research infrastructure that supports cumulative science.

Implementation approaches and incentives: Mentorship plans and training programs to include explicit competencies related to reproducibility, data stewardship, collaborative science, and responsible sharing should be encouraged. These competencies could be incorporated into training grant expectations, career development awards, mentorship plans, progress reports, or recognition of trainee products beyond first-author publications.

Risks and feasibility considerations: Mentorship metrics that privilege large, well-resourced laboratories or rely only on easily counted outcomes such as publication counts, grant awards, or placement in traditional academic roles should be avoided. Measures should account for diverse career paths, institutional contexts, team-based contributions, and the time required to mentor trainees in high-quality stewardship and collaborative research practices.

4. Collaboration: Core recommendation: Recognize collaboration-enabling work as a form of scientific impact when it creates durable infrastructure, standards, governance pathways, shared vocabularies, or trusted networks that allow research findings to extend beyond a single project. Our literature review of data reuse barriers suggests that scientific impact is increasingly generated through networks, not only through individual studies.

Potential indicators: Potential indicators include development or adoption of shared standards, harmonization across studies or datasets, cross-institutional governance models, community coordination activities, shared repositories, common vocabularies, consensus processes, and interdisciplinary tools or documentation resources. NIH could also consider evidence that collaborative infrastructure enabled future studies, cross-study analyses, broader participation, or more generalizable findings.

Implementation approaches and incentives: NIH could encourage applicants and awardees to describe or document specific collaboration-enabling contributions, including standards development, governance alignment, stakeholder coordination, knowledge translation across disciplines, and maintenance of shared resources. These contributions could be recognized in biosketches, progress reports, team science review criteria, center or consortium evaluations, and supplemental funding opportunities.

Risks and feasibility considerations: Measures equating collaboration with the number of co-authors, participating institutions, or number of events/meetings may not reflect the quality, durability, or scientific utility of collaborative contributions. Metrics should account for invisible coordination labor, community-building, governance work, and the contributions of investigators and staff whose roles may not be captured by traditional authorship or leadership indicators.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: Core recommendation: Assess public impact through evidence that research knowledge remains available, understandable, trustworthy, and useful over time. Public value is not determined solely by publication of new findings. Value also depends on whether scientific outputs can inform future research, clinical guidance, public health practice, policy, community decision-making, and responsible innovation.

Potential indicators: Potential indicators include evidence that research outputs informed clinical guidance, public health practice, policy development, future research priorities, methodological standards, community decision-making, or subsequent funded research. NIH could also recognize reuse of research products to address high-burden health conditions, engagement with patients, communities, practitioners, or other end users, and governance practices that improve trust and responsible use.

Implementation approaches and incentives: NIH could encourage applicants and awardees to describe plausible pathways from research outputs to public benefit, including how data, tools, models, standards, or findings may be used by researchers, clinicians, public health practitioners, policymakers, communities, or other stakeholders. NIH could also recognize public-impact contributions that occur over longer time horizons, including foundational infrastructure, replication, stewardship, and community-engaged work that builds trust and enables future advances.

Risks and feasibility considerations: Public impact should not be interpreted only as near-term commercialization, immediate clinical translation, or highly visible media attention. Foundational research, infrastructure-building, replication, data stewardship, and responsible governance may create substantial public value over longer periods. Metrics should also avoid encouraging overstatement of public benefit or disadvantaging fields where impact emerges gradually, through cumulative evidence, or through improved trust and responsible use of sensitive research resources.

8. Other Areas: N/A

267. University of Arizona

I am responding: On behalf of an organization

Name: Kim Patten

Name of Organization: University of Arizona

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings.

NIH should reinforce principles that reward methodological rigor, transparency, and high-quality reporting. Evaluation systems should place greater emphasis on the robustness of study design, analytical methods, and adherence to rigorous scientific standards, rather than only rewarding positive or novel results. This might use standard metrics like invention disclosures, patents filed/issued, licenses executed (more critical than invention disclosures or patents issued), startups launched, startup survival rates, startup fund raising activity, products on market, etc. Evaluation systems need to include recognition and metrics for open scholarship, such as open access repositories, open access publications, and open-source code.

Replication and reproducibility studies should likewise be incentivized and rewarded, as they play a critical role in validating important findings and strengthening confidence in biomedical research, yet are often undervalued relative to studies that emphasize novelty alone. NIH could have calls for proposals that are specific to replication and reproducibility in specific areas of interest as a means to confirm and certify innovative and promising findings faster in order to get them into the public with strong reliability and confidence for the general public. This will support additional commercialization and translation from bench to market.

Additionally, NIH may offer more expedited ways for advancing clinical trials that may demonstrate promise, efficiency, or safety without the large scale generally required of traditional clinical trials. Further, leveraging existing infrastructure like the NCI's National Clinical Trials Network could expand replication.

2. Data, Software, and Model Sharing: Sharing scientific data and tools accelerates biomedical research discovery, in part, by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies. NIH seeks to incentivize and reward these activities to maximize the value of research investments, catalyze collaborations, and accelerate scientific impact.

The sharing of research protocols, analytical workflows, open-source code, open access data, and open access publications should be required to facilitate verification, replication, and data reuse. In addition, the publication of well-executed studies with null/negative findings should be recognized as an essential contribution to the scientific record, helping to reduce publication bias and improve the accuracy of the evidence base. The systematic integration of open and community-engaged scholarship expectations will prompt faculty to accurately represent the work that they are already doing. It will enhance the perception that faculty receive "credit" for all the ways in which they share their work within and beyond the university.

Moreover, we anticipate that faculty will be more likely to participate in open and community-engaged scholarship when they perceive that it is valued and recognized in evaluation by NIH in grant reviews in explicit and tangible ways, such as including a section on open access scholarship to document not only ways that previous research has been openly shared through open source code, open access data, and open access publications, but to also indicate plans to openly share outcomes from the proposed research project in all areas of research protocols, analytical workflows, open source code, open access databases, and open access publications.

NIH needs to provide training to peer reviewers to increase awareness and knowledge of the importance of open access scholarship broadly, and recommendations about how to evaluate submissions that include open access research protocols, analytical workflows, open source code, open access data and open access publications. In addition, NIH should educate peer reviewers on the importance and evaluation of public impact metrics to demonstrate the value to the public in terms of licenses (including patents) and policy briefs. Patents, licenses, and policy briefs are now trackable through reliable databases, and metrics include repeated use of them through a "citation" type of tracking and have peer evaluation aspects. These include tracking of patents in Google Scholar and tracking policy briefs by Sage at Sage Policy Profiles. These can and should be considered as seriously as peer reviewed publications when it comes to evaluating the impact of previous scientific research and the quality of the P.I. and their team of Senior Key Personnel.

While it is essential that faculty members retain the freedom to choose how and in what venues they will disseminate their research, they will also attain greater awareness of the benefits of open scholarship for not only public access to their work, but also the potential for higher citation rates and visibility in their field. Research published between 2011 and 2021 suggests open scholarship finds more readers and generates more citations, as described below:

- 2.6x-4.4x: Increase in page views when articles are open. (Wang, et al., 2015)
- 115%: Increase in downloads when articles are open. (Davis, 2011)
- 8%, 18%, 1.24x: Increases in citations to open articles. (McCabe, Snyder, 2014), (Piwowar, et al., 2018), (Archambault, et al., 2013)
- 10x: Increase in online mentions when books are open. (Lucraft, 2017)
- 50%: Increase in citations to open books. (Lucraft, 2017)
- 2.7x: Increase in downloads when books are open. (Neylon, et al., 2021)

Such awareness of the impact of open scholarship can help address the challenge of traditional mindsets that link the quality and impact of one's work solely to the historical prestige of a journal title and/or publisher. This would also help reassure junior scholars that the senior scholars who are reviewing their work have been familiarized with and support this commitment to open scholarship.

NIH could include subsections in proposals that require the description of institutional resources that explicitly call out resources for open access repositories on campus and open databases. This may also request information about policy and faculty review processes that explicitly value open scholarship and resources to support innovation and entrepreneurship.

To assist in data generation, NIH should prioritize investment in large data collaborative initiatives, creation of databases for research methodologies and protocols and open-source code. By investing in scientific leaders to create and develop these open scholarship sources who are well-prepared in areas of biomedical research, there will be enhanced credibility and utilization. They can develop new forms of open scholarship with tracking ability to indicate metrics of use and evaluation. This type of approach will reward institutional and investigator contribution by reinvesting back on its contributors to promote data utilization, offer computing resources, technical support, for rigor and reproducibility.

3. Training and Mentorship: NIH recognizes that advancing its mission requires promoting the growth and stability of the biomedical research workforce. While NIH supports training and protects time through existing policies and programs, additional indicators are needed to assess and quantify the impact of these efforts.

Training on the value of open/community-engaged scholarship for the researcher/scholar, for the scientific community and for the public good needs to be a priority for preparing our graduate students, post docs, early career professionals, faculty mentors, and review committees. Integration and alignment of these systems through NIH training grants, and expectations for proposals will increase faculty capacity to assess and adopt various open and community-engaged scholarship practices.

In order for NIH to advance team science, there will need to be an accompanying shift acknowledging that mentoring in team science requires unique skills, and best practices than the traditional 'one-one' mentoring in a hierarchical structure. Translational and team science requires collaborators not only with disciplinary expertise but also a leadership team that represents the diversity of necessary experience, skills, access and experience, coupled with the skill at leading team science and conducting convergence research. Given that not all faculty receive training best practices in mentoring, NIH could require training in team science mentoring for best practices among all funded proposals that work with junior scholars, post-docs and graduate students. The agency must consider how a proposed project will address team and bi-directional mentoring approaches and consider how researchers at all stages of their careers develop new team leadership and mentoring skills that will be required for successful project outcomes and outputs.

Outstanding examples of collaboration among Universities and NIH to support workforce development are the NHLBI PRIME program which focuses on promoting researcher training for early career researchers. Programs such as these may be replicated across institutes and centers supporting a broader pool of talent. Given the limited funding available under F, T, K mechanisms for physician

scientists, it has become increasingly difficult to provide time protection for training, institutions with limited resources are unable to invest in this essential phenotype of researcher. Similarly, salary caps for mentors, and lack of funding for mentors also limits the availability of physician scientists to truly dedicate time, emotional and technical effort towards training the upcoming generations.

In their 2015 report, Enhancing the Effectiveness of Team Science , the National Research Council noted a number of strategies that have been shown to be effective in supporting team science. Researchers submitting proposals to NIH could be asked to provide a narrative section in the proposal and evidence that they use these strategies when managing their research team. Alternatively, NIH could require a narrative paragraph incorporated into their biographical sketch to demonstrate they have experience leading large collaborative teams to successful outcomes. NIH could require evidence of training in best practices with team science mentoring in order for the proposal to be competitive because it demonstrates the likelihood that the team will be successful in achieving outcomes without becoming worn down with internal conflicts or lack of effective team based strategies to move the science forward more quickly with better outcomes for all involved, as well as preparing a broader base of future scientists who are prepared to work in large teams.

The key strategies noted in the report include:
· ensuring that the team has task-relevant diversity
· using task-analytic methods
· getting training in team science approaches
· taking time to developed shared knowledge (particularly important for geographically diverse teams)
· the leader should delegate responsibility to smaller groups
· have a collaboration plan

These are simple and straightforward tangible metrics that can be required or at least evaluated in the evaluations of proposals.

4. Collaboration: As biomedical research becomes increasingly interdisciplinary and convergent, collaboration is growing in both scale and complexity. NIH seeks new approaches for assessing and recognizing contributions to team-based science.

To incentivize team and translational science, NIH must examine to what extent the calls for proposals are written in ways that require and incentivize collaboration, both from an inter- and trans-disciplinary (e.g. including non-academic partners) aspect but also looking at institutional types that would generate even more convergence. This can be considered by movement away from the one P.I. model to more models that support and encourage multiple P.I.s and team led-science. This recommendation is consistent with the team science work that has emerged from the NCI, in that grand challenges require multiple disciplines to solve. This can be motivated through proposal requirements, revised budget categories and allocations, and recognition of different types of team leaders. Teams would no longer be led by one or two disciplinary experts but rather would be composed of co-investigators/co-leaders from the various entities, including community-based organizations, industry partners, and bioinformatics partners that are required to move basic research to translation. Team science has been shown to enhance leadership, collaboration structures, and cross-disciplinary research productivity. This also means to address any other existing barriers to collaboration and sharing of funds and data across

institutions.

NIH could incentivize and recognize the value of co-produced research, driven by identified community needs. This would require projects that authentically and deeply engage interest holders across a diverse spectrum. NIH creation of funding opportunities specifically to create training programs to lead and participate in team based collaborative approaches would advance work in this area of high quality collaboration for scientific success. Such capacity building programs could provide toolkits, promote community-based practices, and provide training frameworks for faculty and researchers. Another way of promoting the use of team science, efforts like University of Texas Medical Branch Team Science training microcredentials and/or badges offer introduction to essential principles that are valued across some but not all institutions.

NIH could start taking into account inter and transdisciplinary features of research as they evaluate proposals. Similar to discussions of multiple PI and their roles, a discussion of specific contributions across disciplines will highlight and emphasize the value of different disciplines working together to solve some of our greatest biomedical challenges. These could be linked across the NIH community to stimulate more team-science and translational project proposals and to ensure investigators are equipped with all relevant skills to lead significant research projects. Concurrently, review functions would have to shift away from strictly biomedical disciplinary experts but also include those with expertise in non-biomedical specific areas required for collaboration (e.g. bioinformatics, social science, evaluation, community engagement.)

To stimulate more translational research projects, NIH could offer phased funding approaches where successful ‘phase 1’ funding comes with opportunities for follow on funding. Not every team with an innovative idea is immediately networked or capable of forming the cross-disciplinary, cross-community, and cross industry team needed to be fully successful in community engaged, translational, and public impact research. It takes time to find the right collaborators and co-develop research questions and methods in true collaboration. A successful phase 1 project could lead to a non-competitive ‘phase 2’ award to fund the implementation of the project, develop outputs that and are suitable for all partner’s needs, and assess outcomes.

5. Entrepreneurship and Translation: To advance the translation of discoveries into products for people, NIH fosters the maturation of discoveries from bench to bedside and promotes innovation across sectors. More can be done to incentivize, assess, and integrate the impact of innovation and entrepreneurial efforts as critical to the NIH mission.

NIH can strengthen its impact on innovation and entrepreneurship by integrating translational and commercialization outcomes into funding mechanisms, metrics, and institutional incentives. Some specific ideas might include:
• Create explicit budget lines within major research funding opportunities for Technology Transfer Offices (TTO) operations and staffing. Provide a small budget allocation of 2-5% of institutional overhead costs to support TTO personnel. A strong TTO infrastructure directly correlates with successful technology transfer.
• Establish pilot project funding mechanisms specifically for feasibility and proof-of-concept studies. Again, providing a small budget allocation of 5-10% for competitive pilot projects focused on early-stage

commercialization validation.
• Allow researchers to request some percentage (maybe up to 10%) for startup-related expenses, including things like business plan development and market research; initial intellectual property filing and prosecution; regulatory pathway assessment and FDA pre-submission meetings; early-stage clinical trial planning and preparation; etc.
• NIH could explicitly require funded institutions recognize excellence in commercialization in their P&T processes. Not something so simple as “patents” but there are several other ideas including products on market, “impact score” for patents, etc.

However, NIH should avoid creating competition between NIH and institutions. Intellectual property protection is essential for entrepreneurship. An important balance may be reached between facilitating replication, expansion, and leveraging of findings while prioritizing gains by the original thinkers and developers. NIH mandates for data sharing may be balanced with support for product development, etc. NIH may have partnership with private industry to facilitate this collaboration which should remain inclusive and protective of researchers and their institutional interest.

6. Foundational Scientific Exploration: Basic research and high-risk, high-reward initiatives represent foundational scientific exploration with an inherent risk of failure but the potential for paradigm-shifting breakthroughs. Further embedding these approaches within the academic enterprise is essential to ensuring continued support for bold and innovative research directions.
NIH's effort to develop metrics that better reflect the rigor, collaboration, and long-term value of modern biomedical research is well-timed and necessary. This means that we need to consider how science outcomes have moved beyond peer reviewed publications and include open scholarship, innovation and entrepreneurship as well as community-oriented reporting and sharing of research projects and outcomes. However, as the agency develops new indicators, it should guard against placing disproportionate emphasis on only translational outcomes.

Translation is an essential part of the biomedical enterprise, but many of the most transformative advances originated from curiosity-driven basic research with no obvious clinical application at the time. For example, CRISPR emerged from studies of bacterial immune systems. GLP-1 therapeutics grew from fundamental investigations of hormone biology. Taq polymerase, the foundation of PCR, arose from research on microorganisms living in extreme environments. In each instance, the breakthrough was not initially a therapy or product, but a new understanding of biology that later enabled major innovations.

Metrics that heavily reward near-term translational impact, commercialization, or clinical implementation may unintentionally favor incremental advances over fundamental discoveries. This creates a risk that researchers, particularly early-career investigators building their portfolios, will pursue projects with predictable outcomes rather than tackle the deeper biological questions that often generate the greatest long-term benefits. This dynamic may be especially acute at institutions without robust industry partnerships or translational infrastructure.

A central challenge in evaluating scientific contributions is that their future value is not always predictable at the time of discovery. Many foundational discoveries reveal their significance only years or decades later. Basic research therefore creates not only knowledge, but opportunity; the possibility of future discoveries, technologies, and therapies that cannot yet be imagined. For this reason, NIH should

explicitly recognize and reward contributions that expand the scientific foundation upon which future advances depend. Measurable indicators could include:
• Tool and method creation: Novel methods, data visualizations, tools, datasets, and model systems adopted by the broader community, as measured by protocol citations, plasmid deposits, software downloads, or reagent requests.
• Resource generation: Open scholarship for shared resources that enable research across multiple groups, as measured by data reuse metrics, secondary publications arising from deposited datasets, and cross-institutional collaborations enabled.
• Field-opening discovery: Discovery-driven research that opens new lines of investigation, as measured by time-lagged citation impact and the initiation of new funding programs or research areas.

These measures should complement, not replace, assessments of translational success. Alternatively, NIH may further differentiate funding types, e.g., the R01 versus the SBIR/STTR funding (and perhaps a version of SBIR/STTR for academic institutions). A healthy biomedical research ecosystem requires both discovery and application. Translation converts knowledge into benefit, but basic science generates the knowledge that makes translation possible. Ultimately, scientific impact should be measured not only by what research immediately accomplishes, but also by what future discoveries it makes possible. Translation is the harvest, but basic science is the seed corn. If the research enterprise rewards only the harvest, it risks undermining the source of future innovation. NIH's assessment framework should ensure that foundational discovery science remains a visible and valued component of scientific success across all career stages, disciplines, and institution types.

7. Public Impact: NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.

In addition to traditional metrics of scientific dissemination, such as conference presentations and peer-reviewed publications, NIH must clearly recognize and reward contributions that demonstrably include community in meaningful ways and that directly benefit society, scientific community, and public health. These contributions may include informing clinical practice guidelines and public health practice/policies; improving patient outcomes; enhancing public understanding of science, health, and medicine; fostering meaningful public engagement in science; developing widely adopted tools, software, databases, and research standards; and advancing the translation of discoveries through licensing, commercialization, and implementation. Recognizing these forms of impact would better align scientific incentives with NIH's mission to improve health, extend life, and generate value for the American public.

Scientific communication has become an increasingly important component of the biomedical research enterprise, yet it remains inadequately recognized and rewarded within current evaluation systems. Public trust in science depends not only on the rigor of the research itself, but also on scientists' willingness and ability to communicate evidence, uncertainty, limitations, and evolving knowledge in ways that are accessible and meaningful to a broad range of audiences. However, many researchers devote limited time to public engagement because such activities are rarely valued and may expose scientists to ideological criticism.

NIH can be a leader by recognizing, incentivizing, as well as facilitating the sharing of activities that

explicitly recognize and incentivize activities that strengthen public understanding and trust in science, including public lectures and outreach, engagement with patient and community groups, science communication training, expert participation in public forums, contributions to evidence-based public discourse, and the development of educational resources and content. Potential indicators of impact could include participation in NIH-supported science communication initiatives, contributions to public educational materials, evidence-based expert engagement in public discussions, public media (e.g., podcasts and social media influence) development of community-engaged research partnerships, and demonstrated reach and effectiveness of outreach efforts. Importantly, these activities should be evaluated based on their quality, accuracy, educational value, and societal benefit rather than on popularity metrics such as social media engagement alone.

NIH could document the ways that NIH-funded research has been taken up and used in practice and by practitioners. Citation of research in policy documents, such as through the Overton Index is one approach. Incorporating narrative descriptions of impact, such as used by the REF, is appropriate in many cases because it allows researchers to explain the context and significance of the use of their work. The UK's Research Excellent Framework (REF) includes a well-established method for documenting the societal impacts of research from across all disciplines.

8. Other Areas: Scientific inquiry can be shaped by changing policy and administrative priorities. These shifts may discourage work in areas with significant potential impact and may reduce attention to factors that influence health, access, outcomes, and broader societal effects. Establishing processes that protect rigorous, evidence-based research from influences that unnecessarily limit progress or application is therefore essential.

268. Gerontological Society of America

I am responding: On behalf of an organization

Name: James Appleby

Name of Organization: Gerontological Society of America

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/260819-GSA-Response-NOT-OD-26-087.pdf

Description: On behalf of the Gerontological Society of America (GSA), the nation's oldest and largest interdisciplinary organization devoted to the study of aging, I am pleased to submit these comments in response to NOT-OD-26-087.

GERONTOLOGICAL SOCIETY OF AMERICA®

1101 14th Street NW, Suite 1220
Washington, DC 20005
1.202.842.1275
www.geron.org

Meaningful Lives As We Age

August 19, 2026

National Institutes of Health Office of Science Policy

Re: Request for Information on Measuring and Rewarding Scientific Impact (NOT-OD-26-087)
Submitted electronically via https://osp.od.nih.gov/comment-form-measuring-and-rewarding-scientific-impact/

Dear Colleagues at the NIH Office of Science Policy:

On behalf of the Gerontological Society of America (GSA), the nation's oldest and largest interdisciplinary organization devoted to the study of aging, I am pleased to submit these comments in response to NOT-OD-26-087. GSA's more than 6,000 members span the biological, behavioral, social, and health sciences and include researchers, clinicians, and educators working across the full spectrum of the aging enterprise. This response also draws on input from GSA-affiliated researchers with direct experience in NIH-funded science and policy. We commend NIH for undertaking this reflection and for seeking input from the community that will ultimately be governed by any new indicators.

Aging research is an instructive test case for this effort. It is inherently longitudinal, team-based, and translational, and it sits at the intersection of nearly every discipline NIH supports. Citation and publication counts, in particular, can be poorly correlated with genuine improvements in health and longevity. Alzheimer's disease research offers a cautionary example: for many years, imaging of amyloid plaques was treated as the gold standard for staging disease, and publications and citations surged as more researchers gained access to imaging facilities. Over time, however, the field learned that small soluble amyloid fibrils (monomers and dimers of amyloid protein), rather than the plaques themselves, appear to be the more direct driver of disease progression, and that multiple genetic and biological pathways contribute to the disease.

More recently, blood-based biomarkers for Alzheimer's disease have been developed. Blood has the advantage that many markers can be assessed simultaneously, offering the potential to build a pattern of markers that better signals disease progression and suggests causal pathways in ways the amyloid plaque signal that long dominated the field could not. Yet if success continues to be measured through publication counts and citations, it is still the older imaging research on amyloid plaques that dominates the top of those tables.

A single dominant metric, in other words, can crowd out messier but more accurate science for years at a time. We believe any new framework for measuring and rewarding scientific impact should be built with, not around, fields like ours, and should be designed with an eye toward this kind of risk. We offer the following comments organized around the categories identified in the notice.

Rigor and Reproducibility

Much of the foundational evidence base in aging research comes from long-running cohort studies, some spanning decades, involving older adults with heterogeneous health trajectories, multiple comorbidities, and evolving diagnostic criteria. Standard replication metrics developed with single-endpoint, short-duration studies in mind may not fit this reality well. We also note a structural weakness in replication as currently practiced: replication attempts are frequently motivated by an initial challenge to a finding, and successful replications are less likely to be published than failures to replicate, which can distort the literature over time. We recommend that NIH:

- Reframe replication initiatives as replication-and-generalization initiatives, incentivizing researchers to test how far an effect extends to other models, populations, or contexts rather than only whether it can

be repeated under near-identical conditions. This shifts the incentive from simply challenging a finding to understanding how far it extends, and treats generalization across clearly defined populations of older adults, including by age cohort, sex, health status, care setting, socioeconomic conditions, geography, and, where relevant to the scientific question, race and ethnicity, as an important contribution to rigorous science.

- Support harmonization and re-analysis studies across existing longitudinal cohorts (e.g., the Health and Retirement Study, the National Health and Aging Trends Study, the Baltimore Longitudinal Study of Aging) as a cost-effective and high-value form of reproducibility work.
- Require or strongly encourage sharing of unpublished as well as published replication attempts, to counteract the file-drawer problem in which only failures to replicate see print.
- Clarify how causal inference will be defined and credited in observational aging research, where randomized designs are frequently infeasible or unethical, so that rigorous quasi-experimental and natural-experiment approaches are not disadvantaged relative to trial-based work.

Data, Software, and Model Sharing

GSA members rely heavily on shared infrastructure such as the National Archive of Computerized Data on Aging and public-use files from federally supported cohort studies, and NIH's sustained investment in that infrastructure has been invaluable. The harder challenge lies in incentivizing sharing of individually developed data, software, and models, where researchers may reasonably see career or financial risk in releasing materials before they have fully mined them. We encourage NIH to:

- Require that investigator biographical sketches identify, for each listed publication, where associated data, models, or software can be accessed. Applied consistently over time, this would normalize sharing as a routine expectation of being an NIH-funded researcher rather than a special favor.
- Credit the sustained curation, documentation, and technical support that keep shared aging datasets usable over decades, not only the initial act of data deposit.
- Continue and periodically revisit NIH's work on standardized outcome measures that allow results to be combined or compared across studies, recognizing that instruments and research directions evolve and that standard measures need scheduled review rather than indefinite fixity.
- Ensure any new incentive structure accounts for the additional privacy, consent, and re-identification safeguards required when sharing data on older adults, particularly those with cognitive impairment or residing in long-term care settings.

Training and Mentorship

The biomedical and behavioral aging research workforce is not growing quickly enough to meet the needs of an aging population, and current challenges facing early-career investigators make sustained, high-quality mentorship more important, and more strained, than at any point in recent memory. Mentorship of the kind that produces strong independent researchers takes substantial time, yet it carries no direct financial reward and competes with grant-writing, supervision of research staff, and administrative demands. We recommend that NIH:

- Work with academic and institutional partners to formally recognize mentoring of junior researchers as part of the teaching role for senior investigators with established research records, giving institutions a concrete mechanism to protect mentors' time rather than relying on goodwill alone.
- Track the career outcomes of trainees and mentees over time, including retention in aging-focused research, rather than relying solely on point-in-time productivity counts.

- Credit mentors for sustained, longitudinal investment in early-career investigators, including those supported through T32 and K-series awards, and for mentoring across career stages and disciplinary boundaries, including clinician-scientists and other dual-role trainees whose paths to independence differ from PhD-only investigators.

Collaboration

Advancing the science of aging almost always requires integrating biological, clinical, behavioral, and social science expertise, and NIH has taken real steps to recognize this through the multiple-PI mechanism and collaboration-focused review criteria. Recognition at the point of reward, however, still accrues overwhelmingly to individuals rather than teams: NIH's signature individual-investigator honors, such as MERIT and Director's Pioneer awards, have no team-based counterpart. We recommend that NIH:

- Establish a team-based analogue to its flagship individual awards, recognizing that the most consequential advances increasingly emerge from social and scientific interaction among researchers rather than from any single investigator working alone.
- Develop team-science metrics that credit distinct forms of contribution (e.g., cohort stewardship, biostatistical design, implementation science, community engagement) rather than defaulting to author-order or single-PI conventions poorly suited to large, multidisciplinary teams.
- Pilot early-career-stage convenings, organized around career stage rather than shared topic, that bring together researchers from different fields and departments with substantial unstructured time for discussion. Topic-focused workshops tend to draw researchers who already share interests and language; career-stage convenings are more likely to spark collaborations across genuinely different fields.
- Recognize contributions to practice-translation collaboratives that bring researchers together with health systems and community-based organizations, such as models built on frameworks like the Age-Friendly Health Systems initiative's 4Ms (What Matters, Medication, Mentation, and Mobility).

Entrepreneurship and Translation

Translation in aging research is not limited to commercial products, and even when a researcher successfully develops a genuine practical advance, a host of non-scientific obstacles can still keep it from reaching patients or consumers, from market access barriers to entrenched competitors. We encourage NIH to broaden its definition of translational impact to include the adoption of evidence-based practices, care models, and clinical workflows by health systems, payers, and community organizations, not only patents, licenses, or start-up formation, which will otherwise systematically undercount translational work in geriatrics, gerontology, and long-term services and supports. We also encourage NIH to consider a modest expansion of program staff capacity dedicated to post-discovery translation, distinct from core review responsibilities, whose role would be to identify research portfolios nearing practical application and connect investigators with the business, regulatory, or dissemination expertise needed to carry findings the rest of the way.

Foundational Scientific Exploration

Basic biology of aging research, including work on the hallmarks of aging and geroscience, exemplifies high-risk, high-reward science with a long horizon to clinical relevance, and scientific curiosity should remain a legitimate driving force in what NIH funds. At the same time, some subfields within aging research, caregiving research among them, show signs of having grown repetitive as available tools and paradigms are exhausted. NIH's existing High-Risk/High-Reward program is well positioned to help revitalize such areas, and we recommend NIH consider expanding it along two lines particularly relevant to aging research: a multiple-PI variant requiring senior investigators from different fields, intended to spark new approaches through the collision of disciplinary

perspectives; and a dedicated early-stage-investigator track, since researchers earlier in their careers are often less invested in established research pathways and better positioned to pursue genuinely novel directions.

Public Impact

Loss of public trust in biomedical research, most visibly around vaccine science in recent years, is among the most serious challenges facing the research enterprise, and it is a challenge NIH cannot address through communications alone. GSA has invested substantially in this space through our National Center to Reframe Aging, which applies peer-reviewed communications science to help the public and policymakers understand the value of aging research and correct persistent misconceptions about aging and older adults. We believe NIH can extend this kind of public-facing translation down to the level of the individual grant, and recommend that NIH:

- Require grant applications to describe a plan for sharing findings with the public, whether through social media, community presentations, or other direct engagement, and require progress reports and renewals to report on audience reach, engagement, and any resulting feedback.
- Recognize measurable shifts in public or provider understanding, adoption of research-informed communications by other organizations, and downstream effects on health outcomes, cost of care, and public trust in science as legitimate indicators of scientific impact in their own right.

Closing

GSA appreciates NIH's commitment to modernizing how scientific impact is measured and rewarded, and we would welcome the opportunity to serve as a resource as this work continues. If you have any questions or want any additional information, please contact Patricia D'Antonio, VP, Policy & Professional Affairs at pdantonio@geron.org.

Respectfully submitted,



James C. Appleby, BSPharm, MPH, ScD (hon)
Chief Executive Officer

269. The Sijbrandij Foundation

I am responding: On behalf of an organization

Name:

Name of Organization: The Sijbrandij Foundation

Type of Organization: Non-Profit Research Organization

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: At the Sijbrandij Foundation, we have mapped a concrete set of Patient First reforms, of which the initiatives referenced below are examples. We would welcome the opportunity to share the full agenda and the supporting evidence for each.
Reproducibility matters most where it protects patients, and its standards should fit the setting. In rare and advanced disease, where any one center sees too few cases to power a conventional study, NIH should reward fit-for-purpose, context-driven evidence and the replication or validation that de-risks a finding on the critical path to a therapy or diagnostic, rather than treating large-N replication of already-cited results as the default unit of rigor.
Measurable indicator: whether a replication or validation study changed a downstream decision to advance or halt a candidate that patients could receive.

2. Data, Software, and Model Sharing: NIH should measure whether health and research data are usable by the individuals who generated them: whether patients can obtain the underlying data, direct it to clinicians and researchers of their choice, combine it across institutions, and contribute it to research. Interoperability and patient-directed data release should be treated as scientific infrastructure, not merely administrative functions.
This is a concrete part of the Foundation's agenda:
patient control of their own tissue and its molecular data
HIPAA modernization that makes it easier for patients to gather and move their records
EMR interoperability so systems like Epic and Cerner exchange data with others
patient-directed release that lets a person share their data with any clinician or researcher they choose
Measurable indicators: time to complete a data release; completeness of the released record; portability across systems; and documented reuse in research or care.
Safeguards: patient-submitted data must be matched by rigorous consent, privacy protection, and data-quality safeguards, and incentives should reward returning value to the contributors.

3. Training and Mentorship: Workforce metrics should reward the mentors who build national capacity to bring discoveries to patients, including clinician-scientists, trialists, and regulatory and manufacturing scientists, not only those who train the most prolific publishers.

Measurable indicator: trainees' later contributions to patient-facing outcomes, such as trials led and therapies and diagnostics advanced, tracked over time.
Implementation: extend the trainee information NIH already collects to include progress toward new

treatments, tests, and other patient benefits, and recognize mentorship of trainees who pursue this work in career evaluations. This can be phased in using data NIH already holds.

4. Collaboration: Team science should credit the full chain that moves a discovery to patients, including patients, advocacy organizations, and the regulatory and manufacturing partners without whom access does not happen , using a contributor taxonomy that names these roles. Our substantive recommendations sit under Translation, Data Sharing, and Public Impact.

5. Entrepreneurship and Translation: NIH should recognize that translational impact includes not only developing a new therapy or diagnostic, but removing the regulatory, manufacturing, reimbursement, and institutional barriers that prevent an existing scientific capability from reaching people. For patients with rare or advanced disease, that path is often the binding constraint on benefit.
The Foundation is advancing a concrete agenda of such reforms, which serve as worked examples of translation-enabling change:
modernizing early-phase trial pathways — a notification pathway for Phase 1 safety studies under IRB oversight, and investigator choice of any federally compliant IRB;
expanded access after Phase 1, including conditional approval, so promising therapies are not lost in "invisible graveyards";
fit-for-purpose ("GMP-light") manufacturing for low-volume personalized products, verifying the output rather than only the production process;
regulatory pathways for approved algorithms and processes, not only approved molecules, to enable truly personalized medicines; and
context-driven evidence for reimbursing rare-disease therapies.
NIH should treat this policy-translation work as a creditable form of translational impact, and assess it through verifiable artifacts along a clear arc: a published evidence base, a drafted regulatory or legislative change, documented endorsement from affected professional and patient communities, and ultimately adoption or implementation.
Measurable indicators: time from discovery to a benefit a patient can receive; the number of patients who benefit, including through earlier or expanded access; contributions that make trials faster and easier to scale, such as making the real-time clinical-trials pilot permanent; and documented movement from an evidence base to an adopted change.
Implementation: fold these contributions into the biosketch and progress report NIH already collects, and give reviewers explicit authorization to weigh them alongside, or above, publication and citation counts.
Unintended consequence to avoid: over-weighting near-term commercial return would discourage unprofitable but high-value work, including for rare diseases. Translation measures should be read alongside the access and equity measures below.

6. Foundational Scientific Exploration: NIH should credit the mentors who build national capacity to bring discoveries to patients, such as clinician-scientists, trialists, and regulatory and manufacturing scientists, assessed by trainees' later contributions, such as trials led and therapies or diagnostics advanced. This area is secondary to our submission, and we defer to the workforce community for detailed indicators.

7. Public Impact: A discovery has limited public impact if people cannot learn that it exists, access the relevant trial or diagnostic, or reach the clinicians able to use it. NIH should therefore measure not only

whether an advance was developed, but whether the systems around it allow people to benefit from it. Concrete dimensions of access the Foundation works on include:
Whether patients being told about relevant trials, including in oncology;
coverage of exploratory diagnostics such as raw genomic sequencing;
the freedom to see an out-of-state doctor or nurse, so that expertise is not bounded by geography
Measurable indicators: whether affected patients could find and reach an advance (trial and diagnostic awareness and access); equity of that reach across populations and geographies; and adoption into clinical practice — reported publicly and prominently, so that the headline of NIH's impact is health delivered to patients, not papers published.

8. Other Areas: The ultimate purpose of biomedical research is not to maximize publications, approvals, or institutional compliance; it is to improve the health and lives of individuals. In clinical settings, this means orienting the system to maximize patient survival and benefit rather than to minimize institutional or practitioner liability. NIH's impact framework should give greater weight to work as it moves closer to meaningful health benefit, while preserving appropriate measures for foundational research whose effects emerge over longer horizons.
Three cross-cutting safeguards should govern any new indicator:
Feasibility across career stage and institution. Impact should be assessed at the level an investigator can reasonably influence — early-career researchers credited for their contribution to a larger translational chain; foundational scientists judged on rigor and originality; and outcomes like approval or population health assessed at the team, institution, or portfolio level.
Guard against proxy metrics. NIH should not replace citation counts with new activity counts. Datasets shared, collaborations formed, or policy proposals written should matter only when there is credible evidence they contributed to reuse, adoption, access, or health benefit.
Involve patients in defining the metrics. NIH should include patients and advocacy organizations in designing the indicators it will reward, so the endpoints reflect what patients actually value.

270. American College of Rheumatology

I am responding: On behalf of an organization

Name: Connie Herndon

Name of Organization: American College of Rheumatology

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Vice President, Research, Training & Workforce

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

271. American College of Rheumatology

I am responding: On behalf of an organization

Name: Connie Herndon

Name of Organization: American College of Rheumatology

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Vice President, Research, Training & Workforce

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/2026-NIH-RFI-on-Meauring-Scientific-Impact.docx

AMERICAN COLLEGE of RHEUMATOLOGY
Empowering Rheumatology Professionals

Date:	August 19, 2026
Subject:	Response to NIH Request for Information (RFI) on Measuring and Rewarding Scientific Impact
Notice Number:	NOT-OD-26-087
Submitted By:	Linda Hiraki, MD, ScD, Chair, ACR Committee on Research
Contact Info:	Linda Hiraki Email: linda.hiraki@sickkids.ca
	Connie Herndon (ACR Staff) Email: cherndon@rheumatology.org

Thank you for the opportunity to provide comments on the National Institutes of Health (NIH) Request for Information, Measuring and Rewarding Scientific Impact (NOT-OD-26-087). On behalf of the American College of Rheumatology (ACR), we appreciate NIH's commitment to strengthening the biomedical research enterprise by promoting transparent reporting, rigorous replication, careful interpretation of research methods and analyses, and the ongoing refinement of scientific knowledge. These principles are fundamental to maintaining public trust in science and ensuring that research findings are reliable, reproducible, and ultimately beneficial to patients and society.

The ACR recognizes the importance of meaningful approaches to assessing scientific impact that extend beyond traditional metrics such as publication counts, citation indices, or funding awards. We agree that a comprehensive evaluation of research contributions should also consider factors such as mentorship, data sharing, development of scientific resources, contributions to team science, and the translation of research findings into clinical practice and improved health outcomes. This comprehensive evaluation is especially important for fostering collaboration among investigators in the U.S. and internationally, ensuring a broad and equitable approach to scientific questions.

At the same time, it is important to recognize that academic institutions already devote substantial resources to oversight and training researchers in these areas. Through NIH-supported training programs, including T32 institutional training grants and related career development initiatives, trainees receive extensive instruction in rigorous research design,

reproducibility, responsible conduct of research, scientific communication, and data stewardship. In addition, institutions have established faculty evaluation and promotion processes that assess research productivity, scholarly impact, educational contributions, and service to the scientific community.

While the ACR supports broadening the definition of scientific impact, we note that there is considerable variation across institutions in how faculty performance and research contributions are evaluated. Institutional missions, organizational structures, available resources, and disciplinary priorities differ significantly, leading to diverse evaluation frameworks. As a result, the development of a single, uniform NIH-driven approach to measuring scientific impact would prove difficult to implement equitably across the biomedical research landscape. Such an effort could also create additional administrative and reporting requirements that increase the burden on investigators and institutions without necessarily improving the quality of assessment.

The ACR also has concerns regarding how measures of public impact would be defined, assessed, and applied. Public engagement and societal influence are important outcomes of scientific research; however, these measures can be difficult to quantify objectively and may vary substantially across research disciplines, career stages, and institutional settings. Clear, transparent, and validated criteria would be necessary to ensure that assessments of public impact are fair, consistent, and do not inadvertently disadvantage researchers whose work contributes significantly to scientific advancement but whose impact may not be immediately accessible to broader audiences.

Overall, the ACR believes that academic institutions are best positioned to evaluate the full scope of faculty contributions and research impact because they have direct knowledge of institutional missions, disciplinary expectations, and individual career trajectories. While NIH plays a critical role in supporting and advancing biomedical research, we believe the primary responsibility for measuring and evaluating these outcomes should remain with institutions. NIH should continue to encourage and support best practices in research assessment while avoiding the creation of duplicative processes that could increase administrative burden and divert time and resources away from scientific discovery.

Thank you for considering the ACR's comments on this important issue. We appreciate NIH's ongoing efforts to strengthen the biomedical research workforce and promote high-quality, impactful research.

272. American Association for the Advancement of Science

I am responding: On behalf of an organization

Name:

Name of Organization: American Association for the Advancement of Science

Type of Organization: Professional Organization/Association

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: • Recognize both reproducibility and replicability: Assessment frameworks should recognize contributions that strengthen both reproducibility and replicability, which are distinct dimensions of scientific rigor. Such frameworks include data and code sharing, transparent reporting, and independent replication efforts.

• Build on existing community standards: NIH should work with publishers, funders, scientific societies, and organizations such as the Center for Open Science to build upon established community efforts rather than creating new, duplicative requirements. Existing initiatives, including the Transparency and Openness Promotion (TOP) Guidelines, provide a strong foundation for advancing reproducibility across disciplines. So do reporting frameworks like MDAR, STAR methods, or field-specific efforts like CONSORT for clinical trials.

• Reward transparent communication of evidence: NIH should reward practices that make the strength of evidence transparent. This starts with sound experimental design and inclusion of the full range of high-quality data needed to reproduce a study. Study limitations, for a given work, should be explicitly communicated in manuscripts and abstracts. Journals play an important role in helping authors accurately communicate uncertainty and contextualize findings before publication, strengthening both scientific rigor and public understanding. (Research shows that making limitations clear is among the key contributions of journals, for papers that were previously on preprints.) Such efforts in communicating uncertainty can be further supported by communication teams committed to accurate dissemination (i.e., When “Opposite” Scientific Findings Collide; Replacing Public Doubt with Public Confidence: Experiments in Building Trust at Science).

• Encourage correction of the literature, where needed: Authors should not be punished for correcting errors. Such correction is courageous; it also helps build public trust. Journals that are improving policies and workflows that prioritize a reliable correction process play a key role in this process, especially when they emphasize public communication.

• Reward replication as scientific discovery: Replication should not be viewed solely as a mechanism for confirming prior findings or identifying errors. When well-designed studies yield different results, they often reveal previously unrecognized complexity or methodological limitations that deepen scientific understanding. Such findings ultimately strengthen the scientific record. Research assessment systems

should recognize high-quality replication studies, including those that do not reproduce earlier results, as valuable contributions. When part of the published record, such findings enable future researchers to build on a more complete and reliable evidence base.

• Recognize disciplinary differences in reproducibility: Approaches to reproducibility vary across disciplines. NIH should avoid creating universal metrics that disadvantage fields with different research norms and practices. Through ongoing work to improve structured reporting across Science journals, AAAS has found that the information needed to support reproducibility, or the approach by which that information is gathered, can differ by discipline. Providing better support for authors to improve the completeness and quality of the scientific record should be flexible enough to reflect related differences.

2. Data, Software, and Model Sharing: • Data should be evaluated not only by whether they are shared, but also by whether they are discoverable, reusable, and ultimately used to enable new science.

• Reduce burden while improving reporting: Incentives for data and code sharing should be paired with investments in tools that reduce the administrative burden of compliance. At Science, AAAS is evaluating AI-assisted workflows that automatically generate structured reporting (e.g., MDAR checklists) from submitted manuscripts for editorial review. Early experience suggests that automated reporting can improve the completeness and consistency of research documentation while reducing effort for authors and editors. NIH should encourage innovations that make transparent reporting easier to achieve, not simply another compliance requirement.

• Invest in data infrastructure, not just data sharing: Data sharing alone is not sufficient. To maximize the value of federally funded research, NIH should continue to encourage deposition of data into trusted, discipline-specific repositories that curate submissions, apply standardized metadata, and enable broad discovery and reuse. Resources such as the Protein Data Bank demonstrate how well-curated community databases can accelerate scientific progress, including enabling advances in AI-based protein structure prediction. Publishers play an important complementary role by encouraging or requiring deposition in these repositories.

• Reward data quality, not just data availability: NIH should consider indicators that assess not only whether data and code are shared at the time of publication, but also whether they remain Findable, Accessible, Interoperable, and Reusable (FAIR) over time. Advances in automated assessment now make it increasingly feasible to evaluate FAIR compliance, metadata completeness, and data usability at scale, providing more meaningful measures of scientific value than indicators of data availability alone.

• Measure reuse as scientific impact: As data sharing becomes more common, NIH should look beyond whether data are deposited to whether they are reused. Data citations provide a practical means of recognizing the scientific value of shared datasets and could complement other indicators of research impact. Tracking data citations would also recognize researchers whose datasets enable new discoveries, extending research assessment beyond traditional publication and citation metrics.

3. Training and Mentorship: • AAAS appreciates that training and mentorship are essential dimensions of scientific impact. Biomedical research advances through the findings generated by individual investigators and the development of people, teams, and institutional environments capable of

producing rigorous, ethical, collaborative, and publicly valuable science over time. NIH should therefore recognize effective training and mentoring as substantive contributions to its mission.

• At the same time, AAAS urges NIH to proceed carefully in developing indicators for these activities. The fact that an activity or outcome can be counted does not necessarily mean that it is a valid measure of quality or impact. Training and mentorship often produce effects that are developmental, relational, cumulative, and visible only over long periods. Easily quantified indicators, such as the number of trainees supervised, trainee publication counts, placement in academic positions, or subsequent receipt of NIH funding, may provide some information, but none is sufficient to establish effective training and mentorship occurred. If treated as proxies for quality, these indicators could also create unintended incentives and obscure the contributions of teams and institutions.

• The goal should not be to produce a universal "mentorship score." Rather, NIH should develop credible and context-sensitive evidence that enables the agency to recognize excellent mentorship, identify opportunities for improvement, and hold investigators, research teams, programs, and institutions appropriately responsible for developing the biomedical workforce. We offer the following recommendations toward this end:

1. Recognizing training and mentorship as scientific impact. Training and mentorship help determine whether the biomedical research enterprise can produce reliable knowledge, adapt to emerging scientific challenges, and translate discoveries into benefits for the public. Effective mentors transmit more than technical knowledge; they help developing scientists strengthen research judgement, responsible research practices, communication and collaboration skills, professional networks, scientific independence, and capacity to work across sectors and disciplines.

NIH should distinguish among related but different functions: (i) Research supervision includes oversight of work, safety, and compliance. (ii) Training develops scientific, technical, ethical, and professional competencies (e.g., science communication). (iii) Mentorship provides sustained career and psychological support, including guidance, feedback, sponsorship, and support for increasing independence.

These functions often overlap but employing or supervising a trainee should not by itself be treated as evidence of effective mentorship. NIH should also recognize that mentorship is often provided by a network that may include principal investigators, other faculty members, laboratory staff, peers, former trainees, professional staff, and collaborators inside and outside of the institution. Assessment and incentives should recognize these distributed contributions rather than attributing all mentoring outcomes to a single principal investigator.

2. Use a multidimensional portfolio of evidence. No single indicator can adequately capture mentoring quality or its effects. NIH should consider a portfolio organized around four complementary domains: mentoring practices, trainee experiences, developmental outcomes, and institutional conditions.

Mentoring Practices. Relevant evidence could include: Development and periodic review of individualized mentoring or professional-development plans. Clear expectations concerning

responsibility, authorship, publication, data stewardship, intellectual credit, communication, and conflict resolution. Mentor participation in evidence-informed preparation or professional development. Use of mentoring teams, peer mentoring, and access to career-specific mentors rather than reliance on a singular individual. Availability of a confidential and workable process for changing mentors or addressing serious mentoring problems.

These indicators are relatively feasible to collect, but NIH should avoid reducing them to compliance checks. The presence of a mentoring plan, for example, reveals little unless the plan is used, revisited, and experienced as helpful by the trainee.

Trainee Experiences. Confidential trainee feedback should be a central component of assessment. Possible indicators include trainees' perceptions of clarity of expectations and fairness in the allocation of credit; support for scientific independence and intellectual ownership; access to professional networks, sponsorship, and career information; supportive and safe environment that provides for the ability to raise questions, acknowledge mistakes, or express concerns without retaliation; and support for professional goals across the full range of biomedical careers.
Mentor self-reports should not be the sole evidence of effectiveness. At the same time, NIH will need to protect confidentiality, particularly in small laboratories, departments, and training programs where critical responses may be identifiable. In many circumstances, trainee-experience data may initially be most appropriate for aggregation and improvement at the program, department, or institutional level.

Developmental Outcomes. NIH should focus on whether trainees are developing the capabilities and agency needed to pursue their goals. Possible indicators include growth in research, technical, communication, teamwork, leadership, and career-management competencies; growth in adjacent skill development relevant to trainees' goals (e.g., science communication, community engagement, science policy, etc.); understanding and application of rigorous and responsible research practices; increasing scientific independence, responsibility, and confidence; progress toward appropriate educational or training milestones; and movement into careers aligned with trainees' interest, skills, and goals.
Outcome evidence should be interpreted in context and over an appropriate period. A developmental outcome observed several years later may be more meaningful, but it is often difficult to attribute to a single mentor, grant, or training experience.

Institutional Conditions and Accountability. Mentorship quality is shaped by departments, training programs, and institutions, not solely by individual effort. Relevant indicators could include whether workload, promotion, and recognition systems meaningfully value mentorship; whether mentors receive sufficient time, preparation, and institutional support; availability and quality of mentor-development programs; access to independent ombuds, grievance, and mentor-rematching processes; and support for team, peer, cross-disciplinary, and networked mentoring models.
Assessing institutional conditions is important because NIH should not reward individual mentors for outcomes produced by well-resourced environments while penalizing others for structural conditions outside of their control. Conversely, institutions should not receive credit merely for adopting formal policies that are not consistently implemented.

3. Define career success broadly. NIH should avoid defining successful training primarily through

academic placement, publication output, or subsequent NIH funding. The biomedical research enterprise depends on professionals working across academic research, industry, government, clinical practice, public health, nonprofit organizations, policy, education, research administration, communication, and other areas.

Treating entry into a tenure-track academic position or receipt of an NIH award as the preferred outcomes would undervalue many careers through which biomedical training benefits the public. It could also encourage mentors and institutions to steer trainees toward a narrow range of paths or to disregard trainees whose goals lie outside of academic research.

A more appropriate approach would assess whether trainees received strong preparation, understood the range of opportunities available to them, developed relevant competencies and networks, exercised informed agency in their career decisions, and moved toward work aligned with their goals. NIH should recognize that career paths are increasingly nonlinear and that the public value of research training may emerge through many occupations and sectors.

4. Pilot and validate indicators before using them for high-stakes decisions. AAAS recommends that NIH develop a limited common core of measures while allowing for different career stages, disciplines, research settings, grant mechanisms, and institutional types. Mentoring an undergraduate researcher, doctoral student, postdoctoral scholar, clinical-scientists, staff scientist, or early-career faculty member involves different objectives and constraints. Laboratory-based, clinical, computational, community-engaged, and team-science environments may also require different evidence.

NIH could take the following ordered steps:
o Develop a concise common framework through consultation with trainees, mentors, institutional leaders, workforce researchers, social scientists who conduct mentorship research, and professional societies.
o Pilot the framework in selected institutional training grants, career-development awards, and research programs.
o Combine confidential trainee feedback with mentor, program, institutional, and administrative evidence.
o Test measures for validity, reliability, administrative burden, privacy risks, susceptibility to gaming, and differential effects across institutional types.
o Examine whether indicators predict meaningful developmental outcomes rather than merely correlating with institutional prestige or existing resources.
o Use early results primarily for formative feedback and program improvement.
o Introduce consequential uses in funding or renewal decisions only after the measures have demonstrated sufficient validity and fairness.
NIH should also evaluate the measurement system itself over time. Indicators can change behavior, sometimes in ways that undermine the underlying goal. A measure that initially appears useful may become less informative once it becomes a high-stakes target.

5. Reward quality and improvement while guarding against unintended consequences. NIH could recognize training and mentorship through structured narrative or portfolio-based evidence in relevant

grant applications and progress reports. Reviewers could be asked to consider the quality of mentoring practices, evidence of trainee development, responsiveness to feedback, and contributions to team and institutional mentoring. NIH could also support protected effort, mentor development, institutional capacity-building, and recognition for demonstrated excellence or meaningful improvement.

Indicators should inform expert judgement rather than be mechanically combined into a numerical score or ranking. NIH should explicitly guard against several foreseeable consequences:
o Box Checking: Completion of mentor training or a mentoring plan becomes a substitute for effective practice.
o Gaming: Investigators document meetings or activities that trainees do not experience as meaningful.
o Selection bias: Mentors become reluctant to work with trainees who may need additional support or whose career paths could lower conventional outcome metrics.
o Narrow definitions of success: Academic placement, publications, or NIH funding are treated as the preferred result.
o Attribution errors: Outcomes produced by multiple mentors, teams, and institutions are attributed to one principal investigator.
o Privacy and retaliation risks: Trainees believe that candid feedback can be identified and used against them.
o Institutional inequities: Institutions with greater administrative capacity appear more effective because they can collect and present more sophisticated data.
o Added uncompensated labor: New expectations place disproportionate burdens on individuals already performing substantial mentoring service.
o Short-term outlook: Immediate, easily counted outputs displace attention to development that occurs over longer periods of time.
o Metric substitution: NIH replaces an overreliance on publication and citation metrics with another reductive set of quantitative targets.
Guardrails should include multiple sources of evidence, confidential trainee input, contextual review, recognition of team mentorship, consideration of career stage and opportunity, technical assistance for institutions with fewer resources, and mechanisms for identifying and correcting adverse effects.
The most useful framework will combine evidence about mentoring practices, trainee experience, developmental outcomes, and institutional responsibility while acknowledging that some of the most significant effects of mentorship are relational, cumulative, and visible only over time.

AAAS encourages NIH to adopt a measured, evidence-building approach – define the qualities it seeks to promote, pilot a concise and multidimensional framework, validate indicators across research contexts, protect trainees, and use the resulting evidence to encourage improvements as well as accountability. This approach would better align incentives with NIH's mission while avoiding the creation of another narrow metric that rewards what is easiest to count rather than what matters most.

4. Collaboration: • One mechanism is through authorship metrics, which are typically defined as scientists who have made substantial contributions to the conception or design of the research; the acquisition, analysis, or interpretation of data; or the creation of new software used in the work. NIH should consider using more detailed contribution summaries that outline specifically the manner to which an individual contributed to the research, such as the Contributor Roles Taxonomy (CRediT) model encouraged in Science publications.

• Not all collaborative projects result in authored papers, and collaboration can be measured both qualitatively (knowledge sharing, publication, citations, new funding awards, products) and quantitatively (number of institutions, types of institutions, disciplines, countries, funding sources, etc.).

• Qualitative metrics may require a broad array of approaches that are unique and dependent on the nature or stage of research. We recommend that NIH take this opportunity to support social science research into measuring collaboration before determining appropriate metrics – both qualitative and quantitative.

• With the prospect of artificial intelligence becoming a tool in the conduct of research, one can argue that AI itself can be considered a "collaborator." While NIH has existing policies and guidelines with respect to the use of AI in the drafting of grant proposals and writing of research papers, NIH should consider developing guidance on measuring the use of AI in the conduct of research. For example, the compilation of data, tables and graphics, and the analysis of data, tables, and graphics.

5. Entrepreneurship and Translation: • Translating and integrating new discoveries and technologies into products for people may be due to a need to incubate research for a longer period or require an advancement in a specific technology or set of technologies to translate from a conceptual to a production phase. This is further encumbered by the fact that the processes for advancing translational research through various stages (early-stage, mid-stage, late-stage) will be unique and dependent on the specific experiment.

• Gaps for moving from bench to bedside may fall outside the research lab. Furthermore, the difficulty of scaling from the research bench to a commercial product may be due to the lack of advanced manufacturing to move from conceptual to the production stage. These are issues that have often little to do with the research that served as the foundational innovation.

• Metrics for success come in different forms: reducing risk, improving efficacy, opening new pathways for research, advancing scalability, patentability, publication metrics, number of patients served, number of start-ups created, jobs created, etc.

• We urge caution when developing metrics that can be applied across translational sciences as they may be misleading, ineffective, and even damaging. This is especially true when applying metrics that are ineffective or misleading for determining grant awards.

• The federal government, however, does support programs like STTR, SBIR, ARPA-H, technology transfer, NCATS) that are developed to bridge the so-called "Valley of Death." A landscape analysis of how effectively differing incentives can bridge this divide on different institutions, infrastructure needs, fiscal gaps (both public and private), and expertise of individual researcher, and/or stage of career may serve as a useful first step.

6. Foundational Scientific Exploration: • Academia performs both foundational research and high-risk, high-reward research supported by specific federal programs (e.g., Common Fund, ARPA-H), and it is important that metrics for both pathways remain separate.

• Both basic and high-risk, high-reward research contain some form of uncertainty, but in different ways that necessitate different metrics. With basic research, the uncertainty is due to exploring the edge of human knowledge – we simply do not know what can be learned and what benefits it may bring in other areas outside the realm of initial exploration. In this case success is broad: increasing fundamental scientific knowledge and understanding of the underlying foundations of phenomena and observable facts or refining our understanding of something we've only begun to explore.

• With high-risk, high-reward research, the goal is generally known, but how to achieve that goal involves risks and uncertainty undertaken by the researcher, and thus success can be more narrowly defined. Success in this case could be accomplishing the goal in the initial proposal or simply advancing a few steps closer to solving the problem.

• Conflating the metrics of success for these two types, or research would mean undercounting the advantages each one brings to the United States. Metrics for high-risk, high-reward research typically follow contributions to core knowledge (e.g., theoretical, empirical); level of uncertainty whether for risk of failure or success; the novelty and creativity; the length of time needed for the research; and the potential outputs such as commercial viability.

• Measuring novelty may depend on whether it is applied to novel knowledge that expands existing or introduces new scientific frontiers of understanding or to the application of foundational research in novel ways (e.g., honeybee algorithm) that are not necessarily considered high-risk but involve creative approaches. Each comes with different risks and uncertainties.

7. Public Impact: N/A

8. Other Areas: N/A

273. University of Pittsburgh

I am responding: On behalf of an organization

Name: Michael Holland

Name of Organization: University of Pittsburgh

Type of Organization: Not Applicable

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: The priority for understanding the efficacy of NIH's training and mentorship programs needs to be systematic, comparable data across institutions. The Institute for Research on Innovation and Science (IRIS) has built the infrastructure necessary to directly and dynamically measure the exact number, roles, and compensation of individuals supported by federal grants, including undergraduate and graduate students, postdoctoral researchers, and technical staff. Through its formal partnerships with U.S. Census Bureau infrastructure for privacy-restricted microdata, IRIS is able to link university grant and administrative data to IRS W-2 earnings data and Social Security Administration files. This permits tracking the longitudinal career paths and economic impacts of NIH-funded researchers and trainees, thus directly quantifying the impact of NIH biomedical research workforce development programs with policy-relevant resolution for the academic, industrial, governmental and non-profit sectors. Rather than continuing to seed the creation of new measures or small-scale, non-comparable studies, NIH needs to provide stable, long-term operational support to scale appropriately designed research infrastructure (whether that is IRIS, RTI International, ORISE or OSU's Ohio Education Research Center) that can rigorously examine the efficacy of NIH's training and mentorship portfolio across all NIH-funded institutions. Currently, IRIS is forced to fund itself inefficiently through a model of institutional memberships funded by Senior Research Officers and a mix of discrete sponsored research projects. A stably funded analytic infrastructure would allow NIH and research institutions to better tailor training and mentorship programs for the more than 60% of biomedical research workforce with advanced degrees that now works outside of academia according to NSF's recent Survey of Doctorate Recipients.

4. Collaboration: N/A

5. Entrepreneurship and Translation: As discussed in the recent NSF-funded "Workshop on Identifying and Removing Gaps and Barriers to Translating Research to Practice" that included broad federal agency participation, four recurring "valleys of death" were identified that commercial translational efforts face: [1] spinning out a technology and forming a startup; [2] finding product-market fit and launching a minimum viable product in a complex market; [3] demonstrating the first full-scale commercial product or facility; and [4] proving that the product is stable, profitable, and scalable.
Moving discoveries along these translational barriers requires replacing disconnected handoffs with a

coordinated system across academia, industry, philanthropy, government, and nonprofits. NIH has developed several effective mechanisms, including NIAID's Antiviral Drug Discovery (AViDD) Centers for Pathogens of Pandemic Concern program and NIDCR's Accelerating Product Excellence in Innovation and for Clinical Adoption (APEx) program, that integrate innovation and entrepreneurial efforts within an institute's mission. Other Institutes can adapt these mechanisms to their particular missions and needs. Additional options for NIH to support systematic coordination across these barriers are organized into the following four pillars. These can replace disconnected handoffs with a continuous loop of learning, feedback, adaptation, and application.

Pillar 1: Strengthen Connections to Market Signals and Industry Vetting

• Integrate industry-informed priority setting at the earliest research stages where goals and objectives are set.

• Co-develop challenge statements with industry input that can directly shape federally-funded research agendas

• Use milestone-based funding to facilitate industry engagement in translational pathways

• Embed industry fellows within research and startup ecosystems to provide continuous, insider market validation

Pillar 2: Expand Researcher and Student Exposure to Real-World Challenges and Entrepreneurship

• Expand support for experiential learning programs that connect research with industry

• Support adoption of Department of Defense's "Hacking for X" course methodology to teach students and faculty members how to use startup lean innovation and rapid prototyping to solve the real-world challenges industry faces in bringing products to market.

Pillar 3: Strengthen Translational Pathways and Coordination Across the Innovation Lifecycle

• Link follow-on sector-specific translational programs into integrated, lifecycle-spanning funding pathways. NIDCR's APEx is a particularly strong example of this.

• Develop translational pathways to impact for behavioral health interventions beyond traditional patent and startup models

Pillar 4: De-Risk Translational Progress and Enable Scalable Investment Pathways

• Provide rapid-access administrative supplements that enable investigators to quickly demonstrate specific commercialization-relevant milestones. Retaining consulting engineers in the review process can help agencies assess whether these milestones reflect real development, production, integration, and adoption requirements rather than academic indicators of progress.

• Establish risk-sharing approaches, including milestone-based supplements and co-investment models

• Develop predetermined global IP strategies and partnerships with shared cost and return structures. NSF's longstanding Industry-University Cooperative Research Centers program's approach to IP is worth examining.

The central challenge is therefore not simply to help more technologies cross individual valleys of death, but to build an innovation ecosystem capable of learning, adapting, and carrying promising ideas across the full journey to impact. The report can be found at https://tinyurl.com/2vwudhm8.

6. Foundational Scientific Exploration: Bottom-up, mission-driven strategic planning essential to build the support within the research community and the policy community for continued funding of bold and innovative research directions. Substantive strategic plans for basic research are required if policymakers and agency leaders are to justify major federal R&D investments, shape funding allocations, and design large-scale initiatives.

The NIH-Wide Strategic Plan outlines peer-review mechanisms, broad translational pipelines, and grant

lifecycle adjustments (like eliminating letters of intent or capping direct costs). It is a guide for how the agency distributes research awards, rather than a scientific strategy document. NIH strategic planning at the level of the Institutes and Centers can be characterized as top-down, disease-focused, and operational. Institute-level strategic plans focus on justifying the broad remit of an entire institute's mission area, defining very high-level areas of opportunity that lightly shape individual investigator-initiated R01 grant funding priorities and have a goal of balancing the competing interests of patient advocacy groups and academic researchers. Change in public health burden via domestic disease prevalence, life expectancy, clinical trial translations, and structural health equity has been only loosely correlated with the funded portfolio of awards (see Gao, T., Han, O., Li, G. et al. Allocation of the National Institutes of Health funding and burden of disease in 2008–2021 in the United States. npj Health Syst. 2, 41 (2025). https://doi.org/10.1038/s44401-025-00048-x).

Each NIH Institute and Center needs to engage in bottom-up, mission-driven strategic planning. These plans will enable budget formulation that aligns resources with specific scientific campaigns. They can scope and provide evaluation criteria for the mid-scale and large-scale investments that reduce scientific uncertainty and advance each Institute's progress on attaining its mission. Bottom-up community-wide planning does not inherently shift basic research to applied, but it forges sub-discipline alignment and helps small, disparate research groups come into greater agreement on technical definitions and help frame essential research infrastructure needs, whether that is the development of new instrumentation or the funding of national or regional-scale research cores that support broad user access.

The Department of Energy's Basic Energy Sciences Advisory Committee (BESAC) "Basic Research Needs" workshop reports provide a clear example for the NIH Institutes and Centers to model their bottom-up strategic planning efforts after. Aside from having the practical utility of defending budget requests for the DOE Office of Science, these reports successfully build a unified community voice that resonates well in federal science policy circles and are regularly used to highlight competitive gaps against international rivals (such as Europe and China) in key fields.

7. Public Impact: The Entrepreneurship and Translation topic above is framed explicitly in terms of products. There is no debate that drugs, devices, and clinical therapies are essential, but patient self-management of chronic diseases is increasingly essential to improve health behaviors, health outcomes, and quality of life and, in some cases, has demonstrated effectiveness for reducing health care utilization and the societal cost burden of chronic conditions (Allegrante JP, Wells MT, Peterson JC. Interventions to Support Behavioral Self-Management of Chronic Diseases. Annu Rev Public Health. 2019 Apr 1;40:127-146. doi: 10.1146/annurev-publhealth-040218-044008. Epub 2019 Jan 2. PMID: 30601717; PMCID: PMC6684026).

For community-engaged research that produces programs, interventions, services, or civic technologies that can transform lives rather than drugs, devices, or clinical interventions, their social impact often falls short because most universities' innovation infrastructure assume for-profit commercialization. Bench science researchers (often in medicine, natural sciences and engineering) have access to well-defined support systems that improve the likelihood that their promising research outputs can sustain, scale, and spread for maximum impact. Existing support systems (e.g., NIH's Clinical and Translational Science Institutes, NSF's Innovation Corps, Small Business Innovation Research/Technology Transfer) facilitate the translation pathway for products. As a result, public impact of bench science is optimized, funder investments show greater return, and research teams are able to hand off their research-based products and direct their energy to the next pressing research question.

Existing translational pathway support mechanisms of entrepreneurship training, seed funds, licensing

agreements, gap funds for prototyping or regulatory pathway planning, and coaching resources are not comfortable fits for community-engaged researchers (often in the social sciences) who want to see their research products and programs progress from idea to evidence-based output to sustainable public impact (Burgio, L. D., Disentangling the translational sciences: A social science perspective. Research & Theory for Nursing Practice, 24(1), 56-63 (2010); Sheridan, S. M., DiLillo, D., Hansen, D. J., DeKraai, M., Koenig-Kellas, J., Swearer, S. M., & Wheeler, L. A. Enhancing Lives via Interdisciplinary Translational Science (EnLITS): A circumplex model for the social-behavioral-educational fields, CYFS Working Paper No. 2016-1 (2016). Available at cyfs.unl.edu).

Not only are most VCs not interested in funding non-profit companies or technologies intended for the government sector, the range of sustainable business models for programs and services is larger than that for a technology-based startup company. Social entrepreneurs need to understand non-profit and government business models, they need help in developing funding strategies that will be complex mixes of fee for services rendered, philanthropic funding, and non-research agency state or federal grants. This lack of support and clear steps in the community-engaged translation pathway leads to under realized public impact, low return on funding agency investments, and university researchers expending undue time and resources on promising outputs neither they nor their partners are equipped to sustain.

There is a need to develop and scale a fit-for-purpose, implementation approach that can scale and sustain community engaged programs, inspired by the translational infrastructure for STEM innovations, that will support a “translation mindset” from the earliest stages of community-engaged researchers’ projects.

8. Other Areas: N/A

274. Association of American Universities

I am responding: On behalf of an organization

Name: Lizbet Boroughs

Name of Organization: Association of American Universities

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Senior Associate Vice President, Government Relations and Public Policy

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: The ability to reproduce research findings is essential to the integrity of biomedical science. However, a range of interconnected, systemic challenges can make replicating important studies difficult. Overcoming these obstacles requires a comprehensive approach that addresses every stage of the research process, from study design to data processing. Scientific standards serve as a foundational tool in advancing and ensuring research reproducibility.

AAU applauds recent standardization initiatives, including Stanford's Center for Open and REproducible Science (CORES), whose work, such as its recent symposium, is actively redefining how open data, code sharing, and transparent methodologies are taught and embedded into research workflows. Another example is the recent conference at Frederick National Laboratory for Cancer Research to promote standardization and greater sharing of data through its Scientific Standards Hub initiative.

In response to this RFI, AAU recommends the following:

• NIH should increase its investments in metascience and portfolio analysis to obtain additional clarity on areas of research where replication efforts will have the greatest scientific and public health impact. NIH's Pilot of the Science of Scholars Program should be expanded to include funding support and an opportunity for scholars to publish their findings.
• NIH should continue its investments in shared infrastructure such as data repositories, development of tools, and promulgation of baseline data requirements to encourage replication.
• NIH should build on its existing efforts to encourage professional recognition for reproducibility intramurally, through extramural research institutions, and collaboration with scientific societies.
• NIH should continue to incentivize replication adoption through prizes such as the DataWorks! Prize and Replication Prize to boost current replication efforts and encourage a culture change in biomedical research that results in replication activities being normalized as part of conducting research.

2. Data, Software, and Model Sharing: AAU supports NIH's goals of catalyzing collaborations and accelerating scientific impact. Data is at the core of fostering collaborations and accelerating scientific impact. The successful sharing of scientific data, software, and models is predicated on NIH maintaining its commitment to critical infrastructure support. Data support beyond the term of a grant is necessary to foster access to high-value data sets and promote data reuse.

AAU Recommends:
• NIH is encouraged to support the institutional policies and practices articulated in the AAU-APLU Accelerating Public Access to Research Data Initiative. This report discusses the strategies and approaches to establishing community data-sharing norms and expectations; developing discipline-specific data management plan templates and supporting trusted disciplinary repositories; promoting common metadata and interoperability standards; establishing long-term preservation and governance systems; and creating incentives and recognition for data sharing and data citations.
• NIH should adopt the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, already reflected in its Data Management and Sharing Policy, as part of organizing a framework for future data and model sharing indicators, and calibrate any new reporting requirements to this existing standard rather than establishing a duplicative metric structure.
• NIH is encouraged to continue collaborating with the National Institute of Standards and Technology and its efforts to develop a research data framework. Developing and adopting standard metadata approaches could facilitate the use of metadata across datasets and disciplines, reducing barriers to data sharing and reuse – as AAU shared in its comments on NIH's Enhancing Public Access to the Results of NIH-Supported Research.
• NIH is encouraged to align any changes to its Draft Controlled-Access Data (CAD) policy and proposed revision to NIH Genomic Data Sharing (GDS) policy to reduce barriers to data access. AAU's comments to NIH on its Draft Controlled-Access Data (CAD) policy and proposed revision to NIH Genomic Data Sharing (GDS) policy, submitted March 18, 2026, include additional details.
• NIH's recent Implementation Update on its data management sharing (DMS) plan was welcome. AAU shares NIH's hope that implementing this new format will ensure NIH's data-sharing goals are achieved while reducing administrative burden and improving reporting efficiency. AAU welcomes continued collaboration with NIH on reducing administrative burden.

3. Training and Mentorship: AAU encourages NIH to leverage the agency's long-established investments in mentor training and in career and professional development programs; build on established policies; and advance evidence-informed recommendations from dedicated reports. Across AAU institutions, effective mentorship is viewed as an essential component of graduate and postdoctoral training, and many institutions have developed department-level metrics and assessment processes to identify, evaluate, and strengthen mentoring effectiveness.

AAU Recommends:
• NIH should invest in scaling proven evidence-based mentoring and career and professional development within T-32s grant mechanisms and other programs. For example, the NIH Broadening Experiences in Scientific Training (BEST) funding mechanism was established to "seek, identify and support bold and innovative approaches to broaden graduate and postdoctoral training, such that training programs reflect the range of career required for a robust biomedical, behavioral, social and clinical research enterprise."
o It is notable that BEST programs are not intended to be "alternative career programs," but rather are designed to provide critical career development skills that will benefit all trainees, regardless of career trajectory.
o The program enabled 17 institutions (10 in 2013 and 7 in 2014) to build the infrastructure, curriculum, internships, and training opportunities to expose trainees to the myriad of career options in research

and associated career tracks, while emphasizing workforce sustainability through evaluation of longitudinal outcomes.
o The awardee sites were part of a comprehensive cross-site evaluation designed to understand trainee agency, time to desired careers, and culture changes at academic institutions and have led the way in recent efforts to track career outcomes. Including publishing data and additional information about their graduate programs and graduate student career placements and outcomes.
• NIH should encourage the continued use and advancement of myIDP (My Individual Development Plan). Since the development of myIDP in 2012, myIDP has emerged as an influential career-development tool to help NIH-funded graduate students and postdoctoral scholars assess their skills, interests, and values and explore scientific career options within and beyond the traditional academic pathway.
• NIH should coordinate cross-NIH efforts with established NIGMS activities to support translational pilots that integrate evidence-based practices into mentorship training. These investments could continue to provide NIH with the evidence needed to make strategic decisions about which innovations to scale.
o For example, The pd|hub Mentorship Action Collaborative is a five-year NIH NIGMS-funded project led by Professional Development Hub (pd|hub) in partnership with the Center for the Improvement of Mentored Experiences in Research (CIMER) to develop, test, and disseminate evidence-based resources that strengthen career-related mentoring for graduate students and postdoctoral scholars.
o The initiative aims to improve mentor-mentee career conversations, enhance career resilience and workforce preparation, and integrate mentorship with institutional career-development programs to support diverse biomedical career pathways.
• NIH should support assessment of mentoring effectiveness and capitalize on frameworks being developed to more holistically evaluate faculty work to include various forms of evidence that can recognize effective mentoring in promotion and annual review. For example, the TEval Network, and the American Academy of Arts and Sciences' project on Modernizing Academic Appointment and Advancement

4. Collaboration: AAU supports NIH and the National Science Foundation's (NSF) efforts to promulgate successful approaches to team science, and to explore ways to assess and recognize contributions to team-based science. Many AAU institutions have a significant history of providing structural support for collaborative science, ranging from co-locating equipment for shared departmental use, physically embedding data collection within academic hospitals to improve data access, to recognizing collaborative science in institutional tenure and promotion policies.

Documenting best practices for collaborative science and measuring contributions to team science should include entities outside biomedical research, such as NSF and NASEM. NASEM has shared extensive research in 2015 and 2025 on existing best practices and evaluation gaps for teams and individual contributors.

AAU Recommends the following to foster team science:
• NIH should consider supporting research to measure the implementation of NIH-funded team science and cross-disciplinary models utilizing tools included in NASEM's The Science and Practice of Team Science (2025). Tools exist to measure performance output, project outcomes, team dynamics, and career impact. However, "(f)unders need to provide more funding mechanisms for the scientific study of

science teams. Indeed, major funding bodies such as the National Science Foundation and National Institutes of Health emphasize the critical importance of team science in solving complex global problems, but the support needed to fund the study of science teams is lacking. This disparity is problematic given the complexity of studying science teams, which requires both significant financial resources and specialized methodological approaches tailored to scientific collaboration contexts."
• The report states that assessments of contributions to team science might vary among disciplines, invention reports, patents, and the development of tools with specialized datasets. Funders should work with the scientific community to examine approaches such as "Individual Contribution Portfolios which track specific roles using taxonomies like CRediT (Contributor Roles Taxonomy), verifiable data/software deposits, and mentorship quality rather than just authorship."
• Assessments of individual contributions could include future projects or grants following the initial collaboration, and the development of new technologies.

AAU notes that NIH's Proposal to Cap Simultaneous Research Project Grants per Principal Investigator, if implemented, would likely discourage interdisciplinary and collaborative research. As requested in AAU's comments, clarity from NIH is needed on how the proposed process for team-based and cross-institutional research grants would be measured toward possible caps on simultaneous grants. A working group like the Advisory Committee to the Director's Working Groups on Re-envisioning NIH-Supported Postdoctoral Training should examine how potential caps might affect scientific teams, interdisciplinary research, and data sharing.

5. Entrepreneurship and Translation: AAU supports NIH's goal to "translate discoveries into products for people." AAU consistently supports entrepreneurship, technology transfer, and translation as core mechanisms to convert university research into societal and economic benefits.

For 45 years, the Bayh-Dole Act ((35 U.S.C. § 200 et seq. and 37 C.F.R. Part 401) has served as biomedical research's most enduring metric of entrepreneurship. By allowing universities to retain title to inventions arising from federally funded research and requiring reporting on invention disclosures, licensing activity, and commercialization outcomes, the statute already generates the kind of measurable, longitudinal translation data this RFI seeks. Therefore, NIH does not need a new data architecture to measure entrepreneurship and translation; rather it needs to draw more fully on the invention utilization and licensing data universities already report under the Bayh-Dole Act.

A thriving Bayh-Dole ecosystem depends on three conditions that any new metrics initiative should protect rather than disturb. First, predictability. Technology transfer offices, industry licensees, and venture investors structure years-long commercialization timelines around a stable set of statutory obligations and layering new NIH-specific reporting or performance metrics on top of existing Bayh-Dole compliance risks, fragmenting that framework and discouraging the private investment the statute is designed to attract. Second, patience. Translation from discovery to product routinely spans a decade or more, and metrics that reward near-term commercialization signals over sustained investment in early-stage, higher-risk inventions will distort the discovery pipeline. Third, administrative proportionality. Universities, particularly those with smaller technology transfer staffs, already carry substantial invention reporting burdens, and NIH should build any translation indicators on existing iEdison and annual utilization reporting rather than creating a parallel collection requirement.

AAU has a long-standing position on this point and NIH has repeatedly considered and rightfully declined petitions to exercise march-in rights over drug pricing, concluding in each instance that the statute's "practical application" requirement is satisfied once a patented invention is on the market and available to the public. AAU has explained that exercising march-in rights for reasons other than exigent health or safety needs, particularly if the scope of that authority is broadened beyond Congress's original intent, would introduce substantial market uncertainty and chill university licensing activity, discouraging the private sector investment needed to carry federally funded inventions from the laboratory to the clinic. NIH's metrics initiative should reinforce, rather than complicate, the statute's core bargain: predictable ownership and licensing terms in exchange for the obligation to bring inventions to the public. AAU recommends the following on entrepreneurship and translation:

• NIH should base any future entrepreneurship and translation indicators on existing Bayh-Dole invention utilization and licensing data rather than creating new parallel metrics, and ensure any new indicators preserve the predictability, patience, and administrative proportionality on which a thriving technology transfer ecosystem depends.
• NIH should continue to support entrepreneurship education and workforce development that prepare graduate students and postdoctoral scholars to engage in innovation, commercialization, and venture creation.
• AAU would like to further engage with NIH to reduce administrative barriers to technology transfer by streamlining research contracting, intellectual property management, and public-private partnerships.
• NIH should continue to support and expand innovation/entrepreneurship competitions and comprehensive entrepreneurial ecosystems that include incubators, accelerators, and commercialization and proof-of-concept grants through its SBIR and STTR programs and other innovative initiatives such as the Research Evaluation and Commercialization Hubs (REACH) and the NIH Catalyze programs established by the National Heart, Blood, and Lung Institute.
• AAU recommends that NIH, and the federal government more broadly, protect and strengthen the Bayh-Dole framework that allows universities to retain and commercialize inventions arising from federally funded research, and continue to support university startups and small businesses through Small Business Innovation Research and Small Business Technology Transfer programs and other entrepreneurship-focused funding mechanisms.
• Any framework for measuring and rewarding scientific impact, at NIH or elsewhere in the federal government, should treat that decentralized ownership model as an asset to be strengthened rather than a variable to be re-engineered on an agency-by-agency basis.

• AAU also urges continued investment in SBIR and STTR programs, which remain among the most effective vehicles for converting federally funded discoveries into viable companies. These award-based mechanisms complement Bayh-Dole by supplying the early-stage, non-dilutive capital that allows university spinouts and small businesses to de-risk technologies before they can attract private investment. As other countries deploy substantial state-directed capital to accelerate commercialization of their own research base, sustained and predictable funding for SBIR, STTR, and related entrepreneurship-focused mechanisms across NIH and other federal science agencies is a matter of national competitive urgency, not simply a program-level budget line.
• More broadly, NIH should encourage research institutions to continue to draw on recommendations from PTIE (Promotion & Tenure Innovation & Entrepreneurship), a national initiative that promotes

recognition of innovation, entrepreneurship, commercialization, and other societal impacts within faculty promotion and tenure systems. PTIE provides guidance on aligning academic incentives with public impact by rewarding activities such as technology transfer, startup creation, industry engagement, translational research, and broader societal contributions, alongside traditional publications and grants.

6. Foundational Scientific Exploration: The research ecosystem and academic enterprise depend on talent and talent development. Paradigm-shifting breakthroughs from foundational scientific exploration and high-risk, high-reward research rely on consistent, sustained support for workforce training and development.

Continuing federal investment in the full spectrum of STEM disciplines alongside the critical technologies, recognizing that foundational research in physics, chemistry, biology, computer science, engineering, geosciences, social, behavioral and economic sciences, and mathematics enables breakthrough innovations across all technological areas. Sustained investment in training and technologies is the bedrock of successful scientific exploration. Foundational science is the substrate on which technological and scientific progress is built to produce knowledge as a public good, and it remains largely federally funded.

Many AAU institutions have broadened their respective faculty policies over the past decade to recognize and encourage research engagement beyond the historic “publish or perish” model. Broadened policies encourage graduate students, postdoctoral scholars, and faculty to consider undertaking higher-risk research with institutional support for longer commitments.

AAU recommends for foundational science:
• NIH should continue its High-Risk, High-Reward program. The program recognizes established and early-career investigators and continues to serve as a strong model for academic institutions.
• NIH should continue its engagement with the scientific community to explore options for funding support to encourage community- and policy-engaged research.
• NIH should monitor its Center for Scientific Review’s recent policy realignments to ensure the process emphasizes scientific merit and reduces reputational bias.

7. Public Impact: AAU supports NIH’s efforts to explore approaches for assessing and recognizing the public impact of research. Developing such assessments is inherently complex. In biomedical research, for example, the pathway from a scientific discovery to widespread clinical adoption often spans decades. As a result, equating measures of public impact solely with direct contributions to near-term health outcomes could unintentionally discourage investment in foundational research that underpins future breakthroughs.

History demonstrates that transformative advances often emerge from long-term, fundamental scientific inquiry. Longitudinal studies, basic science investigations, and discoveries in seemingly unrelated fields frequently lay the groundwork for major medical innovations. More than 75 years after its inception, the Framingham Heart Study continues to shape evidence-based approaches to cardiovascular health. Similarly, NIH-funded research on mRNA technology required more than two decades of sustained scientific progress before enabling the rapid development of vaccines, while advances in HIV research contributed to the emergence of breakthrough cancer immunotherapies. Scientific discoveries often

generate benefits well beyond their original field of inquiry, underscoring the importance of recognizing the broad and often indirect pathways through which research creates public value.

Many of the necessary processes for a particular discovery to yield "public impact" involve factors beyond NIH's scope. AAU supports efforts to forge stronger connections among universities, industry, and government to accelerate discovery, innovation, and solutions to major societal challenges. The National Science Foundation studies the economic and societal impacts of scientific funding, including biotechnology and biomedical-adjacent fields. The Agency for Healthcare Research and Quality conducts implementation science. The Center for Medicare and Medicaid Innovation assesses health care delivery and payment reforms that can incentivize widespread adoption of clinical practices.

AAU recommends the following on public impact:
• NIH should explore elevating reports and national networks aimed at advancing public impact research, such as highlighting policies and practices advanced by Advancing Research Impact in Society (ARIS), a national organization that strengthens the research ecosystem to support meaningful real world impact, including public impact, by fostering Broader Impacts (BI), Research Impact (RI), and Societal Impact (SI) across all disciplines.
• NIH could consider elevating pathways modeled after institutional programs in the spirit of projects selected for ARIS Impact Innovation Award, such as Penn State's Evidence to Impact Collaborative. In addition, these projects foster transdisciplinary research, research translation, and reducing barriers between basic research and application.
• NIH should consult recent research by the Pew Charitable Trusts that highlights the need to expand faculty reward systems so that promotion and tenure processes value societal impact alongside scholarly impact, including research that informs policy, engages communities, advances innovation, and addresses real-world challenges
• Public impact indicators should be designed as distinct from, not duplicative of, the entrepreneurship and translation indicators addressed elsewhere in this response. Commercialization success, measured through licensing revenue, startup formation, or patent counts, captures the pathway by which a discovery reaches the public, not whether the public actually benefited from it. Treating the two categories interchangeably risks rewarding research for its commercial traction rather than its demonstrated contribution to health outcomes and could create conflicting incentives within NIH's own metrics framework.
• NIH should also account for the fact that public impact often depends on institutional infrastructure, including data management, regulatory compliance, and grants administration capacity. AAU recommends NIH define public impact indicators separately from entrepreneurship and translation metrics to avoid conflicting incentives. NIH should elevate recommendations in the American Academy of Arts & Sciences series of ARISE (Advancing Research in Science and Engineering) initiatives aimed at organizing research to more effectively translate discoveries into advances in health, energy, environment, security, and economic competitiveness. NIH's ability to heighten community awareness of existing recommendations could encourage the cross-sector collaborations necessary to operationalize advances.

8. Other Areas: Measuring and rewarding scientific innovation involves a complex ecosystem that extends well beyond federally sponsored research. While NIH is a critical participant in that ecosystem, other federal sectors, including the National Science Foundation, the U.S. Department of Energy, and the

White House Office of Science and Technology Policy, also play a role. The impact of any federally proposed scientific metrics and incentives has ramifications well beyond the biomedical sector.

AAU suggests NIH sponsor conversations to collaborate with the National Academies of Sciences, Engineering, and Medicine (NASEM) and the American Academy of Arts and Sciences, scientific societies, and associations that establish professional standards, academic institutions that conduct research and train future scientists, and industry leaders to ensure fulsome discussions regarding potential scientific metrics and incentives.

AAU also suggests that the agency consider two additional levers to incentivize the cultural changes needed to achieve its objectives through existing and newly created funding programs: 1) build flexibility into grant programs to encourage behaviors that promote the broader societal and public impact of research, similar to the NSF's Broader Impacts. Such incentives could include enhanced mentoring, improved instruction, greater data access and sharing, entrepreneurship, public engagement, and related activities: and 2) award grants to entire academic departments, rather than individual investigators, to pilot and expand incentives and faculty reward structures in areas outlined in this RFI. We believe that both recommendations would help achieve the desired outcomes by encouraging institutional change, developing more meaningful metrics, and advancing the priorities identified in this RFI.

Finally, as new metrics are promulgated by NIH, AAU requests that a thoughtful consensus process be undertaken to ensure the feasibility of, and mechanisms for, federal reporting. Further, AAU requests that the current funding climate be integrated into consideration of any new mandatory metrics and incentives. The current research and teaching climate is challenged by delays in Congressional appropriations and agency allocations, declining trends in graduate and PhD admissions, driven in part by research funding uncertainty, changes to federal student loans, and concerns about potential new grant requirements under OMB's Uniform Guidance.

Thank you for the opportunity to share AAU's views on Metrics and Incentives for Scientific Innovation. AAU greatly appreciates NIH's efforts to engage the biomedical research community, and we welcome further collaboration. If there are questions about AAU's comments, please contact Lizbet.Boroughs@aau.edu.

275. American Urological Association (AUA)

I am responding: On behalf of an organization

Name: Vanessa Cramer

Name of Organization: American Urological Association (AUA)

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Manager

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/AUA_NIH-comment-letter_FINAL.pdf

American Urological Association

BOARD OF DIRECTORS

Officers

Eugene Y. Rhee, MD, MBA
President

Thomas F. Stringer, MD
President-elect

Lane S. Palmer, MD
Immediate Past President

David F. Penson, MD, MPH
Secretary

Victor W. Nitti, MD
Secretary-elect

Jennifer U. Miles-Thomas, MD, MBA
Treasurer

SECTION REPRESENTATIVES

Arthur L. Burnett II, MD, MBA
Mid-Atlantic

Adam S. Kibel, MD, MHCM
New England

Jay A. Motola, MD
New York

Tobias S. Köhler, MD, MPH
North Central

Ronald P. Kaufman Jr., MD
Northeastern

Tomas L. Griebling, MD, MPH
South Central

Glenn M. Preminger, MD
Southeastern

Christopher R. Porter, MD
Western

Headquarters

Michael T. Sheppard, CPA, CAE
Chief Executive Officer

1000 Corporate Boulevard
Linthicum, MD 21090

U.S. Toll Free: 1-866-RING-AUA (1-866-746-4282)
Phone: 410-689-3700
Fax: 410-689-3800
Email: AUA@AUAnet.org
Websites: AUAnet.org
UrologyHealth.org
UrologicHistory.museum

Advancing Urology®

August 19, 2026

Jayanta "Jay" Bhattacharya, M.D., Ph.D
Director, National Institutes of Health (NIH)
Office of Science Policy
9000 Rockville Pike
Bethesda, MD 20892

RE: Request for Information on Measuring and Rewarding Scientific Impact
Notice Number: NOT-OD-26-087

Dear Dr. Bhattacharya:

The American Urological Association (AUA) commends the National Institutes of Health (NIH) for issuing RFI NOT-OD-26-087 to modernize how scientific impact is recognized and incentivized and appreciates the opportunity to provide comments.

The AUA is a leading advocate for the specialty of urology and has more than 24,000 members throughout the world. Our members represent the world's largest collection of expertise and insight into the treatment of urologic disease. Of the total AUA membership, more than 17,000 are based in the United States and provide invaluable support to the urologic community by fostering the highest standards of urologic care through education, research, and the formulation of health policy.

As the world's largest urological association, the AUA recognizes that traditional publication and citation metrics fail to adequately reflect the value of translating research into patient care. The focus areas identified by NIH in the RFI reflect the increasingly interconnected nature of the biomedical research enterprise. Rigor and reproducibility, data and resource sharing, training and mentorship, collaboration, translation, foundational discovery, and public impact should not be viewed as discrete activities, but rather as mutually reinforcing components of a research ecosystem that advances knowledge and ultimately improves human health.

Below are the AUA's comments on select focus areas:

FOCUS AREA **– Rigor and Reproducibility.** *Reproducible and replicable biomedical approaches are foundational to rigorous research. New strategies are needed to identify areas ripe for replication and reproducibility efforts, as well as incentives and infrastructure to support the reproduction and replication of existing findings. NIH also seeks comment on rigor, including with regard to defining causal research.*

Both NIH and the Congressionally Directed Medical Research Programs (CDMRPs) at the Department of Defense require Research Sharing Plans in their grants applications. These planning documents explain how investigators will make data and resources

AUA 2027 San Diego MAY 21-24
AUAnet.org/AnnualMeeting

accessible to the public, government, and scientific community. It must detail what data is shared, when it is shared, and justify any exceptions.

A robust data management and sharing plan should clearly describe how final research data, software, model organisms, research tools, and materials will be made accessible to the scientific community while adhering to applicable legal, regulatory, and ethical requirements. The plan should also outline measures to protect participant privacy, proprietary commercial information, and third-party intellectual property. When data or resources cannot be shared, investigators should provide clear justification, such as limitations related to cost, feasibility, confidentiality, or intellectual property considerations.

Application submission and tracking begin through the eBRAP system for pre-applications and are later processed through Grants.gov for full proposal submission. Data management and sharing plans are expected to align with the FAIR principles, ensuring that research data are Findable, Accessible, Interoperable, and Reusable in accordance with federal and defense-funded medical research standards.

FOCUS AREA **– Training and Mentorship.** *Indicators to quantify and reward the effort put into stabilizing and growing the biomedical workforce through effective mentoring.*

Training and mentorship serve as critical tools essential to sustaining the future of the biomedical research workforce in the United States. The ability to quantify research training program impact has been traditionally measured through individual-focused metrics such as publications, citations, and receipt of federal research funding of a mentor's subsequent mentees. As the landscape of biomedical research and clinical care delivery continues to rapidly evolve, respective evolution in the way we assess our ability to educate, train, and equip the next generation of clinician-investigators and scientists is paramount.

Firstly, valuable transformations to the current model should incorporate emerging research outputs that have not been captured and considered in the traditional model. These outputs are well established and well-regarded avenues for research productivity in the modern healthcare ecosystem. Potential new indicators to quantify and reward effort put forth into effective mentoring including measurement of mentee activities including:

- Unfunded research collaborations with patient advocacy and support networks
- Initiation of unfunded multi-disciplinary research collaborations
- Participation in population health research initiatives
- Participation in prevention health research
- Investigator initiated studies and clinical trial participation
- Private industry, foundation or professional society research funding

A second prospective method to enhance the ability to evaluate research mentoring impact could involve design and implementation of a research skill milestone

assessment for distribution for research mentees participating in federally funded research training and education programs such as NIH F32, R13, R25, T35, T32 programs. This would allow for standardized feedback from mentees on progression along landmark research milestones designated by leading research education experts. This concept stems from the foundational theories of clinical competency measurement implemented by the Accreditation Council for Graduate Medical Education.

Unintended ramifications of implementing some of these proposed changes could include a deemphasis on peer-reviewed publication in established scientific journals and degradation of the scientific peer-review process outside of federal funding mechanisms. These unintended consequences could be mitigated by continuing to incorporate the landmark pillars of peer-reviewed publication and federal funding in research training and mentorship evaluation and focusing on expanding, not removing, traditional measurement metrics.

The RFI notes that NIH is particularly interested in specific examples and to that end, the AUA offers the following snapshot:

- Six out of seven of the AUA's Office of Research grant mechanisms require mentorship, beginning with medical students and continuing through the career continuum to early-career investigators.
- For some awards, the AUA requires mentors to submit a List of Previous Mentees as our peer reviewers want to see the success of the mentor's previous trainees (e.g., remain in academic positions, success in securing independent funding).
- For some awards, the AUA weighs the mentor equally or more so in evaluation of the application. Some of our grants are truly about mentorship as opposed to the research project, supporting the building of a successful urologist and physician scientist workforce.

In addition, the AUA supports physician scientists to remain in research with the RSTAR mechanism. This award provides salary support to clinicians with career development awards to encourage them to be physician scientists. We also have a recognition award, the Distinguished Mentor Award, to honor longstanding, successful mentors of the AUA.

FOCUS AREA **– Collaboration.** *New frameworks for recognizing and valuing contributions to large-scale, interdisciplinary team science.*

In contemplating this focus area, the AUA offers lessons learned from our Office of Research's solicitation of feedback from members on current research funding mechanisms and what types of research they would like to see.

The results of this survey led in part to the development of the upcoming Urology Care Foundation (UCF) Transformative Research Award (TRA). This new award mechanism provides two years of funding to support high-risk, paradigm-shifting, and collaborative

research. One of the ways that applicants can do this is through cross-disciplinary collaborations with other fields, such as gynecology, neurology, cardiology, and radiology. The AUA's other funding mechanisms all require descriptions of the research team, including important roles such as biostatisticians and subject matter experts in other fields when warranted.

Drawing on these lessons, the NIH could incrementally strengthen team science by incorporating cross-disciplinary collaboration and team dynamics into grant review criteria, placing greater emphasis on a project's potential to drive scientific advancement than on preliminary data alone. Additionally, Research Performance Progress Reports (RPPRs) could shift from focusing primarily on publication counts to highlighting team-based outcomes, including shared data resources, collaborative intellectual property development, cross-disciplinary partnerships, and mentorship contributions.

FOCUS AREA **– Entrepreneurship and Translation.** *Incentives for researchers moving discoveries from bench to bedside.*

In contemplating this area of focus, the AUA offers lessons from its Innovation Nexus program. The AUA Innovation Nexus is a year-round incubator and accelerator program launched by the AUA to support urology startups. It connects innovators, clinicians, researchers, investors, and strategic partners to move ideas from concept to commercialization.

Grounded in operational insights from the Office of Research and the Nexus Program, the AUA has determined the following specific indicators, implementation strategies, equity considerations and risk mitigation frameworks to assist the NIH in establishing modern evaluation standards.

- The NIH should expand research impact metrics beyond traditional publication measures to include objective indicators of translation and commercialization. These could include advancement along Technology Readiness Levels (TRLs), demonstrating progress from early concept validation to clinical prototype development; intellectual property and commercialization outcomes, such as patent filings, granted patents, and licensed technologies; de-risking and capital formation milestones, including participation in translational training programs, investor showcases, SBIR/STTR funding, and private investment; and regulatory and clinical achievements, such as FDA IDE or IND approvals, clinical trial initiation, and early implementation within healthcare systems.
- The NIH can integrate recognition of translational research into existing funding and review processes with minimal administrative burden. This could include adding an optional "Translational and Entrepreneurial Outcomes" section to the biosketch, allowing investigators to highlight patents, licensed technologies, startups, regulatory approvals, and commercialized products alongside traditional publications.

- Peer review criteria could also place greater emphasis on translational progress by evaluating technology de-risking and commercialization potential within the Impact and Approach scoring categories.
- A modernized evaluation framework should account for differences in career stage and institutional resources. For early-career investigators, translational and commercialization achievements can provide meaningful evidence of impact before long-term citation records are established.
- The framework should also recognize that researchers at resource-limited institutions may have less access to technology transfer and commercialization infrastructure. National specialty society programs, such as the AUA Innovation Nexus, can help level the playing field by providing mentorship, investor access, regulatory guidance, and industry connections that support innovation regardless of an investigator's home institution.

Potential drawbacks of expanding translational research metrics should be addressed through thoughtful safeguards. Increased emphasis on entrepreneurship may create real or perceived conflicts of interest, which can be mitigated through standardized disclosure requirements and by separating scientific review from commercial evaluation. Finally, while translational impact should be recognized, these metrics should be applied primarily to applied and translational research programs, ensuring that foundational basic science continues to be evaluated based on its potential for scientific discovery and long-term impact.

FOCUS AREA **– Foundational Scientific Exploration.** *Safely embedding support for high-risk, high-reward basic research, which carries a risk of failure but can yield massive breakthroughs.*

The aforementioned TRA is also relevant here, as the AUA is encouraging researchers to think about old urology problems and develop new tactics to address them through this mechanism.

FOCUS AREA **– Public Impact.** *Novel approaches to measure a study's broader societal footprint, such as its effect on actual health outcomes, public trust, accessibility, and clinical practice.*

A potential drawback to these new metrics is that different communities require different outcomes. For example, rural urology patients would be interested in research that measures the number of patients treated or helps to increase the number of physicians and their access to those patients. However, populations in more urologist-dense areas of the country may benefit from more subtle metrics, such as the specific skills of a physician or hospital system, and both would benefit from patient education improvements. Confining all research within one area of urology to the same metrics might allow for direct comparisons but could inadvertently fail to show the benefits of some of the projects to certain populations. There are also no universally agreed upon metrics on things such as public trust, accessibility, and clinical practice.

American Urological Association

Notably, UCF Humanitarian grants require an estimate of how many people were reached/impacted through this work as a deliverable. For AUA research grants, we ask the following of our awardees in their progress reports:

- What is the impact or potential impact of the results of the project for the practice of urology and for urology patients?
- Describe how findings, results, and/or techniques that were developed or other products may make an impact on the base of knowledge in urology. Summarize using language for the lay audience.

***

The AUA thanks NIH for the opportunity to comment on this important RFI and welcomes continued engagement.

Sincerely,



Mark T. Edney, MD, MBA
Chair, Public Policy Council, American Urological Association (AUA)



H. Henry Lai, MD
Chair, Research Council, American Urological Association (AUA)

276. B. Ian Hutchins

I am responding: On behalf of myself

Name: B. Ian Hutchins

Name of Organization: University of Wisconsin-Madison

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: It is now possible to flag areas of science at high risk of retractions before they occur. Individual topics have structural vulnerabilities to cascading errors. These are measurable based on the network structure of papers, citations, authors, and retractions. A model built on detecting these vulnerabilities predicts a topic's future retraction rate (Pearson r = 0.714). These forecasts can help target the limited resources for reproducibility research into the areas where they are most needed. The attached response describes these vulnerability measures and predictive model in more detail.

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: We can identify which papers are likely to inform future innovation and patenting, years in advance of the decade long path from paper to patent. Models using trajectories of knowledge flow can segment scientific articles into groups where those with the highest prediction scores are patented at ten times the rate of the bottom group. As an article-level metric, it can be used for program evaluation, portfolio analysis, and policy forecasting. The attached response describes the model and its time-series performance.

6. Foundational Scientific Exploration: We can identify signals of nascent fields that will lead to a breakthrough, before landmark studies are published. Network signals in the scientific literature, such as scientists flocking into a new topic and network dynamic instability, often precede the breakthrough itself. Two models, one developed by NIH and one from an independent research group, validate this approach. These signals can measure progress toward breakthroughs and could inform funding toward the most promising scientific topics. The attached response details both models and further use cases.

7. Public Impact: It is now possible to forecast which lines of biomedical research will yield an FDA-approved therapy roughly a decade before approval, in most cases before the pivotal phase 2 study begins. Recently published prediction models accomplish this with an F1 scores of 0.75-0.84. These forecasts can measure portfolio progress toward this applied health impact, while there is still time for decisions to make a difference. The attached response details the modeling approach in more detail.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/RFI_NIH_Impact_2026_Hutchins.docx

Description: Detailed response from B. Ian Hutchins at the University of Wisconsin-Madison on the subjects of Rigor and Reproducibility, Entrepreneurship and Translation, Foundational Scientific Exploration, and Public Impact

From: B. Ian Hutchins, Ph.D., Assistant Professor, Information School, University of Wisconsin-Madison

To: Office of Science Policy, Office of the Director, National Institutes of Health

08/19/2026

Recent advances have made it possible to forecast the impact of biomedical research in advance. Researchers have created methods to forecast which lines of research will lead to novel therapies a decade before approval, which nascent fields are most likely to lead to transformative breakthroughs, what areas of science are headed for a wave of future retractions, and which papers will fuel downstream innovation and patents. This response highlights four cutting-edge methods that accomplish these goals and align with the impact areas requested in this RFI.

This is not the first time that I have developed new methods for measuring biomedical research impact. My previous measures of scientific influence (the Relative Citation Ratio) and early indicators of clinical translation (the Approximate Potential to Translate) are now built into the iCite and iSearch tools that program staff use regularly. The methods highlighted below are extensions of that work into some of the harder questions highlighted in this RFI.

All four make extensive use of prediction for a structural reason: there is a fundamental disconnect between agency decision-making timelines (typically 4-5 years) and the long timelines for these decisions to bear fruit (10+ years for innovation and patents, 15-20 years for therapeutic development, and longer for foundational basic science breakthroughs). In addition to identifying the most promising research investments, prediction is necessary for measures that indicate whether research initiatives are moving in the right direction within the necessary decision-making timeframes at NIH. **The payoff from this year's funding will not be directly visible for 10-20 years, so prediction is how to measure it within a decision cycle**.

While I do not provide explicit policy recommendations, I hope to present valuable information and resources that can help NIH to measure the impact of biomedical research investments.

Executive Summary

1. It is now possible to forecast which avenues of research are most likely to yield FDA-approved therapeutics. Our recently published machine learning method [1] predicts which lines of research into human disease will have novel therapies approved, approximately a decade in advance, forecasting future **Public Impact** of biomedical research investments.
2. Recent advances in **Foundational Scientific Exploration** methods have made it possible to identify scientific topics that are poised for breakthrough discoveries. NIH's own model [2] predicts breakthrough discoveries in topic areas that later culminate in Nobel Prize-

Information School
University of Wisconsin–Madison 330 N. Orchard St. Madison, Wisconsin 53715
608-262-5672 Email: bihutchins@wisc.edu hutchinslab.github.io

winning studies; another model [3] using a similar general approach forecasts exceptional growth of research topics, a phenomenon correlated with future breakthrough discoveries.

3. **Rigor and Reproducibility** initiatives have drawn an equivalence between transparency and rigor; my lab has pursued a complementary avenue of attempting to identify areas of science that are particularly at risk for having articles retracted in the future. We have identified vulnerability measures that flag which areas of science have elevated risk of unreliable information transmission through a combination of concentrated author-citation networks and previous retractions [4]. These have enabled us to generate a predictive model that estimates the future rate of retraction in different scientific topics—a modeling approach that might help target limited reproducibility resources to areas where it would be most valuable.
4. By leveraging artificial intelligence approaches, our lab has developed **Entrepreneurship and Translation** forecasting tools that identify which papers are most likely to be used and cited in patents granted by the USPTO [5]. These models identify patterns of cross-domain knowledge flow that typically precede patents.

Public Impact

We can predict an FDA-approved therapy a decade in advance

The development of novel therapeutics is a lengthy process, typically taking 15-20 years, that draws from fundamental research discoveries to inform downstream innovation and applied clinical work. In principle, targeted efforts by the scientific community and science funders could accelerate the discoveries that drive therapeutic development. In practice, it is challenging for scientists, funders, and pharmaceutical companies to identify the most promising research avenues at an early stage.

It is now possible to predict which research will yield an FDA-approved therapy a decade in advance, most often before the relevant phase 2 trial has begun. In order to identify which lines of research will yield future FDA-approved therapeutics [1], we took a three-stage approach: we (a) pruned the entire biomedical knowledge graph of paper-to-paper citations, then (b) partitioned this using machine learning into time-series topic clusters, and (c) generated a predictive model to forecast, based on historical data, which of these topics would later contain the pivotal papers used in FDA decision-making to approve a new therapy. This approach gives models with an F1 score of 0.84 and predicts 65% of therapies before their phase 2 trials have even been initiated (Figure 1).

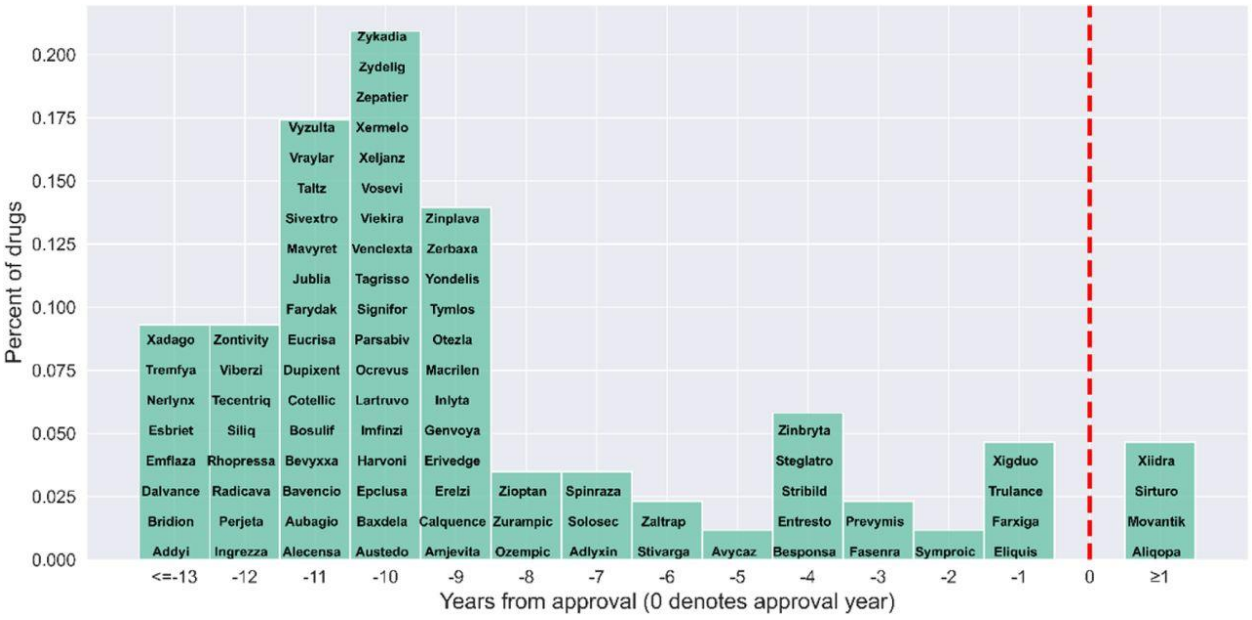


Figure 1. Drug approval forecast that were detected before FDA approval (left of dotted line) as well as the number of years pre-approval that each therapy was predicted [1].

This type of modeling approach gives information early enough that decisions about which research investments are the most promising can still be made, while there is still time to accelerate discovery. In addition, such probabilistic forecasts can be used as an early measure of whether a given part of the research portfolio is likely to yield this type of applied Public Impact outcome in the foreseeable future.

Foundational Scientific Exploration

We can identify a transformative breakthrough while it is still forming

Only a small subset of basic research ideas ultimately goes on to make major transformative breakthroughs. Most such breakthroughs are not immediately recognized even by peers—it is only after broad adoption by the community that recognition eventually accrues. However, the process by which this occurs, involving substantial time, effort, and resources to vet an idea and convert it to broad practice, leaves signatures in the scientific knowledge network. Some of these are quantifiable before landmark studies appear, and indicate the active preparation of a nascent field toward a breakthrough.

The NIH model predicting future transformative breakthroughs uses co-citation networks to partition the literature into time-series topics [2] (financial disclosure: I am a co-inventor on patent no. 12,536,216 covering this methodology). This approach sharpens the signals of future breakthrough that are embedded in the knowledge network compared to the direct-citation approach used in our FDA prediction study. In fact, as in the cancer immunotherapy example below that won a Nobel Prize, network signals like the flocking of scientists into a new scientific topic or high levels of dynamic instability precede the prize-winning work itself (Figure 2). This means that we can identify a transformative breakthrough while it is still forming.

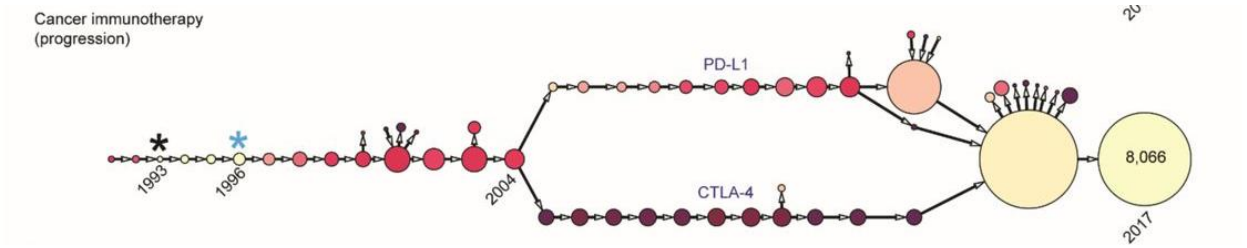


Figure 2. Time series of co-citation network topic clusters for cancer immunotherapy breakthrough [2]. First signal detection in 1993 (black asterisk), seminal paper appeared years later (blue asterisk). This lineage split into distinct research trajectories that covered the PD-L1 and CTLA-4 proteins before re-merging into a large cancer immunotherapy cluster concomitant with the validation of these lineages for clinical use.

Another independent research group has used a similar approach to detect a phenomenon related to breakthroughs: extreme growth of research topic clusters [3]. This is based on the observation that research topics derived from citation knowledge networks typically undergo large growth spurts alongside their emerging policy-relevance. This identification was one of the goals of the Foresight and Understanding from Scientific Exposition initiative, and this type of modeling approach is very effective at forecasting such extreme growth (Figure 3). While this modeling approach was not as tuned to biomedical research as the NIH model, it performed well across diverse areas of science including those outside biomedicine.

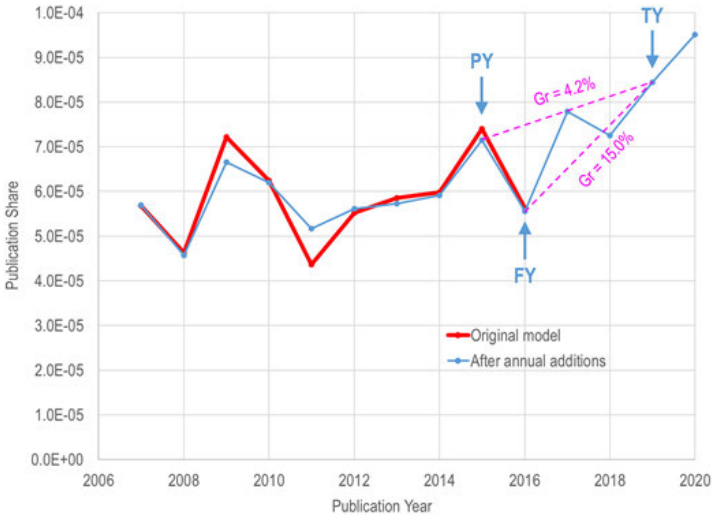


Figure 3. Forecasting extreme growth of topic clusters [3]. Target Year (TY) forecasts made from information available in the Forecast Year (FY) predict extreme topic growth with high accuracy.

There are three new possibilities raised by these studies. First, contributions to emerging breakthroughs in the existing scientific portfolio can be measured with these approaches. Second, methods to identify breakthrough foundational science can enable funding to be delivered to the most exciting scientific opportunities if the signals are used in decision-making. Finally, the fact that there are multiple distinct modeling approaches enables a direct test of a prevailing hypothesis among funders: that targeted funding can rapidly accelerate breakthroughs. If a randomized, controlled trial approach is employed, the effectiveness of targeted funding at accelerating science can be measured using one model to detect breakthrough signals, and the other to measure the ongoing effects of targeted funding in that randomized design.

Rigor and Reproducibility

We can detect which areas of science are headed for a wave of retractions

Much of the recent effort on scientific rigor and reproducibility has focused on ensuring transparency in scientific articles. This is valuable, but transparency alone does not guarantee rigor; studies can be flawed and transparent at the same time. Retractions are a highly visible marker of flawed studies, and the community has rightly highlighted the risk of cascading errors through subsequent use and citation of retracted works. Despite this awareness, prior work has not measured the systematic vulnerabilities in the literature to such cascading failures.

Our recent work developed network science methods to assess two particular vulnerabilities arising from retracted articles [4]. The first draws on the observation that individual scientists tend to publish multiple articles that are later retracted, either due to research misconduct, inadequate methods, or insufficient supervision. Scientific topics within a field, if sufficiently exposed through citation connections to a body of work from an individual scientist whose work is later retracted, may in effect be building a house of cards. In addition, topics represent the collective effort of scientists with similar goals, research norms, and common analytical approaches. Because of this, vulnerabilities inherent to these collective properties may be measured empirically as concentrated retractions occurring within individual topics independent of multiple-retraction author risk.

Our recent analysis employing this framework shows both vulnerabilities: at the extremes, research topics within ten fields of biomedical research that we examined had between 39% and 94% of their papers citing work from an individual author, making the topic extremely vulnerable if that author's work should later come into question. We similarly observe a wide variance in the percentage of topic papers citing retracted works. This is partly due to multiple retractions from individual authors, but much of the effect remains even if multiple retractions per author are removed, suggesting other vulnerabilities at play [4]. Building on this analytical framework, we have developed a machine learning model to predict what fraction of articles within a topic will be retracted in the coming years (Figure 4). This shows that **we can detect which areas of science are headed for a wave of retractions before they happen**.

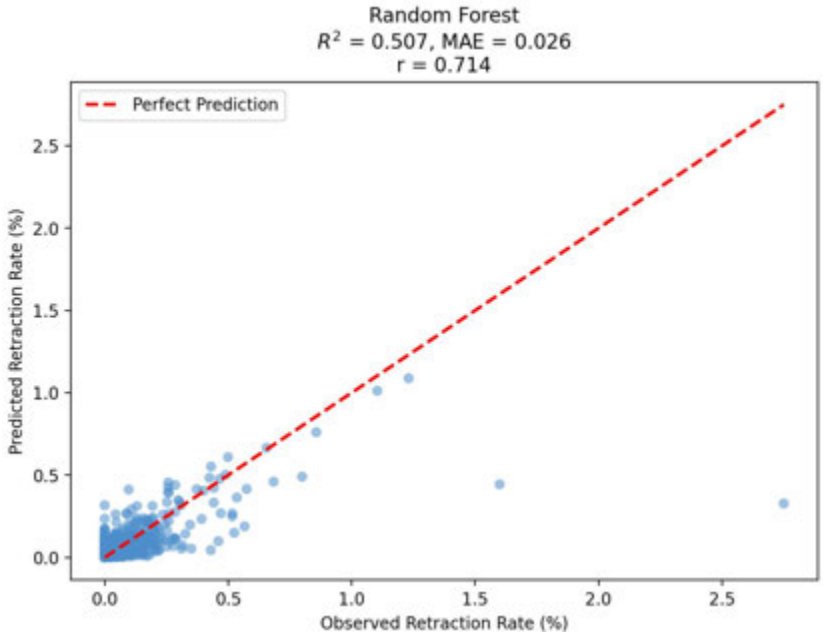


Figure 4. Machine learning predicts the future retraction rate of papers within topic clusters with high accuracy (each point, one topic cluster).

This is a highly accurate prediction system (Pearson $r = 0.714$) that may be valuable for identifying topics that are at high risk of having papers retracted in the future. Of course, determining the reproducibility of high value discoveries is an important complementary approach. However, reproducibility research is expensive, and this type of approach could be used to further target reproducibility efforts: looking at the intersection of high value discoveries that are also uniquely vulnerable.

Entrepreneurship and Translation

We can tell which papers are going to fuel future patents

Only a minority of papers are ever used and cited in a patent. Furthermore, unless the authors of a paper reduce their discoveries to practice themselves, this process takes on the order of a decade to translate. Building on my prior work developing the Approximate Potential to Translate measure, we have combined content analysis of scientific articles with network analysis and machine learning to predict which papers are most likely to contribute to future innovation and patents. Our machine learning model takes advantage of a novel property of scientific articles: the pattern of which types of papers cite a given article and how quickly that happens carries significant predictive power about whether a patent building on this work is likely in the future. Our model [5] shows that, between articles in the group with the lowest machine learning prediction scores and those with the highest, there is a tenfold difference in the observed number of patented papers in the subsequent years (Figure 5).

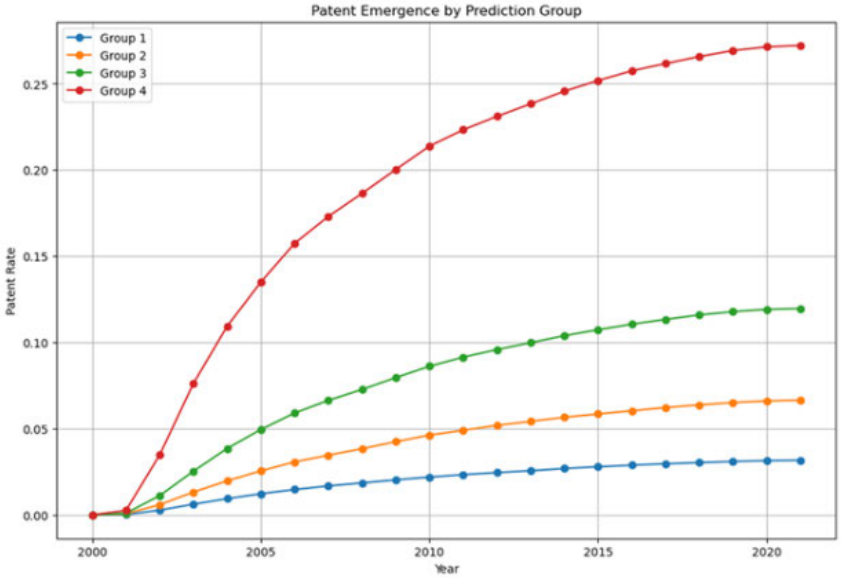


Figure 5. Observed fraction of patented papers, stratified by groups with differing patent prediction scores (group 1 lowest, group 4 highest).

We can tell which papers are going to fuel future patents and which are not: the top group translates to ten times as many patents per paper as the bottom group. Being able to forecast which discoveries are likely to contribute to downstream innovation would be a valuable impact measure. Like other article-level measures that I have developed in the past, this type of measure can be used in program evaluation, portfolio analysis, and policy forecasting.

Analytical Use

As this RFI acknowledges, the value of biomedical research is multifaceted, and the community needs multiple measures to reflect that. One common feature of many of the measures described here is that they examine scientific topic clusters rather than individual papers or grant applications. This is because many of the questions posed in this RFI are more complex than simpler measures of research success like number of papers or citation impact. Development of novel therapeutics and emerging scientific breakthroughs are fundamentally collective efforts of groups of researchers working toward common goals. These are best captured by larger clusters of research of the types used in these studies [1-3]. Likewise, reproducibility (or research integrity and retraction risk, a negative facet of that concept) is in part due to the research norms and common analytical approaches in these communities [4]. This cannot be fully captured by looking at individual papers, but rather by examining the surrounding scientific context as well. Importantly, capturing this context does not require new data collection. Existing NIH data infrastructure can be easily adapted to integrate this topic cluster-level information and map it onto the appropriate papers and applications. That is because the links needed to map topic clusters to these documents are already present in their reference lists. NIH has already prototyped topic cluster analysis [2], so this could be readily incorporated into the iCite or iSearch platforms, as the Relative Citation Ratio and Approximate Potential to Translate have.

References

1. Arabi S and Hutchins BI (2026) Forecasting novel therapeutic development in biomedical research. *Communications Medicine*. doi: 10.1038/s43856-026-01838-6
2. Davis MT, Busse BL, Arabi S, Meyer P, Hoppe TA, Meseroll RA, Hutchins BI, Willis KA, and Santangelo GM (2025) Prediction of transformative breakthroughs in biomedicine. *bioRxiv*. doi: 10.64898/2025.12.16.694385 US Patent no. 12,536,216
3. Boyack KW and Klavans R (2022) An improved practical approach to forecasting exceptional growth in research. *Quantitative Science Studies*. doi: 10.1162/qss_a_00202
4. Pan Z and Hutchins BI (2026) Quantifying structural vulnerabilities and resilience to research integrity risks in biomedical research networks. *bioRxiv*. doi: 10.64898/2026.06.26.734852
5. Yu B and Hutchins BI (2026) Knowledge in Transit: Cross-Domain Citation Flow Predicts Patent Translation. *International Conference on Science of Science and Innovation*.

277. Physicians Committee for Responsible Medicine

I am responding: On behalf of an organization

Name: Mikalah Singer, JD, LLM

Name of Organization: Physicians Committee for Responsible Medicine

Type of Organization: Non-Profit Research Organization

Role: Other

Role - Other: Science policy

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Physicians-Committee-NOT-OD-26-087-Impact-RFI-2026-08-19.pdf

PhysiciansCommittee
for Responsible Medicine
PCRM.ORG
5100 Wisconsin Ave. NW, Suite 400 • Washington, DC 20016 • Tel: 202-686-2210 • Fax: 202-686-2216 • pcrm@pcrm.org

August 19, 2026

Re: NOT-OD-26-087. Request for Information on Measuring and Rewarding Scientific Impact

Submitted via comment form: https://osp.od.nih.gov/comment-form-measuring-and-rewarding-scientific-impact/

Dear Director Jorgenson:

On behalf of the Physicians Committee for Responsible Medicine, a 501(c)(3) nonprofit organization supported by nearly one million members and supporters worldwide working for effective, efficient, and ethical medical research and product testing, thank you for the opportunity to provide input on how to incentivize and reward rigorous, reliable research that delivers value to the American public.

Biomedical research still focuses heavily on the biology of other species, with as many as 100 million dogs, cats, monkeys, mice, rats, and other animals being used in U.S. laboratories each year.[1] But animals do not make reliable models of human biology and diseases and do not reliably predict human clinical outcomes, contributing to failures and wasteful spending in drug development and regulatory testing, and putting clinical trial participants at risk by failing to predict unsafe or ineffective products.[2] Investment in animal experimentation does not deliver value to the American public.

Fortunately, human-based research methods, including new approach methodologies (NAMs), have rapidly advanced and continue to improve, providing more reliable models of human biology and disease and more predictive testing than animal experiments. Models such as organoids and organs-on-chips are derived from donated human cells and tissues and can replicate the structural, functional, mechanical, and

[1] Veening-Griffioen DH, Ferreira GS, Boon WPC, et al. Tradition, not science, is the basis of animal model selection in translational and applied research. *ALTEX*. 2021;38(1):49-62. https://doi.org/10.14573/altex.2003301; Carbone L. Estimating mouse and rat use in American laboratories by extrapolation from Animal Welfare Act-regulated species. *Sci Rep*. 2021; 11(1):493. https://doi.org/10.1038/s41598-020-79961-0

[2] Ewart L, Apostolou A. Briggs SA, et al. Performance assessment and economic analysis of a human Liver-Chip for predictive toxicology. *Commun Med*. 2022;2(1):1-16. https://doi.org/10.1038/s43856-022-00209-1

other biophysical properties of the human body at the tissue and multi-organ levels.[3] In the last decade, human-based methods have improved in their degrees of biological complexity and their breadth of application, making them powerful tools to better understand and treat disease.

Recognizing the failures of animal experimentation and promise of nonanimal, human-based approaches, in April 2025, the NIH launched an initiative to prioritize human-based research and reduce animal use.[4] We applaud the steps the NIH has already taken to implement this initiative, including: establishing the Office of Research Innovation, Validation, and Application (ORIVA),[5] investing $150 million in human-based research projects under the Complement-ARIE program,[6] and developing infrastructure like the Standardized Organoid Modeling Center.[7]

To build on agency-wide efforts in this regard, we encourage the NIH to further incentivize and reward the replacement of animal experimentation with innovative, ethical, and effective nonanimal, human-based approaches. The recommendations below elaborate on this central aim, helping to ensure that the agency supports the most impactful science and delivers optimal value to the American public.

Thank you for your attention to these recommendations. We welcome further dialogue to aid their implementation.

Sincerely,

Mikalah Singer, JD, LLM
Science Policy Specialist
Physicians Committee for Responsible Medicine

---

[3] Ingber DE. Human organs-on-chips for disease modeling, drug development and personalized medicine. *Nat Rev Genetics*. 2022; 23:467-491. https://doi.org/10.1038/s41576-022-00466-9

[4] NIH to prioritize human-based research technologies. National Institutes of Health. April 29, 2025. Accessed August 10, 2026. https://www.nih.gov/nih-prioritize-human-based-research-technologies

[5] National Institutes of Health Division of Program Coordination, Planning, and Strategic Initiatives. Proposed Reorganization: New Offices within DPCPSI. Published January 14, 2026. Accessed August 14, 2026. https://dpcpsi.nih.gov/proposed-reorg-establish-oriva-orepa

[6] National Institutes of Health. NIH invests $150 million in human-based research to reduce use of animal models. Published March 18, 2026. Accessed August 6, 2026. https://www.nih.gov/news-events/news-releases/nih-invests-150-million-human-based-research-reduce-use-animal-models

[7] National Institutes of Health. NIH Standardized Organoid Modeling (SOM) Center. Accessed August 6, 2026. https://www.nih.gov/som

1. RIGOR AND REPRODUCIBILITY

The Physicians Committee supports the NIH's increasing focus on rigorous and reproducible biomedical research—particularly its recognition that translatability is a key related factor. The August 2025 Leading in Gold Standard Science implementation plan highlighted bolstering human-based research technologies and expanding efforts of the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) as ways to advance more reproducible and replicable biomedical approaches.[8] Animal experiments have historically suffered from poor rigor and reproducibility. Although this impacts translation rates, no number of measures to improve the rigor and reproducibility of animal experimentation will make it optimally translatable due to inherent limitations in model validity.

Thus, shifting to nonanimal, human-based approaches is crucial for improving translatability. As this shift unfolds, we must ensure that the development and adoption of NAMs is not hindered by the unwarranted insistence to benchmark against animal experiments and that the problems of rigor and reproducibility that have plagued animal experimentation are not repeated. *To ensure that human-based methods are optimally reproducible, the NIH should support standardized validation efforts, infrastructure projects and collaboration to facilitate centralized and accessible resource sharing, update in vitro reporting guidelines, and encourage preregistration.*

Recommendation 1.1. Establish standardized validation efforts for all model types.

Because of the novelty of NAMs, further investment is needed to establish confidence in their ability to model human biology and clinical characteristics. Historically, benchmarking nonanimal methods has involved comparisons to the results of animal tests, but stakeholders increasingly recognize this as problematic. Formal validation practices for NAMs in regulatory toxicology are moving away from comparisons to non-human animal test results and instead are incorporating human data, resulting in the acceptance of human-specific approaches that outperform animal tests.[9] Patient-derived data should be used as benchmarks to corroborate findings from human disease models and to establish model validity. Other considerations that are crucial for

[8] National Institutes of Health. Leading in Gold Standard Science An NIH Implementation Plan. Published August 22, 2025. Accessed August 10, 2026. https://www.nih.gov/sites/default/files/2025-08/2025-gss.pdf

[9] van der Zalm AJ, Barroso J, Browne P, et al. A framework for establishing scientific confidence in new approach methodologies. Archives of Toxicology. 2022;96(11):2865-2879. https://doi.org/10.1007/s00204-022-03365-4; Kleinstreuer NC, Hoffmann S, Alépée N, et al. Non-animal methods to predict skin sensitization (II): an assessment of defined approaches. Critical Review Toxicology 2018;48(5):359-374. https://doi.org/10.1080/10408444.2018.1429386; Clippinger AJ, Raabe HA, Allen DG, et al. Human-relevant approaches to assess eye corrosion/irritation potential of agrochemical formulations. Cutan Ocul Toxicol. 2021;40(2):145-167. https://doi.org/10.1080/15569527.2021.1910291

validation efforts include fitness for purpose, human biological relevance, technical characterization, data integrity and transparency, and independent review.[10]

Validation of human disease models should continue to be refined and tailored according to each specific use, to help build confidence among researchers, industry, and regulators. Sequencing, imaging, and other data-driven approaches can accelerate model validation by enabling deep resolution and high-throughput comparisons between experimental platform readouts and gold standard clinical characteristics.

The need to validate NAMs is a frequent topic of interest as federal scientific agencies and other regulatory bodies around the globe increasingly turn to their use, but animal-based methods have never been held to the same standards—until now. The August 4th Guidance on Applications for Studies with Animal Models published by National Cancer Institute (NCI) Director, Anthony Letai, highlights new requirements for grant applicants to include plans to validate animal experiments with human-based models.[11] *Building on NCI's guidance, we recommend that the NIH prioritize the development of standardized validation efforts for animal- and nonanimal-based research models, ensuring that the framework emphasizes human biological relevance, reliability, and fitness for purpose, rather than concordance with historical animal data. Measurable indicators can include the number of new, validated, and human-based methods.*

Recommendation 1.2. Establish *in vitro* reporting guidelines.

Reproducibility across experiments and research labs is crucial for building trust in NAMs and other human-based methods and maximizing the translatability of findings. *To improve reproducibility, the NIH should establish standard* in vitro *reporting guidelines, require all grantees to follow these standards, and encourage publishers to require these standards.* Several examples of *in vitro* reporting guidelines exist that the NIH could build from or point toward.[12] Measurable indicators could include the number of updated reporting guidelines implemented by publishers and improved reproducibility of published studies.

---

[10] van der Zalm AJ, Barroso J, Browne P, et al. A framework for establishing scientific confidence in new approach methodologies. Archives of Toxicology. 2022;96(11):2865-2879. https://doi.org/10.1007/s00204-022-03365-4

[11] Letai A. Guidance on studies with Animal Models. National Cancer Institute. August 4, 2026. Accessed August 6, 2026. https://www.cancer.gov/grants-training/nci-bottom-line-blog/2026/guidance-animal-models

[12] National Centre for the Replacement Refinement & Reduction of Animals in Research. The DRIVER recommendations. Accessed August 18, 2026. https://nc3rs.org.uk/3rs-resources/driver-recommendations; Hartung T, De Vries R, Hoffmann S, et al. Toward Good In Vitro Reporting Standards. ALTEX. 2019;36(1):3-17. https://doi.org/10.14573/altex.1812191; OECD. *Guidance Document on Good In Vitro Method Practices (GIVIMP)*. Organisation for Economic Co-operation and Development; 2018. Accessed February 24, 2023. https://www.oecd-ilibrary.org/environment/guidance-document-on-good-in-vitro-method-practices-givimp_9789264304796-en;jsessionid=IcFbIMGZO8qR-_Hk2btN6Zt0IMML8oWdMBQ0gGcK.ip-10-240-5-51; Pamies D, Leist M, Coecke S, et al. Guidance document on Good Cell and Tissue Culture Practice 2.0 (GCCP 2.0). ALTEX. 2022;39(1):30-70. https://doi.org/10.14573/altex.2111011

Recommendation 1.3. Encourage prospective study registration.

To ensure the rigor and reproducibility of studies that use NAMs, the NIH should encourage investigators to prospectively register study designs and analysis plans. (Prospective registration includes preregistration in a registry database and the publishing of registered report article types.) This recommendation is in line with recommendations 2.1 and 2.2 within the final report of the ACD Working Group on Enhancing Rigor, Transparency, and Translatability in Animal Research, which aimed to raise awareness and understanding of prospective registration and to develop and implement a program to evaluate the impacts of prospective registration on research quality.[13] *We support these recommendations and urge the NIH to also implement them in the context of NAMs.* The benefits of prospective registration with regard to research quality, transparency, and reproducibility will help continue to build scientific confidence in human-based methods. Measurable indicators include the number of preregistered study designs.

2. DATA, SOFTWARE, AND MODEL SHARING

NAMs and other human-based methods are powerful tools to accelerate biomedical research discovery. A 2022 survey found that biomedical researchers are eager to use 3D human *in vitro* models, but cite complexity, standardization, entry costs, and lack of ready-to-use systems as barriers.[14] Addressing these issues will enable broader validation, reproducibility, and scaling. We applaud the steps the NIH has already taken toward these goals, but to maximize the value of continued investments, *we recommend placing an increased focus on democratizing human-based technology: supporting coordinated NAMs infrastructure to make human tissues, cells, and data more accessible and make platforms more operational.*

Recommendation 2.1. Expand human-based research specific infrastructure to facilitate resource sharing.

The Office of Research Infrastructure Programs (ORIP) 2026-2030 Strategic Plan builds on the 2025 NIH-wide initiative by setting out a strategic priority to expand infrastructure, and the NIH has already started to provide awards in this area to benefit

[13] ACD Working Group on Enhancing Rigor, Transparency, and Translatability in Animal Research. Final Report. Published June 11, 2021. Accessed August 12, 2026. https://www.acd.od.nih.gov/documents/presentations/06112021_ACD_WorkingGroup_FinalReport.pdf

[14] Busek M, Aizenshtadt A, Amirola-Martinez M, et al. Academic User View: Organ-on-a-Chip Technology. *Biosensors*. 2022;12(2):126. https://doi.org/10.3390/bios12020126

human-based research,[15] including C06 construction grants.[16] To further incentivize the uptake of human-based methods and maximize the value of existing investments, NIH should support funding opportunities within the existing mechanisms. Several P-series mechanisms can be used to support biotechnology resources, facilities, and program development that would incentivize resource sharing and facilitate the creation of high-quality data and tools, maximizing research investments and catalyzing collaborations. Examples include *PN2 (Research and Development Center), P30 (Center Core Grants), and P41 (Biotechnology Resource Grants) mechanisms* that can be used to incentivize new programs, facilities, institutions, groups, and labs dedicated to using human-based methods. Specifically, we recommend increased support for biobanks through such mechanisms, including a national resource for iPSC acquisition and differentiation, with diverse cell and tissue types, disease status, deep phenotypes, genotypes, and demographics. We also urge the NIH to reward investigators for collecting samples and data with community co-leadership, cultural sensitivity, reciprocity, and regular evaluation. Finally, we recommend the agency prioritize infrastructure and resource sharing for nonanimal-derived biomaterials and antibodies. *Measurable indicators can include the number of biobanks, breadth and depth of shared resources, and rates of utilization by the research community.*

Recommendation 2.2. Support artificial intelligence and computational tools to facilitate data sharing.

During drug development, millions of animals are used every year in the United States to test compounds that have already been shown to be safe. Research teams often choose to repeat animal experiments rather than manually search massive existing datasets. Tools such as artificial intelligence will accelerate biomedical research discovery and facilitate more efficient data sharing. *In collaboration with the Food and Drug Administration, NIH should create funding opportunities to promote the development of tools like Insilica's ToxIndex, which promote data reuse and reduce expensive and duplicative animal experiments.*[17]

Recommendation 2.3. Create a centralized database of human-based models.

Existing databases of NAMs are limited to regulatory applications or do not cover the entire spectrum of research areas. For example, the NIH Collection of Alternative

[15] National Institutes of Health Office of Research Infrastructure Programs. Strategic Plan Infrastructure for Innovation Fiscal Years 2026-2030. Accessed August 10, 2026. https://orip.nih.gov/sites/g/files/mnhszr376/files/ORIP-Strat-Plan-Final-11-25-2025-508-updated-compressed-Final-508.pdf

[16] Development of Collaborative Research Facilities or Research-Resource Facilities (C06 Clinical Trial Not Allowed). National Institutes of Health. Published March 6, 2026. Accessed August 10, 2026. https://files.simpler.grants.gov/opportunities/6179eb2c-9239-40ad-9a12-7329291f3509/attachments/f1f13568-9784-4a26-938b-5e0133e0c8e2/PAR-26-106-Revised-Full-Announcement.html

[17] ToxIndex. Insilica. Accessed August 10, 2026. https://toxindex.com/

Methods for Regulatory Application (CAMERA) provides information on 50 NAMs and is searchable by test method type, predicted endpoints, measure effect, and regulatory and test guidelines,[18] but there is a need for a broader centralized information database that aggregates human-based models across all areas of research and applications. This platform would improve uptake of NAMs; enable more accessible retrieval of information about models' capabilities and validation status; and help investigators, peer reviewers, and animal oversight committees better determine appropriate rationales for animal use versus replacement models. *NIH should develop an easily searchable, centralized database of human-based methods across all research areas, characterized by level of development, validation status, and context of use, and linked to publications. Measurable indicators can include creation of this resource, number of entries, and rate of utilization by the research community.*

3. TRAINING AND MENTORSHIP

As NIH continues to embrace human-based research and move away from unreliable animal-based methods, training and education in NAMs and other human-based methods is a key consideration,[19] and according to the 2023 final report of the Advisory Committee to the Director Working Group on Catalyzing the Development and Use of Novel Alternative Methods to Advance Biomedical Research, constitutes one of the highest priority needs for advancing NAMs.[20]

To promote the growth and stability of the biomedical research workforce, we recommend the prioritization of interdisciplinary education and training in biomedical research, bioengineering (including microfabrication and microfluidics), pharmacology, toxicology, computational biology, and AI through the support of curricula and training center development.[21] Because animal-based methodologies were the default in biomedical research for decades, undergraduate and graduate curricula are still largely based on these approaches, creating a workforce that lacks the necessary skills to advance the development, usability, qualification, and application of NAMs. Additionally, budgetary constraints and concerns limit the willingness of institutions and individual PIs to pursue the development of new curricula and implemention of training programs on human-based methods.

---

[18] National Institutes of Health. Collection of Alternative Methods for Regulatory Application. Accessed August 19, 2026. https://camera.od.nih.gov/explore

[19] Moruzzi A, Shroff T, Keller S, Loskill P, Cipriano M. Training the Next Generation of Researchers in the Organ-on-Chip Field. *Educ Sci*. 2023;13(2):144. doi:10.3390/educsci13020144

[20] Chang HY, Jorgenson L, Baines A, et al. *Catalyzing the Development and Use of Novel Alternative Methods*. National Institutes of Health; 2023:42. Accessed May 13, 2026. https://www.acd.od.nih.gov/documents/presentations/Working_Group_Report.pdf

[21] Moruzzi A, Shroff T, Keller S, Loskill P. Training the Next Generation of Researchers in the Organ-on-Chip Field. *Education Sciences*. 2023;13(2):144. https://doi.org/10.3390/educsci13020144

There are several existing mechanisms NIH has in place that will promote the growth and stability of the biomedical research workforce, as the NIH moves away from animal experiments toward a more human-focused landscape. Such mechanisms, in addition to the infrastructure-focused mechanisms in sections 1 and 2 of this response, support education projects (R25), career development (K-series), fellowship (F-series), and training (T-series) grants. Within these mechanisms, NIH can support the agency-wide human-based initiative by providing funding for institutional programs, conferences for training, and independent-investigator career development. *Incentives, including funding, will provide direct benefits to institutions and investigators willing to lead the charge, and address the current educational gap in human-based approaches to better support the scientific workforce.*

We recommend evaluating the number of career development and path-to-independence grants, the plateauing rates of PhDs going into industry versus academia, and the economic value to training PhDs, such as wages and economic productivity, as specific measurable indicators to assess and quantify the impact of training and educational support.

4. COLLABORATION

This administration has highlighted team-based research as critical for leading us into the new age of science.[22] To best support interdisciplinary and convergent collaboration, NIH should place additional focus on transparency of research, accessibility of technologies and data, and incentives that support team-based outcomes rather than individual accolades.

Recommendation 4.1. Continue to leverage public-private partnerships to advance interdisciplinary, cross-sector NAMs development.

Public-private partnerships demonstrate the value of collaboration and are already being used to support the development, validation, and deployment of advanced human-based research technologies, including NAMs. These technologies are often developed by small businesses and industry innovators but require coordinated federal engagement to reach scale and regulatory acceptance. The Validation and Qualification Network (VQN) of NIH's Complement-ARIE program demonstrates how well-

[22] Kratsios M. Science A New Golden Age. Office of Science and Technology Policy. July 2026. Accessed August 10, 2026. https://www.whitehouse.gov/science/

structured public-private partnerships can align federal agencies, regulators, industry, and nonprofit stakeholders around shared technology deployment goals. As human-based methods are increasingly implemented, the Complement-ARIE VQN model is especially valuable because it treats validation and regulatory preparedness for NAMs as a shared responsibility, coordinating across agencies and collaborators to develop common frameworks, standardized reporting, and regulatory-ready data packages. *The public-private partnership model should be used as an example of the importance of collaboration in science, and we recommend that model developers and engineers should collaborate directly with disease researchers to advance human-based science.*

Recommendation 4.2. Incentivize team-based science through high-priority challenge initiatives.

As highlighted in the July 2026 Report to the President, many of the current funding mechanisms favor small teams led by independent investigators, rather than larger-scale, moonshot projects that require collaboration across disciplines and sectors.[23] For example, the Complement-ARIE *NAMs Reduction to Practice Challenge* aimed to accelerate the development and validation of combinatorial NAMs that can transform the way biomedical research, drug development, and clinical trials are conducted. The NIH, in collaboration with the Environmental Protection Agency (EPA) and Food and Drug Administration (FDA), specifically sought multidisciplinary teams to compete for significant funding for their projects. Challenges like this support teams and projects in a way that would not be possible using smaller funding mechanisms.

We recommend prioritizing funding opportunities and challenges that promote collaboration across sectors, disciplines, and career levels. A measurable indicator may include the number of supported awards in areas where collaboration is particularly relevant such as new model development for specific disease areas where engineers and scientists working as a team are paramount.

5. ENTREPRENEURSHIP AND TRANSLATION

The July 2025 Report to the President notes that despite massive increases in biomedical research funding over the past three decades, the rate of significant breakthroughs and drug approvals has slowed down.[24] The following recommendations

[23] Kratsios M. Science A New Golden Age. Office of Science and Technology Policy. July 2026. Accessed August 10, 2026. https://www.whitehouse.gov/science/
[24] Kratsios M. Science A New Golden Age. Office of Science and Technology Policy. July 2026. Accessed August 10, 2026. https://www.whitehouse.gov/science/

will incentivize, assess, and integrate the impact of innovative and entrepreneurial efforts critical to translating research into clinical advances.

Recommendation 5.1. Bolster support for human-based methods to drive innovation and translation of discoveries into products for people.

The NIH can better advance the translation of discoveries into products for people through programs that promote the development and use of human-based methods. Research using animals often does not translate to humans because there are insurmountable species differences in anatomy, physiology, lifespan, disease characteristics, and more, leading to nine out of ten new treatments that appear successful in animals later failing in humans.[25] Additionally, many treatments that may have been safe and effective in humans risk being prematurely delayed or discarded due to misleading results in animals.[26]

We recommend that the NIH continue to prioritize human-based research in order to overcome species-specific translatability issues. The NIH should critically review how animal experiments have led to clinical benefits, and use drug attrition as a measurable indicator for success.

Recommendation 5.2. Promote stronger research assessment reform.

The July 2026 Report to the President notes that the current scientific landscape, particularly within academia, rewards publications over commercialization, which impacts when products and treatments reach patients. Prioritizing publications over scientific impact may affect multiple processes within the scientific enterprise, including grant evaluations, hiring and tenure assessments, and public trust. *We recommend that NIH support stronger research assessment reform, assessing scientists for potential and actual impact, rather than bibliometrics, including during the grantmaking process.*[27] The NIH should use resources created by the Declaration on Research Assessment (DORA), which advances practical and robust approaches to research assessment reform.[28]

---

[25] Thomas D, Chancellor D, Micklus A, et al. Clinical Development Success Rates and Contributing Factors 2011-2020. Published online 2021. https://go.bio.org/rs/490-EHZ-999/images/ClinicalDevelopmentSuccessRates2011_2020.pdf; Pound P, Ritskes-Hoitinga M. Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. *J Transl Med*. 2018’16(1):304. https://doi.org/10.1186/s12967-018-1678-1

[26] Van Norman GA. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is it Time to Rethink Our Current Approach?. *JACC: Basic to Translational Science*. 2019;4(7):845-854. https://doi.org/10.1016/j.jacbts.2019.10.008

[27] Bailey J. Biomedical Research Must Change—But a Shift Toward Human-specific Research Methods Is Only Part of What Is Needed. Alternatives to Laboratory Animals. 2021;49(3): 69-72. https://doi.org/10.1177/02611929211030417

[28] San Francisco Declaration on Research Assessment. DORA. Accessed August 10, 2026. https://sfdora.org/read/

Recommendation 5.3. Increase support for collaborative infrastructure projects to promote human-based model development.

The NIH has already funded the establishment of several regional human-based methods centers through grants and cooperative agreements (U2C and U24), such as the Translational Centers for Microphysiological Systems (TraCe MPS)[29] and Tissue Chip Testing Centers.30 Such centers accelerate the translational use of human-based methods from drug development through regulatory acceptance and adoption for industrial use, as well as provide resources for the validation of methods to improve the evaluation of drug efficacy and toxicity. However, support for the TraCe MPS program is ending, creating a challenge for investigators to secure funding from other sources—including from funding opportunities that involve competition with animal-based infrastructure programs—and threatening to stall important progress toward the broader adoption of these technologies. *We recommend that the NIH expand funding for human-based infrastructure projects and ensure mechanisms are in place for awardees to be eligible for sustained financial support. Measurable indictors can include additional human-focused infrastructure projects supported, the number of models developed at such programs, and the broader uptake of these models by the research community.*

Recommendation 5.4. Incentivize collaboration between industry and academic institutions to democratize human-based models.

Small- and medium-sized businesses (SMBs) are central drivers of innovation in human-based research technologies,[31] and during the past two decades, the human-based methods industry has increased substantially, with growth projections continuing.[32] The Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs support small business entrepreneurs. To bolster the development of human-based technologies and improve translation from bench to bedside, these programs should prioritize projects that replace animal use and generate data suitable for regulatory confidence, with bridging mechanisms between Phase II and later-stage support. Additionally, the NIH should encourage the model developers it supports to market to academic labs just as much as the pharmaceutical industry, to make human-based approaches more accessible at all levels of the scientific enterprise.

---

[29] National Institutes of Health. Translational Centers for Microphysiological Systems (TraCe MPS)(U2C Clinical Trials Not Allowed). National Institutes of Health. Dec 27, 2022. Accessed August 10, 2025. https://grants.nih.gov/grants/guide/rfa-files/RFA-TR-23-001.html

[30] National Institutes of Health. Tissue Chip Testing Centers: Validating Microphysiological Systems (U24). National Institutes of Health. March 17, 2026. Accessed August 10, 2025. https://grants.nih.gov/grants/guide/rfa-files/RFA-TR-16-006.html

[31] Small Business Administration. *Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) Programs: Policy Directive.* Published May 2023. Accessed August 10, 2025. https://www.sbir.gov/sites/default/files/2024-07/SBA%20SBIR_STTR_POLICY_DIRECTIVE_May2023.pdf

[32] Organoids and Spheroids Market (2025-2030). Grand View Research. April 2025. Accessed August 19, 2026. https://www.grandviewresearch.com/industry-analysis/organoids-spheroids-market; Sharma R. Microphysiological System Market. Data Intelo. April 2026. Accessed August 19, 2026. https://dataintelo.com/report/microphysiological-system-market

To incentivize entrepreneurial efforts in human-based research, NIH should prioritize funding opportunities that directly support collaboration between industry and academic institutions or aim to transition promising research and development to market such as RC3 and R43 grants, and U-series cooperative agreements. Measurable indicators of success can include academic utilization of commercial human-based models.

6. FOUNDATIONAL SCIENTIFIC EXPLORATION

The NIH's mission focuses on enhancing health, lengthening life, and reducing illness, including basic research to investigate the causes and mechanisms of human diseases. Human-based methods are paramount to advancing translational research and have already been shown to overcome existing, unavoidable, species-specific differences between non-human animals and humans, but human-based methods also play a vital role in basic research too.

Recommendation 6.1. Prioritize human-centered methods for basic research.

Human-based methods use human cells, tissues, and data to improve our understanding of human biology and disease mechanisms, building the foundation for translational success. *NIH should prioritize increased uptake of human-based methods by embedding these approaches within the academic enterprise and ensuring that all areas of NIH-supported research, including basic research, are human relevant.*

Researchers can be reluctant to onboard new approaches due to the risk involved, whether related to high up-front costs, lack of training, funding, or institutional support, or the inherent process of failing, troubleshooting, and repeatedly trying again. This is all compounded by the fact that most grants require demonstrating preliminary data, which is difficult to generate if an investigator does not yet have funds to implement new approaches. These inherent risks contribute to the slow uptake of otherwise promising human-based methods.

To incentivize researchers to embrace human-based models, the NIH should support investigators to adopt these models and to afford to take risks in their research. Several mechanisms have been used to incentivize high-risk, high-reward research (DP1, DP2), but should be implemented to specifically support projects that use and develop human-based methods. Exploratory grants or smaller pilot program support, including R21 grants (which do not require preexisting data and are smaller than standard R01 grants), may attract scientists with smaller labs or at less prestigious universities which would refocus support for the individual scientist and revitalize America's scientific

enterprise. Additionally, larger-scale pilot programs to support institutions operationalizing human-based methods into departments, research groups, and facilities would take the pressure off individual researchers to adopt new technologies without existing institutional support. To ensure these mechanisms support human-based science, funding opportunity language should specifically state that projects proposing animal use are not eligible.

Research involving human subjects, communities, and populations is also critical for advancing foundational knowledge about human health. *We recommend that the agency shift funds from animal experimentation toward human-centered basic research aimed at understanding disease risk, prevention, and non-biological health influences. Measurable indicators can include relative investment in animal- and human-based basic research projects.*

Recommendation 6.2. Reward researchers using nonanimal methods to answer novel scientific questions.

The NCI Guidance on Applications for Studies with Animal Models notes that scientistic studies should be "inspired by asking great questions with high potential or human impact," and that the "best and most relevant scientific methods" should be used to accomplish that goal.[33] *The NIH should encourage researchers to consider novel scientific questions that have not been asked before or have only been attempted to be answered using animals, and to do so with approaches that reflect human biology rather than nonhuman animals.*

7. PUBLIC IMPACT

NAMs and other human-based research methods provide more reliable tissue and disease modeling than animals, leading to more predictive clinical testing and improving the efficiency of federal research expenditure. Moving away from animal experimentation and supporting a scientific landscape built on human-based methods means that more treatments—that are safer and more effective—will make it to patients and at a faster pace. Ultimately, this shift will improve health outcomes, clinical practice, economic growth, public trust, and societal benefit.

Recommendation 7.1. Facilitate a systematic review of clinical translation of animal experiments.

[33] Letai A. Guidance on Applications for Studies with Animal Models. National Cancer Institute. Published August 4, 2026. Accessed August 10, 2026. https://www.cancer.gov/grants-training/nci-bottom-line-blog/2026/guidance-animal-models

Continued support for projects involving animal experiments impacts clinical translation and harms the American public. To achieve the federal goal of improving clinical translation and reducing the use of animals, the NIH should prioritize synthesizing: the failures of animal experiments to model human biology and translate to clinical advances; and the capabilities and needs of nonanimal approaches. *We recommend that the NIH support systematic reviews on these topics, especially in high-priority areas like infectious diseases and neuroscience, where animals are still perceived by some as scientifically necessary, or where key modeling capabilities need further development.*

Recommendation 7.2. Track and report number of animals used and funding for projects involving animals and human-based methods.

There is growing support from the public to move away from animal use. A 2024 national survey showed that more than 85% of Americans agreed that animal-based research should be phased out in support of methods that do not use animals,[34] and a 2026 Gallup poll found that a majority of Americans view medical testing on animals as morally wrong.[35] In addition to the recommendations made throughout this response, to address this growing public interest, improve public trust in science, and foster transparency and accountability, *we recommend that the NIH accurately track and publicly report of the number of animals used by funded facilities and funding toward projects involving animals and NAMs.*

The Office of Laboratory Animal Welfare currently requests only crude numbers for animal use every four years from U.S. facilities. Yet more than 95% of the animals reported to OLAW are excluded from the federal Animal Welfare Act, which is overseen by the USDA, meaning those animals are never accurately counted by any federal agency. The Federal Animal Research Accountability Act (H.R. 3295) would require the NIH to annually collect from each funded facility the number of animals bred, housed, and used in the previous year—sorted by species and level of pain and distress—and make those reports publicly available. *We recommend that the NIH either proceed with the bill's approach or implement a system requesting only the total number of animals by species (not including pain and distress categories). In either case, the NIH should make individual facility numbers available to the public, as the United States Department of Agriculture does.*

[34] Morning Consult. Physicians Committee for Responsible Medicine. U.S. Gen Pop Beliefs on Animal Testing. 2024. Accessed August 11, 2026. https://pcrm.widen.net/s/qzfxtfh7bw/animal-testing-survey
[35] Brenan M. Moral Acceptability Falls for Several Behaviors. GallUp. June 9, 2026. Accessed August 11, 2026. https://news.gallup.com/poll/710981/moral-acceptability-falls-several-behaviors.aspx

A 2023 NIH presentation showed that competitive awards using NAMs grew steadily each year since 1981, reaching ~8% in 2021.[36] A 2024 GAO report estimated that, in recent years, animals were used in ~41% of intramural projects and ~25% of extramural projects.[37] These data are crucial for justifying large-scale human-based initiatives, tracking progress toward goals, and rebuilding public trust and accountability. *Aggregate funding rates toward projects involving human-based methods and animals should be reported at least annually. Individual grant data regarding NAM and animal use should also be reported in RePORTER—the latter, for example, by incorporating information from grant application form G.220.*

Recommendation 7.3. Require animal use transparency in grant application abstracts.

The National Cancer Institute has already announced new guidance to promote additional transparency related to animal use.[38] Starting October 5, 2026, grant applicants will need to include a justification for animal use and explanation for why nonanimal methods would not be sufficient to address the question at hand *in the abstract of the grant application*, which will make make this information publicly accessible via RePORTER. *NIH should implement this requirement agency-wide to improve transparency and public accountability toward animal replacement goals.*

CONCLUSION

As elaborated throughout the entirety of this response, NIH can support human-based research methods through improving resource and data sharing, increasing training and education opportunities, facilitating collaboration, and incentivizing model use and development. In addition to the aforementioned recommendations, w*e recommend that the NIH take into account all recommendations, implementation approaches, and measurable indicators, including (A) successful clinical trial outcomes, (B) translatability to humans, (C) funding for training in human-based methods, (D) improved accessibility of data and methods, (E) public opinion, and (F) investment in animal experiments compared to NAMs made in this response when determining how to incentivize and reward rigorous research that delivers to the American public.*

---

[36] Chang H and Jorgenson L. Catalyzing the Development of Use of Novel Alternative Methods (NAMs) to Advance Biomedical Research. NIH Advisory Committee to the Director Working Group. Published December 14, 2023. Accessed August 12, 2026. https://acd.od.nih.gov/documents/presentations/12142023_Update_Chang_Jorgenson.pdf

[37] United States Government Accountability Office. National Institutes of Health Assessing Efforts to Improve Animal Research Could Lead to Greater Human Health Benefits. Published December 2024. Accessed August 12, 2026. https://www.gao.gov/assets/gao-25-107140.pdf

[38] Letai A. Guidance on Applications for Studies with Animal Models. National Cancer Institute. Published August 4, 2026. Accessed August 10, 2026. https://www.cancer.gov/grants-training/nci-bottom-line-blog/2026/guidance-animal-models

278. Woods Hole Center for Oceans and Human Health

I am responding: On behalf of an organization

Name: Neelakanteswar Aluru

Name of Organization: Woods Hole Center for Oceans and Human Health

Type of Organization: Non-Profit Research Organization

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Response-to-NOT-081926.docx

Response to NOT-OD-26-087: Measuring and Rewarding Scientific Impact

Submitted by: Woods Hole Center for Oceans and Human Health (COHH), Woods Hole Oceanographic Institution

COHH sits at the intersection of oceanography, toxicology and public health — a model of exactly the convergent, team-based science this RFI is asking NIH to better measure. Our comments draw on that vantage point: NIH-funded centers where marine science directly protects human health (harmful algal bloom toxins, seafood safety, weather-linked ocean hazards, coastal community exposure).

Rigor and Reproducibility

- Environmental and oceanographic exposure data are inherently variable (seasonal, geographic, climate-driven); metrics should credit field/monitoring protocols and standardized reference methods as rigor indicators. Current rigor metrics lean heavily on replication, which is not always feasible with environmental data. Instead, metrics should also credit the practices that make rigor achievable *without* literal replication.
- Reward development and adoption of community reference materials and interlaboratory calibration exercises (e.g., algal toxin standards) that make replication across labs feasible.
- Distinguish reproducibility expectations for controlled lab studies vs. field/environmental causal research, where true replication is often impossible and triangulation across methods should count instead. i.e., testing the same causal question with multiple independent methods where convergent findings across these approaches substitute for direct replication.

Data, Software, and Model Sharing

- Invest in long-term environmental monitoring datasets (e.g., HAB toxin time series, coastal water quality) as high-value shared infrastructure, even when they don't generate a single high-impact paper.
- Credit contributions to shared data standards and interoperable databases that link oceanographic, environmental, and health/exposure data (a persistent bottleneck in oceans-and-human-health work).
- Value tool/model reuse metrics (downloads, citations of software/datasets) alongside traditional publication metrics.

Training and Mentorship

- COHH-style centers train scientists who must be fluent across disciplines (marine science/oceanography + epidemiology/toxicology); metrics should recognize cross-disciplinary training investment, not just trainee publication counts.
- Track placement and career trajectories of trainees into public health, regulatory, and translational roles, not only academic tenure-track outcomes.
- Recognize mentorship of early-career investigators bridging environmental science and health, a workforce pipeline NIH doesn't currently track well.

Collaboration

- Team science in Oceans and Human Health requires oceanographers, chemists, toxicologists, clinicians, and community/public-health partners; authorship-order-based credit systems undervalue this. Support adoption of contributor role taxonomies (e.g., CRediT) in NIH reporting/review.
- Credit infrastructure-building roles (shared field stations, monitoring networks, sample repositories) that enable others' science but rarely appear as first/last-author output.
- Recognize interdisciplinary collaboration and capacity building as research outcomes; sustained OHH partnerships across disciplines and communities develop expertise, methods, and response capacity that enable more effective science and public-health action.
- Recognize center-level and network-level impact (e.g., NIEHS/NSF Oceans and Human Health Centers) as a unit of assessment, not only individual PI metrics.
- NIH should continue to support researchers to collaborate broadly (including foreign collaborations) and to attend scientific meetings where ideas are exchanged, which improves the research.

Entrepreneurship and Translation

- Unlike the traditional "bench to bedside" model of translation, in the field of oceans and human health research translation often takes the form of providing information and methods that county, state or federal agencies can use to carry out their mission to protect human health. For example, OHH research has improved tools and information for monitoring, forecasting, and responding to harmful algal blooms (HABs), including methods for species and toxin detection that inform shellfish closures and other public health actions. Related to this, the collaborative, interdisciplinary networks established through OHH research programs have built regional capacity and strengthened decision-making and coordination for HAB response. Metrics should capture the transfer and adoption of policies and practices, not just commercialization.
- Credit collaboration with NOAA, EPA, FDA, and state health departments as a translation pathway distinct from industry partnership.

Foundational Scientific Exploration

- High-risk work on emerging marine hazards (novel toxins, range expansion of pathogens/HABs) often lack preliminary data to demonstrate feasibility. Under current review criteria that weigh preliminary data heavily, this puts exploratory research on emerging threats at a disadvantage. Metrics and review criteria should support and invest in high-risk research.

Public Impact

- To better assess and recognize the public impact of research such as that conducted by interdisciplinary programs like the OHH program, NIH should collect information about actions and impacts that have been enabled by the research, such as coastal community health advisories, seafood safety guidance, reduced illness from HAB exposure, and public trust built through community-engaged monitoring.
- Consider community and stakeholder engagement metrics (co-developed research questions, public data dashboards and portals, advisory contributions, co-created outreach materials) as evidence of public impact.

Cross-cutting considerations for NIH to weigh

- Metrics should be field-appropriate: what counts as rigor/impact in bench biomedical science doesn't map cleanly onto environmental/field-based research that benefits human health.
- Interdisciplinary contributions should be explicitly valued: Addressing complex environmental health challenges requires sustained collaboration across disciplines with different methods and outputs; NIH evaluation should recognize the added value and effort required to integrate diverse forms of expertise.

279. Babak Heydari

I am responding: On behalf of myself

Name: Babak Heydari

Name of Organization: Northeastern University

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: AI-enabled scientific production creates a production–verification asymmetry: literature searches, candidate hypotheses, code, analyses, and manuscripts can be generated increasingly quickly, while independent validation, provenance review, causal interpretation, and responsibility for consequential judgments remain scarce. NIH should therefore treat verification capacity—not simply output production—as a central constraint on scientific impact.

NIH could identify areas especially suitable for replication based on their systemic importance and evidentiary risk. Relevant considerations include the extent to which other studies or clinical decisions depend on a finding; whether an influential result has been independently validated; the sensitivity of the finding to analytical choices; conflicts among existing results; reliance on difficult-to-audit data or AI-generated workflows; and the potential consequences of an erroneous conclusion. Citation counts alone should not determine replication priority because visibility may reflect cumulative advantage or collective attention rather than evidentiary importance.

Indicators of rigor should include independent replication and validation; robustness across reasonable specifications; transparent provenance of data, software, models, and analytical decisions; correction of errors; publication of credible negative findings; and evidence that consequential AI-generated outputs were understood and independently evaluated by accountable human researchers. Replication attempts should receive credit for the quality of the test and the information they generate, regardless of whether the original result is confirmed.

NIH should support replication, correction, and validation as first-class scientific activities. Possible incentives include dedicated replication supplements, renewal credit for projects that enable or complete independent validation, support for maintaining reproducible research artifacts, and recognition of researchers who identify and responsibly correct consequential errors. Review criteria should distinguish scrutiny and stewardship from the production of additional claims.

These indicators should not become another count-based checklist. The number of replications, robustness checks, or deposited files is not by itself evidence of rigor. Applicants should instead provide a concise impact-pathway statement identifying the relevant contribution, the validation performed, the evidence supporting each claimed link, and the remaining uncertainty. Links could be labeled “observed,” “projected,” or “not yet observable.”

NIH could pilot this approach in selected programs, train reviewers with anchored examples, and monitor reviewer agreement, reporting burden, strategic behavior, and whether the criteria actually increase independent validation and correction. Indicators that encourage superficial compliance, excessive burden, or avoidance of uncertain research should be revised or retired.

This concern is increasingly important under AI because methodological convenience can be mistaken for epistemic gain. See, for example, Messeri and Crockett, "Artificial Intelligence and Illusions of Understanding in Scientific Research," Nature 627:49–58 (2024), https://doi.org/10.1038/s41586-024-07146-0.

2. Data, Software, and Model Sharing: Shared data, software, and models are scientific public goods. Their value often materializes outside the originating team through independent reuse, validation, avoided duplication, methodological diffusion, and the creation of new research capabilities. Because the originating researchers capture only part of this value, conventional individual-level metrics predictably produce underinvestment in documentation, maintenance, interoperability, and stewardship.

NIH should therefore assess sharing through evidence of an impact pathway rather than through the mere existence or popularity of a deposited artifact. Relevant indicators include the proportion of users who are independent of the originating team; cross-institutional and cross-disciplinary reuse; use in replication or validation; downstream studies or tools enabled; sustained availability; version histories and documented corrections; interoperability with other resources; and evidence that the artifact reduced duplication or enabled work that would otherwise have been infeasible. Downloads, repository stars, mentions, or raw user counts should not stand alone.

AI makes this distinction more important. AI can rapidly generate code, models, datasets, and documentation of uneven quality. The volume of shared artifacts may therefore rise while provenance, validation, security, maintenance, and accountability become more difficult to assess. NIH should request disclosure of consequential AI use, the human validation performed, known limitations, the accountable contributors, and the conditions under which a shared model or dataset should and should not be used.

Incentives should address the underlying spillover problem. NIH could provide supplemental maintenance funding or renewal credit when consequential independent reuse is verified; recognize curation, software engineering, model validation, and stewardship as substantive contributions; support independent audits of widely used resources; and provide continued support for infrastructure whose value becomes evident only after the originating award ends.

My empirical research on multi-sided digital platforms illustrates the broader boundary problem underlying this recommendation. Our longitudinal study of Airbnb in Boston showed how activities creating value for direct platform participants can also reshape neighborhood social organization and affect people outside the focal transaction. The substantive setting differs from scientific sharing, but the general measurement lesson is relevant: evaluating only direct participants misses important spillovers and affected stakeholders. See Ke, O'Brien, and Heydari, "Airbnb and Neighborhood Crime," PLOS ONE

16:e0253315 (2021), https://doi.org/10.1371/journal.pone.0253315.

NIH should pilot these indicators proportionately. Expectations should vary by field, resource type, privacy constraints, career stage, and institutional capacity. The objective should be to recognize verified enabling value—not to impose uniform adoption thresholds or disadvantage necessary resources serving small research communities.

3. Training and Mentorship: Training and mentorship create scientific capabilities whose value is distributed across people, teams, institutions, and time. These contributions may influence the research ecosystem long after the originating grant ends, yet they are poorly represented by publication counts, grant dollars, or the subsequent prestige of a small number of trainees.

NIH should evaluate training through contextualized evidence of capability formation. Relevant indicators could include the development of methodological, computational, clinical, collaborative, and translational capabilities; trainee progression toward greater intellectual independence; transfer of knowledge across disciplinary or institutional boundaries; adoption of rigorous and reproducible practices; preparation for responsible use and validation of AI; sustained mentoring relationships; and the creation of inclusive research environments in which trainees can contribute consequential judgment rather than only execute assigned tasks.

AI makes this broader conception of mentorship more necessary. AI tools can expand access to sophisticated methods and reduce some barriers to participation, but they can also conceal gaps in understanding or substitute automated production for the development of scientific judgment. Effective mentorship should therefore include the ability to interrogate generated methods, recognize uncertainty, document provenance, validate outputs independently, and understand when human expertise remains indispensable.

Assessment should distinguish among mentoring roles. Primary advisors, co-advisors, research staff, collaborators, and peers may contribute differently to technical instruction, professional development, interdisciplinary integration, sponsorship, and research culture. A concise structured description, informed by contributor-role approaches such as CRediT, would be more informative than assigning all trainee outcomes to a single principal investigator.

NIH should recognize mentoring and capability-building through role-specific review credit, support for protected mentoring time, continuation funding for demonstrably effective training infrastructures, and team-level recognition where multiple contributors created the capability. Evidence should be appropriate to career stage and time horizon. Early indicators may concern emerging skills, independence, and reproducible practices; later indicators may include sustained scientific, clinical, public-sector, or entrepreneurial contributions.

Raw numbers of trainees, completion rates, publications, or placements should not stand alone. These outcomes depend on field conditions, institutional resources, labor markets, trainee interests, and prior advantages. NIH should also avoid creating incentives to select only trainees who appear most likely to produce easily measured outcomes.

A pilot could ask applicants for a short training-impact pathway: mentoring contribution -> capability developed -> evidence of increasing independence or responsible practice -> subsequent scientific or public value. Reviewers would assess the credibility of relevant links and acknowledge that downstream outcomes may not yet be observable.

4. Collaboration: Modern collaboration creates an attribution and information problem. A visible output may combine conceptual work, data collection, software development, model construction, prompting, AI-generated analysis, human correction, project coordination, and consequential scientific judgment. Publication position and principal-investigator status reveal increasingly little about this division of labor.

NIH should require concise, structured descriptions of accountable contributions. CRediT's contributor-role taxonomy provides a useful foundation, but it should be extended to include consequential AI use, human validation of generated work, responsibility for key decisions, and changes in roles over the life of a project. The purpose should not be to reward every listed role equally. It should be to make contribution, provenance, and accountability more observable.

Collaboration should also be assessed beyond the focal team. Relevant evidence includes the integration of otherwise disconnected disciplines or institutions; the creation of reusable interfaces, methods, or shared language; diffusion of knowledge across organizational boundaries; resolution of coordination problems; and the development of capabilities that allow others to collaborate effectively. Such bridge-building may generate substantial network value even when it does not produce a large number of papers for the person performing it.

My research on organizational innovation networks provides empirical support for evaluating collaboration contextually. In "Core or Periphery: Examining Where to Allocate Heterogeneous Inventors and the Impact on Firms' Innovation," my coauthors and I show that the value of locating heterogeneous inventors in core or peripheral network positions depends on problem complexity and actor characteristics. The implication for research assessment is that there is no universally optimal collaborative profile: the contribution of a role depends on the architecture and task of the team. Strategic Management Journal 47:331–363 (2026), https://doi.org/10.1002/smj.70005.

NIH should match recognition to the level at which collaborative value is created. Individual reviews can recognize accountable roles and judgment; team or project reviews can recognize integration, validation, and shared infrastructure; and program-level assessment can examine institutional concentration, bridging, and the diversity of methods and perspectives across a portfolio.

Structured role descriptions can themselves be gamed. NIH should therefore avoid converting contribution labels into a point system. Reviewers should consider the relationship between claimed roles, research artifacts, decisions, and outcomes. NIH should pilot the approach, monitor contribution-label inflation and administrative burden, and revise the taxonomy as human–AI research arrangements evolve.

5. Entrepreneurship and Translation: Entrepreneurial and translational impact should be assessed as a staged pathway, not inferred from patent counts, startup formation, licensing revenue, or other highly visible endpoints alone. These outputs may be useful evidence, but they do not necessarily establish

technical validity, clinical value, adoption, accessibility, or contribution to the NIH mission.

A translational impact pathway might proceed from a reliable result or capability to independent validation, technical maturation, regulatory or clinical milestones, adoption, and ultimately health or public value. Not every project will traverse the entire pathway, and downstream outcomes may require many organizations and years. Applicants should identify which links are observed, projected, or not yet observable and provide evidence appropriate to the stage of development.

Potential indicators include independent technical validation; demonstrated reproducibility outside the originating team; clarification of a consequential clinical or user need; achievement of regulatory, manufacturing, interoperability, or implementation milestones; adoption by clinical, public-sector, or industry partners; evidence of accessibility and usability; and learning that credibly reduces technical, regulatory, or market uncertainty. A well-validated failure may create substantial value by preventing costly downstream investment or redirecting a field.

NIH should match incentives to the relevant bottleneck. Early projects may need support for validation and de-risking; intermediate projects may need assistance crossing organizational or regulatory boundaries; and mature projects may warrant recognition for verified adoption, accessibility, or public benefit. Researchers who create enabling methods, data, standards, or validation capabilities should receive credit even when another organization completes commercialization.

AI increases the need for staged validation. AI can make prototypes, analyses, claims, and business narratives cheaper to produce, but it does not automatically resolve reliability, provenance, accountability, regulatory readiness, or real-world adoption. Visible production can therefore overstate translational maturity.

NIH should also distinguish private value from public value. Commercial success can coexist with unequal access, displacement of costs, or effects on people outside the focal transaction. Translation assessment should consider who benefits, who bears risks or costs, and whether the innovation advances health outcomes and accessibility.

These criteria should remain contextual and should not be combined into a universal translational score. Appropriate pathways differ across therapeutics, devices, diagnostics, software, policy, and service innovations. NIH could pilot pathway-based assessment in selected translational programs and compare it with current milestone and output measures.

6. Foundational Scientific Exploration: Foundational exploration produces uncertain, delayed, and often widely distributed benefits. A project may create a new method, problem formulation, dataset, research community, or well-validated negative result long before its downstream impact becomes observable. Assessment based primarily on publications, citations, or near-term translation can therefore shift effort toward safer and more immediately legible work.

AI may intensify this bias. By lowering the cost of searching, generating, and refining work in well-represented areas, AI can increase individual productivity while directing researchers toward similar data, methods, benchmarks, and questions. Hao et al. report an association between scientists' use of AI

and increased publication and citation impact alongside contraction in the collective range of topics studied. Although observational, this finding illustrates why individual productivity and portfolio-level exploration should be assessed separately. See Hao et al., Nature 649:1237–1243 (2026), https://doi.org/10.1038/s41586-025-09922-y.

NIH should recognize the option value created by foundational work. Relevant evidence may include a new problem formulation; a method or capability that opens previously inaccessible questions; a credible result that rules out an important path; bridges between previously disconnected research communities; independent follow-on exploration; or preservation of scientific options in areas that would otherwise be neglected.

Because these outcomes have different time horizons, applicants should distinguish observed contributions from projected pathways and outcomes that are not yet observable. Reviewers should evaluate the quality of the question, the credibility of the search or test, the capability created, and the information gained—not require a speculative claim of eventual clinical impact.

Incentives should include protected portfolio space for exploratory and high-risk research, longer evidence windows, support for replication of surprising foundational findings, and recognition of well-documented failure. Milestone reviews should focus on learning and capability formation rather than require that uncertainty disappear according to a predetermined schedule.

Portfolio diversity should be managed at the program level, not converted into a score assigned to individual investigators. NIH could monitor concentration in topics, methods, institutions, data sources, and scientific approaches while preserving expert judgment about field-specific opportunities. Large and small teams, central and peripheral network positions, and established and emerging fields can play complementary roles.

The objective is not to reward novelty as an abstract property. It is to prevent assessment criteria—especially in an AI-enabled environment—from systematically eliminating uncertain but potentially transformative scientific options.

7. Public Impact: Public impact emerges through a pathway connecting scientific contribution, reliable knowledge or capability, independent validation, adoption, and public value. Because this pathway is distributed across organizations and time, neither citations nor a single downstream outcome can adequately attribute impact to the originating research.

NIH should permit multiple forms of evidence while requiring respondents to identify the relevant causal links and remaining uncertainty. Potential indicators include incorporation into clinical guidelines or practice; verified use by health systems, public agencies, or community organizations; improvements in accessibility or affordability; contributions to health outcomes; independently documented economic or workforce effects; adoption of reusable tools or standards; public access to understandable research information; and evidence supporting public trust. Applicants should distinguish direct evidence from plausible projections.

Public-impact assessment should also consider distribution and externalities. An intervention may

generate aggregate benefits while leaving some communities unable to access it or imposing risks and costs on people outside the focal project. NIH should ask who benefits, who remains excluded, which stakeholders are affected, and whether claimed impact depends on complementary infrastructures or institutions.

Collective perception is itself relevant to public impact. Citations, rankings, media narratives, and apparent scientific consensus influence what researchers, policymakers, clinicians, and communities believe is promising or trustworthy. AI-generated summaries and evaluation narratives can accelerate such feedback before underlying claims are adequately verified.

My research with Ilami, Comola, and Prina on networked risk perception shows in another setting how beliefs and behavior can become community-level, mutually reinforcing phenomena rather than independent individual responses. The same general mechanism may affect scientific and public attention, although it requires direct empirical testing in this context. See Ilami et al., "Networked Risk Perception and Behavioral Bubbles," arXiv:2606.22599 (2026), https://doi.org/10.48550/arXiv.2606.22599.

NIH should recognize activities that support credible public impact, including transparent communication, correction of errors, community engagement, accessible resources, implementation partnerships, and responsible stewardship after an award ends. These activities should not be judged through communication volume or media attention alone.

Time horizons and expectations should vary by research type and career stage. Foundational researchers should not be required to demonstrate immediate public adoption, while translational claims should be supported by stage-appropriate validation. Public-impact indicators should inform contextual expert judgment rather than be aggregated into a universal score.

8. Other Areas: CROSS-CUTTING FRAMEWORK: SCIENTIFIC-IMPACT ASSESSMENT AS AN ADAPTIVE INTERVENTION IN AN AI-TRANSFORMED RESEARCH ECOSYSTEM

Central recommendation

NIH should treat scientific-impact assessment as an adaptive intervention in a multilevel research ecosystem, not as a search for a more comprehensive scalar metric. AI is rapidly changing the production economics and information structure of science, exposing weaknesses in output-based measures. NIH should assess contextualized pathways of contribution, validation, diffusion, and public value; match indicators and incentives to the level at which value is created; and revise criteria as the scientific system responds.

Why AI makes this necessary

AI lowers the marginal cost of searching literature, generating hypotheses, writing code and text, exploring parameter spaces, and producing candidate analyses. These capabilities can expand productivity and access to sophisticated methods, but costs fall unevenly. Verification, provenance, causal interpretation, high-quality data, judgment, and responsibility for consequential decisions remain

scarce. The result is a production–verification asymmetry: more claims and artifacts can be generated than the scientific system can independently evaluate.

AI also increases attribution and information asymmetry. Evaluators may not observe how AI was used, whether researchers understood a generated method, how much human validation occurred, or which contributors made the consequential decisions. Meanwhile, already measurable outputs become cheaper to produce, while stewardship, contradiction resolution, mentoring, integration, and validation remain difficult to count. Rewarding what becomes abundant because it is measurable can move effort away from what has become scarce.

Individual gains may also diverge from collective outcomes. AI can increase the productivity of individual researchers while concentrating attention on similar questions, methods, data, and benchmarks. Agentic systems may accelerate these dynamics by coordinating multi-step production and reproducing successful strategies at scale.

A multilevel ecosystem perspective

Scientific impact emerges through interactions among researchers, teams, institutions, funders, journals, infrastructures, communities, and AI systems. Assessment is part of this causal system: what NIH measures changes what scientists produce, disclose, share, verify, and avoid. Effects also travel across levels. A shared dataset may look modest at the individual level but generate substantial downstream spillovers. A highly visible line of research may look successful locally while narrowing a portfolio's scientific diversity.

This proposal extends the sociotechnical ecosystem perspective I developed in "A Sociotechnical Ecosystem Perspective of Sharing Economy Platforms," a chapter in the Cambridge University Press volume Reengineering the Sharing Economy: Design, Policy, and Regulation, for which I served as lead editor. That perspective treats system boundaries, externalities, affected stakeholders, information, and governance as interdependent design choices. DOI: https://doi.org/10.1017/9781108865630.003.

The impact-pathway instrument

Rather than requesting a longer inventory of outputs, NIH should pilot a short, structured impact-pathway statement:

accountable contribution -> reliable result or capability -> independent validation -> reuse or adoption -> NIH mission or public value.

Respondents would identify which links are relevant and label each as "observed," "projected," or "not yet observable." They would name the evidence supporting each claimed link and state important uncertainties. Reviewers would judge the credibility and strength of the relevant links, not reward the mere number of links completed.

Five broad evidence categories could make the pathway operational:

1. Accountable contribution and provenance, including human and AI roles, human validation, and responsibility for consequential decisions.

2. Validation, correction, and stewardship, including replication, robustness, negative findings, error correction, version histories, and maintained research artifacts.

3. Spillovers, diffusion, and bridging, including independent reuse, avoided duplication, cross-institutional adoption, and knowledge crossing disciplinary or sectoral boundaries.

4. Capability, exploration, and portfolio diversity, including workforce development, protected exploratory options, new methods, and bridges across research communities.

5. Translation and public value, including validated technical, regulatory, clinical, policy, entrepreneurial, accessibility, and adoption pathways.

Diagnosis-to-design logic

Each instrument should answer a specific failure mechanism. Attribution and information asymmetry require structured contribution descriptions and AI provenance. Verification scarcity requires recognition and funding for replication, correction, and stewardship. Spillovers and network effects require following value beyond the originating team through independent reuse, bridging, diffusion, and adoption. Cumulative advantage and collective perception require program-level attention to concentration, methodological and topical diversity, and the distribution of opportunity. Delayed and cross-level outcomes require matching the evidence horizon and reward to the level at which value is created. Indicator endogeneity requires treating assessment criteria as revisable interventions.

Use a menu, not a checklist. Relevant evidence will depend on the contribution, field, time horizon, and career stage. Where quantification is appropriate, NIH should prefer verifiable and denominated signals, such as the share of users independent of the originating team, rather than raw downloads, mentions, patents, startups, or revenue. No universal set of indicators should be collapsed into a cross-field league table.

Implementation

NIH can begin without creating an entirely new reporting architecture. The current Biographical Sketch Supplement already collects narrative Contributions to Science. NIH could pilot a concise impact-pathway module in one shared-resource program, one team-science program, and one investigator-centered mechanism.

Each pilot could:

1. Request an impact-pathway statement of no more than 500 words, with evidence and an observation-status label for each relevant link.

2. Train reviewers using anchored examples and ask them to identify strong links and inferential gaps.

3. Match recognition or funding consequences to the evidenced link and appropriate level. For example, portfolio diversity should guide program management rather than become an individual score.

4. Monitor burden, reviewer agreement, team and institutional distributions, independent reuse and replication, topic and method diversity, and signs of gaming.

5. Compare outcomes over two or three cycles using staggered introduction or matched programs where feasible, publish aggregate findings, and revise or retire indicators that produce burden, inequity, homogenization, or strategic distortion.

This adaptive approach is consistent with my research on information modulation in multi-agent AI systems, which demonstrates how targeted changes in an information environment can alter collective outcomes without controlling each actor or changing the underlying interaction network. See Chen, Ilami, Lorè, and Heydari, "Adaptive Information Modulation: Designing Governance Mechanisms for Multi-Agent AI Systems," Journal of Mechanical Design 148:041708 (2026), https://doi.org/10.1115/1.4070755.

The goal should not be a perfect metric. It should be an assessment system that creates better incentives, recognizes enabling and distributed contributions, detects its own behavioral and distributional effects, and learns as science and AI coevolve.

Description: This attachment provides an integrated response to the RFI, proposing that NIH treat scientific-impact assessment as an adaptive intervention in a multilevel research ecosystem. It explains how AI is changing the economics and information structure of science; examines attribution, verification scarcity, spillovers, network effects, collective perception, and cross-level consequences; and proposes a causal impact-pathway framework, corresponding indicators and incentives, and a feasible NIH pilot.

280. American Medical Informatics Association (AMIA)

I am responding: On behalf of an organization

Name: Tayler Williams

Name of Organization: American Medical Informatics Association (AMIA)

Type of Organization: Professional Organization/Association

Role: Other

Role - Other: Senior Manager, Public Policy

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: AMIA recommends that NIH more explicitly reward research practices that improve rigor, reproducibility, replication, and generalizability. Current incentives often favor novelty and publication volume, while undervaluing replication studies, methods papers, negative or mixed findings, and cross-site validation. NIH defines scientific rigor as the strict application of the scientific method to ensure robust and unbiased design, methodology, analysis, interpretation, and reporting, and NIH recognizes reproducibility as essential to validating scientific findings.

This is particularly important in informatics, where AI models, clinical decision support tools, digital health interventions, and workflow changes may perform differently across health systems, EHR environments, populations, and implementation contexts. AMIA recommends that NIH recognize methods papers, protocols, analytic code, workflow descriptions, model documentation, implementation guides, technical validation reports, and reproducibility packages as important scientific outputs.
NIH should also create dedicated funding opportunities, study sections, or review criteria for replication and reproducibility research. These applications should not be disadvantaged because they are not "novel". Instead, they should be valued for confirming, refining, or challenging prior findings. The Annual Review of Medicine article on reproducibility failure emphasizes that reproducibility problems arise across the conduct, reporting, and review of biomedical research and require interventions aimed at improving research practices and incentives.

AMIA also recommends that NIH pilot adversarial collaboration in areas where legitimate scientific disagreement persists despite accumulated evidence. Adversarial collaboration brings investigators with opposing views together to jointly design, analyze, and publish studies, often with a neutral arbiter, and aligns with open science practices such as transparency, preregistration, and shared analysis. NIH could test low-cost adversarial reanalysis of existing clinical trial, cohort, EHR, or claims data using shared protocols and publication commitments regardless of outcome.

2. Data, Software, and Model Sharing: AMIA supports NIH's efforts to incentivize data, software, and model sharing as core components of scientific impact. NIH's Data Management and Sharing Policy recognizes that sharing scientific data accelerates discovery by enabling validation, access to high-value datasets, and reuse for future studies.

However, sharing is not simply a compliance activity. Preparing data, software, and models for meaningful reuse requires infrastructure, documentation, governance, privacy protections, and long-term stewardship. NIH should allow applicants to budget adequately for data curation, software maintenance, repository fees, documentation, privacy-preserving infrastructure, and technical support.

For informatics, key outputs include well-documented datasets, data dictionaries, open-source software, analytic code, common data models, validated AI models, model cards, implementation guides, APIs, terminology mappings, and reproducible workflows. NIH should recognize these as scholarly contributions when they enable validation, replication, secondary analysis, or future research.

3. Training and Mentorship: AMIA recommends that NIH broaden how it measures training and mentorship impact. Scientific impact includes preparing researchers, clinicians, informaticians, data scientists, implementation leaders, and team-science contributors to conduct rigorous, transparent, and reproducible research.

Informatics training is especially important because modern biomedical research increasingly depends on data quality, interoperability, AI, software, workflow design, human factors, privacy, implementation science, and responsible data stewardship. NIH should reward mentorship that prepares trainees to work in collaborative, data-driven research environments and to apply reproducible methods across settings.

Mentorship impact should not be measured only by the number of trainees who obtain academic faculty positions. NIH should also recognize trainees who contribute to health systems, public health agencies, government, industry, standards organizations, nonprofit research settings, community-based organizations, and science policy roles.

4. Collaboration: AMIA recommends that NIH better reward collaboration and team science. Informatics research often depends on multidisciplinary teams that include clinicians, informaticians, data scientists, statisticians, software developers, implementation scientists, standards experts, patients, caregivers, public health professionals, and health system leaders.

Traditional authorship and citation metrics may not capture essential team contributions such as data engineering, software maintenance, workflow redesign, standards development, implementation leadership, governance, patient engagement, and cross-site coordination. NIH should encourage applicants and reviewers to identify and value these contributions.

AMIA also recommends that NIH support multi-site studies and implementations that improve generalizability and reproducibility. In informatics, a tool that works in one institution may not work elsewhere unless it is tested across different data systems, populations, clinical workflows, and organizational contexts.

NIH should also consider adversarial collaboration as a structured form of team science for contested questions. A recent JAMA Viewpoint notes that adversarial collaboration is most valuable where scientific disagreement is genuine and consequential and where existing studies are perceived as biased

or inadequate. Such collaborations may include joint meta-analyses, mutually accepted reanalysis of large datasets, codesigned prospective studies, or modeling projects that test how conclusions shift under different assumptions.

5. Entrepreneurship and Translation: AMIA recommends that NIH define translation broadly to include not only commercialization, but also implementation into clinical care, public health, research infrastructure, policy, standards, and health system operations.

In informatics, translation may include deployment of clinical decision support, adoption of interoperability standards, implementation of AI governance frameworks, integration of patient-facing digital tools, use of common data models, or incorporation of research findings into health system workflows. These activities may not always result in patents or products, but they can have substantial public value.

NIH should reward research that demonstrates whether innovations work safely, effectively, equitably, and sustainably in real-world settings. For AI and digital health tools, translation should include external validation, workflow integration, usability, safety monitoring, bias assessment, patient and clinician experience, and evidence of sustainability over time.

6. Foundational Scientific Exploration: AMIA recommends that NIH continue to reward foundational research, including high-risk and infrastructure-focused work whose impact may take years to become visible. In informatics, foundational contributions often include data standards, terminology systems, common data models, software platforms, knowledge representation methods, evaluation frameworks, and methods for measuring data quality, bias, provenance, and interoperability.

These contributions may not produce immediate clinical outcomes, but they enable future discovery, replication, implementation, and translation. NIH should ensure that impact assessment does not undervalue foundational informatics work simply because its benefits are indirect, distributed, or long-term.

AMIA also recommends that NIH support greater standardization of outcome reporting in informatics. Meta-analysis remains underused in the field partly because studies often use inconsistent definitions, measures, outcomes, and implementation descriptions. NIH states that common data elements enhance data interoperability and reproducibility, supporting more consistent data collection across studies and sites.

7. Public Impact: AMIA recommends that NIH measure public impact in ways that reflect both near-term and long-term value. Public impact should include contributions to health outcomes, clinical practice, public health, patient safety, accessibility, trust, health equity, economic value, workforce capacity, and responsible use of technology.

In informatics, public impact may include safer clinical decision support, more transparent AI, improved data exchange, reduced documentation burden, better patient access to health information, improved public health reporting, stronger data quality, or tools that help clinicians and patients make better decisions.

NIH should recognize that public impact often takes time to emerge. Foundational work in data standards, software, methods, infrastructure, and reproducibility may enable future breakthroughs even if immediate outcomes are difficult to measure. NIH should therefore avoid relying only on short-term metrics.

8. Other Areas: N/A

281. Greater Gulf Coast Translational Science Alliance

I am responding: On behalf of an organization

Name:

Name of Organization: Greater Gulf Coast Translational Science Alliance

Type of Organization: Other

Type of Organization - Other: CTSA consortium

Role: Organizational Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Recognize Contributions to Rigor and Reproducibility. Rigor and reproducibility should be evaluated based on the quality of the scientific process, rather than primarily on the number of publications, citations, or other traditional measures of academic productivity. Important indicators include appropriate study design; pre-specification of analyses when appropriate; transparent reporting of statistical methods; appropriate handling of missing data; assessment of model assumptions; sensitivity analyses; validation of analytical approaches; independent replication; transparent sharing of data and code when possible; and sufficient documentation to allow others to understand and reproduce an analysis.

Importantly, these practices often require substantial contributions from individuals whose work may be difficult to capture fully using traditional measures of scientific impact. Biostatisticians and data scientists frequently contribute critically to study design, analytic strategy, statistical methodology, data management, software development, validation, and interpretation. These contributions may determine whether a study produces reliable evidence, yet these contributions may be reflected in ways not fully captured by conventional authorship, funding, or citation indicators. NIH should explicitly recognize methodological contributions as important indicators of scientific impact.

NIH could encourage this recognition by asking investigators to document specific contributions to rigor and reproducibility in grant applications, progress reports, and other assessments. Examples could include development of reproducible analytical workflows; creation and validation of statistical methods or software; independent validation of findings; development of data standards or documentation; replication of important findings; and substantive contributions to study design and analytic integrity. Contribution-based assessment would better reflect the increasingly team-based nature of biomedical research.

Measure Reproducibility. Reproducibility should be considered multidimensional. A single metric is unlikely to adequately capture whether research is rigorous or reproducible. NIH should consider a portfolio of indicators that reflect both the practices used to produce research and the subsequent ability of others to reproduce, validate, replicate, or extend the findings.

Potential indicators include availability of analysis code; appropriate documentation of computational workflows; availability and appropriate governance of data; use of version-controlled software; independent validation; replication studies; transparent reporting of deviations from prespecified analyses; and reuse of data or software by independent investigators. Importantly, these measures should distinguish between computational reproducibility, methodological reproducibility, and biological or clinical replication, which are related but not interchangeable concepts.

NIH may wish to design metrics that emphasize meaningful practice rather than checklist-style compliance. Requiring a reproducibility checklist, for example, does not necessarily produce reproducible science. Measures should emphasize meaningful practices and outcomes rather than compliance alone.

Recognize Replication and Negative Findings. Current incentives often place greater emphasis on novel and positive findings. This may provide fewer incentives for replication studies, confirmation of established findings, and publication of well-conducted studies that produce null or negative results. NIH should consider mechanisms to increase the scientific value assigned to these contributions.

A rigorous replication study that confirms or challenges an influential finding may have greater long-term scientific value than a novel but poorly validated result. Similarly, a well-designed study producing a null result can prevent substantial resources from being devoted to an ineffective intervention or unsupported hypothesis. NIH should explicitly recognize these contributions when assessing scientific impact.

2. Data, Software, and Model Sharing: Reward usable and reusable scientific products. NIH should evaluate datasets, software, and models based not only on their availability or citations, but also on their quality, documentation, interoperability, independent reuse, external validation, and contribution to subsequent discoveries or health improvements. Indicators could include the number and diversity of independent users, successful reproductions, incorporation into new studies, validated clinical or public-health applications, and the populations ultimately benefiting from the work. Publication and citation metrics remain valuable, particularly for foundational research, but they should be complemented by measures of translational and public impact.

Fund the full lifecycle of sharing. Data sharing is not a one-time act of depositing files. It requires data engineering, quality control, curation, metadata, privacy and security review, storage, transfer, software maintenance, user support, model monitoring, and eventual archiving or deprecation. Current budgeting approaches may not always fully account for this operational complexity and the substantial contributions of research software engineers, informaticians, data stewards, system administrators, security specialists, and project managers. These personnel and activities should be explicitly recognized as allowable and reviewable scientific costs. Shared infrastructure and post-award stewardship mechanisms would help prevent each project from rebuilding the same capabilities independently.

Develop consistent, machine-actionable federal standards. NIH should promote common requirements for metadata, data dictionaries, provenance, versioning, quality-control reporting, persistent identifiers, and programmatic access. NIH-supported repositories are extraordinarily valuable, but differences in submission, validation, and export requirements can create technical burdens that greater

standardization may reduce. Greater standardization, reusable validation tools, bulk export capabilities, and standards-based application interfaces—including emerging agent-compatible interfaces—would improve accessibility without lowering scientific or privacy standards.

Advance patient-centered governance and consistent privacy protections. Federal policy should move toward comparable privacy, security, transparency, and accountability standards for health-related data held by academic institutions, healthcare organizations, technology companies, and other commercial entities. The objective should not be to weaken protections for clinical data, but to reduce regulatory asymmetries while enabling responsible research. Individuals should have meaningful opportunities to authorize, direct, and, where feasible, revoke research and AI uses of their data across institutions. NIH could support both a strong patient-control principle and carefully evaluated pilot programs for patient-directed authorization, portability, and consent.

Recognize that responsible sharing is context dependent. A single sharing model will not fit omics, imaging, EHR, real-world evidence, clinical trials, and continuously generated device data. Cancer and molecular research may require particular attention to reference versions, biospecimen provenance, genomic identifiability, and reproducible analytic pipelines. Aging, longitudinal EHR, and real-world evidence research may require continuing attention to consent, decision-making capacity, changing health status, repeated observations, data generated outside healthcare settings, and model performance over time. NIH should define common principles while permitting discipline- and modality-appropriate controlled-access approaches.

Balance sharing with commercialization and intellectual property. Incentives should recognize that immediate unrestricted release can conflict with patent protection, licensing, public-private partnerships, and development of products that ultimately benefit patients. NIH could permit clearly defined and time-limited sharing delays, followed by release of appropriate validation data, documentation, and nonproprietary components. Investigators should not be penalized for using justified controlled-access or staged-release models when necessary to protect participants or preserve a credible path to translation.

Modernize scientific credit and reporting. Contributors who build, validate, document, maintain, and support shared resources should receive visible academic and grant-review credit. As AI increasingly assists manuscript production, evaluative emphasis should shift from publication volume and narrative length toward auditable technical artifacts, including versioned methods, quality-control results, software and model documentation, limitations, failure modes, and reproducible workflows. This can be incorporated into existing publication and grant-review practices without displacing the importance of rigorous peer review.

A practical NIH implementation strategy could begin with common core indicators across programs—usability, independent reuse, validation, stewardship, and demonstrated health or scientific benefit—while allowing field-specific measures. Pilot implementation through CTSA hubs and other shared-infrastructure programs could assess feasibility, administrative burden, and unintended consequences before broader adoption.

3. Training and Mentorship: Assessing the impact of NIH-supported training and mentorship programs requires recognition that outcomes often emerge over long periods of time and across diverse career pathways. The value of mentorship and workforce development may not be apparent during the training period itself; therefore, evaluation frameworks should incorporate longitudinal tracking of mentees across multiple career stages and sectors. Success should be measured broadly and not be limited to traditional academic outcomes or NIH funding alone.

Several indicators could help quantify the effectiveness of NIH training and mentorship investments. These include the ability of former trainees to secure research funding during and after their training, including support from NIH, other federal agencies, foundations, and industry partners. Evaluation should also recognize contributions to funded research projects beyond principal investigator roles, as many scientists build successful careers through collaborative and team-based science.

Retention in scientific and biomedical careers is another important outcome. Metrics should capture whether trainees remain engaged in research-related professions, regardless of whether they pursue careers in academia, government, nonprofit organizations, or industry. Tracking the attainment of a first independent position, whether as faculty or as a scientific leader in another sector, would provide an important indicator of career progression.

The long-term influence of mentoring relationships should also be considered. Former trainees who go on to mentor others, serve as mentors on training grants, or support mentees who subsequently receive career development awards or other research funding demonstrate a multiplier effect of NIH's training investments. These downstream impacts may be among the strongest indicators of mentorship success.

Scholarly productivity should be evaluated using measures that reflect the collaborative nature of modern biomedical research. In addition to first- and senior-author publications, contributions to team science efforts, including publications where a trainee serves in a middle-author role, should be recognized. Similarly, presentations at scientific conferences, including posters, platform presentations, invited talks, and presentation-related awards, can provide valuable evidence of professional development and scientific engagement.

Finally, there is significant concern regarding the reduced availability of support for postdoctoral training as T32 training opportunities are phased out or reduced. This change may disproportionately affect postdoctoral researchers without clinical degrees, who often rely on these programs for protected research time, mentorship, and career development. Any assessment of NIH's training and mentorship ecosystem should consider the potential impact of these changes on the development and retention of the future biomedical research workforce.

4. Collaboration: Reward Team Science. Biomedical research is increasingly interdisciplinary, but academic reward systems often continue to emphasize individual achievement. This can create gaps between collaborative scientific practice and traditional evaluation approaches. NIH should place greater emphasis on the contributions of research teams and on the complementary expertise required to answer complex biomedical questions.

Team science could be evaluated using structured contribution statements that identify the scientific,

methodological, computational, data, mentoring, and leadership contributions of individual team members. Such information could complement, rather than replace, traditional publication metrics. NIH could also encourage institutions to incorporate contribution-based assessments into promotion and tenure processes.

Particular attention should be given to scientific infrastructure and methodological contributions. Individuals who develop datasets, software, analytical pipelines, statistical methods, repositories, or other resources may enable hundreds of subsequent studies while their contributions may be difficult to recognize through conventional metrics. These contributions should be viewed as scientific products rather than merely as service activities.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: To strengthen the public impact of biomedical research, NIH should encourage changes in the culture and reward systems that shape investigator behavior and institutional priorities. Current academic promotion and tenure processes often place primary emphasis on individual achievements, grant funding, and publication metrics, which can undervalue activities that contribute substantially to societal benefit. NIH could help drive change by promoting evaluation frameworks that recognize and reward reproducible and transparent research practices, team-based and collaborative science, and community-engaged research. Importantly, these activities should be considered core expectations of high-quality research rather than optional or supplemental contributions. Indicators of success could include demonstrated adoption of open science practices, participation in multidisciplinary research teams, and meaningful engagement with patients, communities, or other stakeholders throughout the research process.

NIH should also consider mechanisms to ensure that research is responsive to the needs of those most likely to benefit from its outcomes. One approach would be to include patients, community representatives, caregivers, public health practitioners, or other end users in grant review processes when appropriate. Their participation could help assess whether proposed projects address community-identified priorities, have a realistic pathway to implementation, and are likely to produce meaningful societal impact. Potential indicators could include evidence of stakeholder involvement in study design, dissemination activities tailored to affected communities, and measurable adoption of research findings in practice, policy, or public health initiatives.

Achieving these goals will also require changes in workforce development. Training and mentorship programs for graduate students, postdoctoral researchers, and early-career faculty should provide skills and experiences that support community engagement, collaborative science, research transparency, and translation of findings into real-world settings. Evaluating the effectiveness of these efforts could include tracking participation in community-engaged research activities, interdisciplinary collaborations, open science initiatives, and other practices that contribute to public impact.

Collectively, these approaches would encourage a research ecosystem that values not only scientific discovery, but also the extent to which research is rigorous, accessible, responsive to societal needs, and ultimately beneficial to the public. As NIH develops public impact metrics, it will be important to ensure

that they are feasible across disciplines, career stages, and institution types, while avoiding unintended consequences such as excessive administrative burden or overly narrow definitions of impact.

8. Other Areas: Training and Mentorship as Scientific Impact

Scientific impact also includes the development of people who will conduct future research. Training and mentorship should therefore be recognized as meaningful scholarly contributions rather than secondary activities.

NIH-supported training programs can have effects that extend well beyond the immediate products of a project. Effective training in experimental design, statistical reasoning, data science, reproducible research, responsible conduct, and scientific communication can influence the quality of research produced by trainees throughout their careers. NIH should consider measures such as trainee progression into research careers, subsequent research contributions, development of methodological expertise, and continued engagement in rigorous and reproducible research.

Importantly, these measures should be feasible across career stages. Expectations for an undergraduate trainee, early-career investigator, established investigator, and senior team leader should not be identical. Evaluation systems should recognize developmental trajectories rather than relying on a single set of productivity metrics.

The Essential Role of Biostatistics and Data Science

The increasing complexity of biomedical research makes methodological expertise essential to scientific rigor. Biostatisticians and data scientists contribute across the research lifecycle, including formulation of research questions, study design, sampling, data management, statistical analysis, computational methods, interpretation, visualization, and reproducibility.

NIH should explicitly recognize these contributions when defining scientific impact. In many collaborative studies, the quality of the statistical design and analysis is foundational to the validity of the scientific conclusions. Yet conventional metrics can make these contributions difficult to document and reward.

One potential solution is for NIH to encourage standardized contribution reporting that goes beyond authorship. For example, grant applications and reports could describe contributions to study design, statistical methodology, computational methods, data resources, software, validation, and reproducibility. Such documentation would provide reviewers and institutions with a more complete picture of an individual's scientific contributions.

Avoiding Unintended Consequences

Any new system for measuring scientific impact should be implemented cautiously. Quantitative metrics are attractive because they are easy to compare, but overreliance on metrics may create incentives that do not fully align with the intended goals. For example, investigators could prioritize activities that are easy to count rather than activities that are scientifically valuable.

Metrics should therefore be used as evidence to inform expert assessment rather than as automatic measures of scientific worth. NIH should also periodically evaluate whether new measures produce unintended consequences, including increased administrative burden, inequities across disciplines, or incentives for checklist-style or low-value compliance.

Feasibility is particularly important. A system requiring extensive documentation for every contribution could disproportionately burden investigators at institutions with fewer administrative resources and could reduce time available for research activities. New measures should be simple enough to implement while still capturing meaningful contributions.

Recommendations

-Evaluate scientific contributions rather than relying primarily on publications and citations.

-Explicitly recognize methodological, statistical, computational, data, software, and infrastructure contributions.

-Reward team science through structured contribution reporting.

-Recognize replication, validation, and well-conducted null or negative studies.

-Measure reproducibility through meaningful research practices and outcomes rather than checklist compliance.

-Recognize training and mentorship as contributions to the long-term scientific enterprise.

-Use multiple measures rather than a single quantitative metric.

-Ensure that measures are appropriate for different disciplines, career stages, and types of scientific contribution.

-Minimize administrative burden and continually evaluate new metrics for unintended consequences.

-Use metrics to support informed expert judgment rather than replacing expert assessment.

The biomedical research enterprise increasingly depends on interdisciplinary teams, sophisticated data resources, computational methods, and methodological expertise. A system for measuring scientific impact should evolve accordingly. Recognizing the people and practices that make research rigorous, reproducible, collaborative, and useful would strengthen the scientific enterprise and better align incentives with NIH's goal of producing trustworthy knowledge that advances human health.

282. Fred Ledley, M.D.

I am responding: On behalf of myself

Name: Fred Ledley, M.D.

Name of Organization: Bentley University, Center for Integration of Science and Industry

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: Rigor and Reproducibility. We agree that reproducible and replicable biomedical approaches are foundational to rigorous research. We caution, however, that a distinction needs to be drawn between exploratory research aimed at hypothesis generation and confirmatory research aimed at hypothesis testing. The National Academy of Sciences describes the distinction thus: "Confirmatory research is research that starts with a well-defined research question and a priori hypotheses before collecting data; confirmatory research can also be called hypothesis testing research. In contrast, researchers pursuing exploratory research collect data and then examine the data for potential variables of interest and relationships among variables, forming a posteriori hypotheses; as such, exploratory research can be considered hypothesis generating research."[23]
One way to describe the distinction is that exploratory research is directed towards "unknown unknowns." Boguslav et al wrote: "Research begins with a question. It progresses through accumulating knowledge such that a previously unexplored subject (an unknown unknown) becomes an active research area exploring the questions (known unknowns), until a body of established facts emerges (known knowns)." [57]From a risk perspective, Derbyshire et al. describe unknown unknowns as "…elements of knowledge that would be of importance to assessing the focal risk if they were known but which are presently unknown to anyone including those tasked with making the risk assessment, who, furthermore, lack awareness of not knowing them."[58] Only when something is "known," can it be subject to rigorous, controlled hypothesis testing or considered for utility in applied or translational research.
From the perspective of reproducibility and replicability, exploratory research may yield observations that are, at least in part, irreproducible or irreplicable since it is impossible to control for unknown unknowns. In fact, failure to replicate observations in exploratory research can be useful in that it may expose the presence of unknown factors and prompt further exploratory studies aimed at their characterization. We caution that arbitrary standards focused on reproducibility or replicability could be inherently ineffective if applied to exploratory research could compromise the process of exploration and create impediments to progress.
We endorse efforts to improve the rigor of confirmatory research aimed at testing discrete hypotheses with potential applications in healthcare or industrial development.
We propose the following:

• NIH should increase its funding for education in research methodologies, particularly at the graduate level;

• NIH should assure that investigators have adequate funding for their proposed research and do not feel pressured to “cut corners” to satisfy budget constraints;
• Free dissemination and critique of ideas without political interference or consequences is central to our communal responsibility to conduct rigorous research;
• Federal health agencies need to set an exemplary standard in promulgating only applications of research that has been replicated, distinguishing early-stage findings that are only exploratory, and calling out information that is merely observational or conjecture.

2. Data, Software, and Model Sharing: Data, Software, and Model Sharing. We agree that “sharing scientific data and tools accelerates biomedical research discovery” and urge the NIH to continue efforts to make data sharing a normative practice. We believe the NIH has long set a high standard for making data widely available for scientific research and applications. The sharing of publicly funded research, data collections, analytical tools, and even is increasingly normative. We would emphasize the following:

• The confidentiality of personal, clinical data must remain an overriding principle without compromise;
• Data and research tools made available by the federal government must set a high standard and must represent a complete and accurate representation of data that is collected, or research results generated with public funds. These shared collections cannot be subject to ideological modification or censoring. We note that these collections must be complete datasets and are likely to include content that is not uniformly considered to be accurate or useful. Only information that is formally retracted after scientific critique should be eliminated.

3. Training and Mentorship: Training and Mentorship. We applaud NIH’s continued commitment to training a professional and technical workforce. We call on the NIH to ensure adequate, continuous funding for these programs.

4. Collaboration: Collaboration. We applaud the NIH’s commitment to collaboration. We recommend that the NIH consider collaborations beyond the traditional disciplines of medicine, biology, biochemistry, psychology, informatics, analytics, and physics and promote collaborations also between scientists and individuals expert in management, economics, finance, and public policy. Such collaborations could significantly help bridge the gap between laboratory science and applications in medicine.

5. Entrepreneurship and Translation: Entrepreneurship and Translation. While most discussions of entrepreneurship and the transition from basic to applied or translational science focus on issues of finance, we believe a critical, overlooked aspects of entrepreneurship and translation is technology readiness.
Evidence in engineering suggests that the maturation of technology has predictive value for subsequent development performance including both technical risk, cost, and timelines. The GAO, for example, finds that entering product development with immature technologies leads to greater costs and schedule delays and have higher development costs than those based on mature technologies. This has led agencies including DOD and NASA to incorporate technology readiness assessments in their determination of whether novel technologies are sufficiently mature to proceed into development and production. Our research suggests that pharmaceutical development is similarly dependent on the maturation of the underlying basic science, that few, if any targeted therapeutics are successfully developed until this science passes a passes a quantitatively-defined threshold of maturation, and that

the timelines of clinical development are significantly longer if clinical development commences before this threshold is reached.[6-9, 22, 59] Our work further suggests that innovation can stall when private capital is invested in immature technologies and is, therefore, dissuaded from subsequent investments in the space as it matures. Given the high risk, limited availability, high cost of investment capital for early-stage development, greater focus on characterizing the readiness of basic science for development could radically reshape the landscape of innovation and promote more rapid translation of the most promising technologies.

6. Foundational Scientific Exploration: Foundational Scientific Exploration Our research suggests that the primary contribution of NIH funding to pharmaceutical innovation is its support of basic science. Basic research conducted with federal funding has an essential role in the innovation ecosystem as the primary driver of discovery and development and a source of economic stimulus.

Basic science is defined as "…experimental or theoretical work undertaken primarily to acquire knowledge….without any particular application or use in view" [60], recognizing that NIH spending is inherently "use-inspired" [61]. Our work further suggests that virtually all of this work is performed in public institutions (as opposed to industry) and as noted above, the maturation of this body of research may be a critical determinant of product development success, timelines, and costs. [6-9, 22, 59] Initiating applied, translational, or development work based on an immature body of basic research increases the risk, timelines, and costs to industry.
We would argue that initiatives focused on specific diseases or conditions can only succeed if there is a commitment to establishing a mature body of basic research before considering translation or development. We have studied innovation following passage of the National Cancer Act of 1972. This Act was long considered to have been a failure due to the absence of new therapies or substantial improvement in outcomes for decades after the Act was passed. Our analysis suggest that the Act was, in fact, highly successful. The War on Cancer was launched at the outset of an era that saw the emergence of molecular biology, genomics, and biotechnology and the Act was instrumental in funding much of this foundational work. Our analysis shows that it was only when the foundational, basic research in these areas reached a threshold of maturation that new product began to appear. The last 20 years have seen an explosion of new therapies for cancer and dramatic advances in both treatments and outcomes. [22] Similarly, many early failures of candidate therapies for Alzheimer's involved candidate products that entered development before research on Amyloid was mature while the recent successes uniformly entered development only after that body of research was established. [62]

We would also call attention to the economic efficiencies created by public support for basic science. As described below, the timeline between basic research on drug targets and first approval of a drug associated with that target can be as much as 25 years. [21, 63, 64] This long timeline, coupled with the high cost of capital in industry (10.5%) makes investment in basic science untenable for industry. It is possible to estimate the cost savings provided to industry by NIH funding for basic science that is then transferred to industry through the public domain, workforce training or licensing. Industry typically calculates drug development costs as the sum of spending on the approved drug, an estimate of spending on failed candidate products, and a 10.5% cost of capital. Applying the same calculation to assess the value provided to industry by public funding of basic research, we estimate that the average cost saving to industry is as much as $2.9 billion, an amount roughly equivalent to reported development costs. [20, 21, 65]

Public funding for basic science also creates economic efficiencies through spillovers. Studies suggest that an average of three novel therapeutics are developed for each biological target. [19, 66] This is evident, for example, by the observation that most NIH-funded projects related to at least one drug approval are typically associated with multiple approved drugs. This means that the public investment in basic science is ultimately distributed across multiple approvals, reducing the effective public investment/drug and multiplying the returns from product development.

We believe that a continued government commitment to funding basic science is essential for the future success of pharmaceutical innovation and commercialization. Initiatives that draw resources away from basic science toward applied or translational research in any specific therapeutic or disease area, no matter how well intentioned, represent a threat to future progress.

7. Public Impact: N/A

8. Other Areas: See uploaded document including Supporting Research and explanatory notes

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Ledley-Comments-on-NOT-OD-26-087-NIH-request-for-information.pdf

Description: Comments on NOT-OD-26-087

Date submitted:	August 19, 2026
Title:	Promoting Scientific Impact at the NIH
Submission to:	National Institutes of Health NIH Office of Science Policy
Response to notice:	NOT-OD-26-087: Request for Information on Measuring and Rewarding Scientific Impact, June 18,2026
Comments by:	Fred D. Ledley, M.D. Director, Center for Integration of Science and Industry Professor, Departments of Natural & Applied Science, Management Bentley University Waltham, MA USA
Comment tracking number:	TBD
Document ID:	TBD

Introduction

We are honored to offer our comments on Promoting Scientific Impact at the NIH in response to NOT-OD-26-087: Request for Information on Measuring and Rewarding Scientific Impact.

These comments are informed by my professional experience as the founding director of the Center for Integration of Science and Industry at Bentley University,[i] which has a mission to **"advance the translation of scientific discoveries for public value by engaging science, business, and policy in transdisciplinary dialogue, scholarship, and education."** Our approach is to bring together the perspectives of natural science, business, analytics, economics, law, and policy to better understand, and advance, the path of translational science and the business models that create public value from the fruits of scientific discovery, primarily through pharmaceutical innovation. **A primary focus of our research is on characterizing sources of innovation capital including funding from the public sector, product sales, equity markets, and debt markets, their contributions to pharmaceutical innovation, and their required returns.**

We applaud the overarching aim of the NIH's initiative to "incentivize and reward rigorous research that delivers value to the American public" and "…better align its measures of success with its core mission; generating knowledge that enhances health, lengthens life, and reduces illness and disability." We also agree that many current metrics of scientific success – as well as many criticisms of the NIH – involve "…indicators of productivity, such as publication counts and citation metrics."

We are, however, concerned about disruptions of the NIH over the past 20 months that have reduced the number of basic, applied and translational research projects awarded, reduced the total amount of research funded by the NIH, interfered with peer review, reduced or eliminated Notices of (new) Funding Opportunities, disrupted research careers and collaborations, interfered with the dissemination of research findings to the public, imposed censorship and political criteria for funding, interrupted workforce and professional training programs, and led to the departure of many experienced and dedicated public servants and loss of their accumulated expertise. **We believe the NIH's top priority must be to restore normal operations including a robust process for soliciting, reviewing, and funding scientifically driven research that has long provided the foundation for advances in American prosperity, influence, health, and wellbeing.**

Our perspectives

Our perspectives are based on three principles:

1. **The NIH has a significant impact in advancing healthcare and health outcomes for Americans.**

   Other commentators are more qualified to comment on the NIH's broad impacts on improving healthcare practice and reducing morbidity and mortality in the United States. From a personal perspective, the table below illustrates some of the clinical outcomes I experienced in my pediatric training and the advances that have been achieved to the present day.[ii]

Selected pediatric outcome	Historical	years	Contemporary	years
Infant mortality rate	13/1,000 live births	1980	5.6/1,000 live births	2023
Likelihood of death before age 20a	2.3% live births	1975	1.1% live births	2022
Minimum gestational age considered potentially viable	28 weeks	1971	22 weeks	2013-2018
Childhood acute lymphoblastic leukemia (ALL) survival	57% 5-year survival	1975	92% 5-year survival	2014-2020
Childhood acute myelogenous leukemia (AML) survival	22% 5-year survival	1975-1984	68% 5-year survival	2005-2014
Cystic fibrosis survival	11 years	1978	>65 years	2021-2025
Critical congenital heart disease survival	68% 1-year survival	1979-1993	83% 1-year survival	births 1994-2005
Phenylketonuria (PKU) cognitive outcome (IQ)	<70	1961b	normal	post-screening
Motor vehicle crash deaths — infants	7/100,000	1980	2/100,000	2018-2019
Sudden infant death syndrome (SIDS)	1.5/1,000 live births	1979-1982	0.4/1,000 live births	2024
Measles deaths	35/year	1971-1975	<3/year	2025
Varicella (chickenpox) deaths	138 deaths/year	1973	18 deaths/year	2017-2019
Respiratory syncytial virus (RSV)-associated pediatric deaths	~500 deaths/year	1979-1997	96 deaths/year	1999-2018
Influenza and pneumonia deaths — all ages	24 deaths/100,000	1980	14 deaths/100,000	2024
Pediatric cancer deaths (age 0-19)	4.6/100,000	1980	2.2/100,000	2022

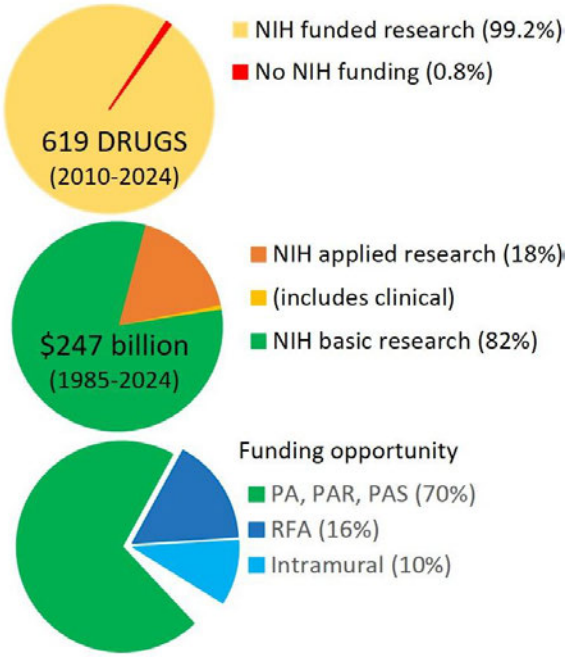
Note: data retrieved with ChatGPTPlus. All data confirmed manually or from second source. accessed August 19. 202. (a) expectancy of survival to age 20: 1975=97.8%; 2022=98.9%; (b) before newborn screening.

   These isolated data are not intended to be representative, but only to provide a snapshot of the profound improvements in pediatric health that have personally impacted American children and families in recent decades. As noted in a 2006 report from the National Academy of Sciences: "*Pediatric health is the cornerstone of the nation's well-being. … These advances would never have happened without the significant and consistent federal investment in pediatric research, supported as a whole or in part by the National Institutes of Health (NIH)."[1]* The purpose of the foregoing is not to focus these comments on pediatrics, particularly in light of longstanding concerns regarding the adequacy of NIH support for pediatric research.[3-6] Similar statements could be made about the role of the NIH in advancing all aspects of American healthcare and health.

2. **The NIH plays an essential role in pharmaceutical innovation, enabling the discovery and development of novel drugs, providing robust data resources and capabilities, and promoting new investment in the biopharmaceutical industry.**

   While industry is primarily responsible for the development, approval, and commercialization of novel pharmaceutical products,[iii] our research (detailed below) demonstrates **that the NIH funds basic or applied research related to >99% of all news drugs before their first FDA approval[iv] with an average NIH cost of ~$1.5 billion/drug, an amount comparable to the investment by industry.** [19-21]

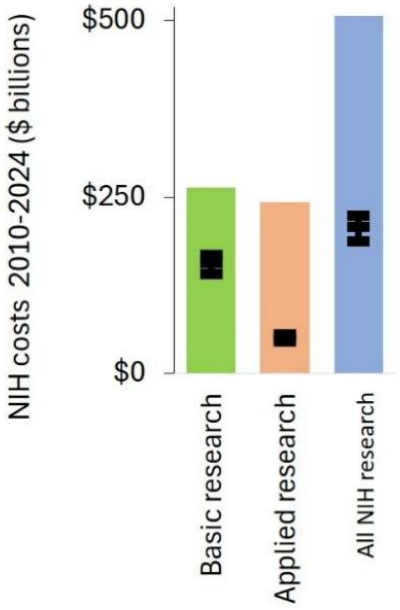
Our work also shows that >80% of this NIH funding involves basic research and that **few, if any novel drugs are approved before the body of basic research achieves a threshold of maturity**. [7-9, 22] This **mature body of basic research is not only essential to assure an useful understanding of drug targets and their involvement in both the normal state and in disease but also to replicate, refute, or refine earlier research and establish the body of reliable evidence required for drug development and regulatory approval** [7, 8, 18, 23].



Little basic research is performed in industry, and the availability of this research in the public domain, through workforce training, or through licensing both **reduces the private investment required to successfully develop new drugs and creates economic efficiencies** for society. Our work also shows that commercialization of these products **generates considerable economic activity, profits greater those of companies in most other sectors of the S&P500,[24] growth of shareholder value,[25, 26] health value for society, job creation, and further R&D activity**.[v] [31-33]

3. **The NIH is highly effective in directing research funding to projects associated with outcomes that advance healthcare practice and health. Changes in this model to prioritize applied or translational research over basic research, move to mission-driven funding mechanisms over investigator-initiated funding, or censor research in specific areas threatens the vitality of the research enterprise that has long provided the foundation for advances in American prosperity, influence, health, and wellbeing. While we endorse a process of continuing assessment and evidence-based improvement of the scientific enterprise, we believe these criticisms are unfounded and that radical reform of the NIH is unwarranted.**

Recent criticisms of the National Institutes of Health (NIH) argue that NIH spending is overly focused on basic rather than applied or translational research and structured around investigator-initiated research rather than mission-driven strategies for addressing the mission of "enhance health, lengthen life, and reduce illness and disability." [37] **Our work (detailed below) demonstrates that ~40% of all NIH funding is productively employed towards new drug approvals including ~60% of all basic science.** [38]



These studies further show that the large majority of NIH funding contributing to drug approvals is provided through investigator-initiated funding mechanisms. Moreover, our analysis suggests that **investigator-initiated mechanisms**

significantly more likely to direct research resources toward drug approvals than program-driven mechanisms. [38]

We would also note that efforts to selectively eliminate funding in selected research areas [39] based on words used in the research proposals runs the risk of creating substantial collateral damage to other research areas. Our analysis suggests that some of the words that have been subject to criticism are significantly associated in funded research projects with words describing other high-priority areas for research and innovation. This research could be inadvertently curtailed if this funding were selectively eliminated based on their word content.[40]

While we recognize there is an uncertain association between NIH funding levels and the US burden of disease, [41-46] **there is a significant association between the number new drug approvals and the US burden of disease [47, 48]** There is also extensive evidence that new drug approvals are significantly associated with increased life expectancy as well as reduced mortality, disability, and functional limitations [49-56]. In this context **the observation that a substantial fraction of NIH funding is productively employed towards drug approvals indicates that the NIH model for funding basic and applied research is effective in directing resources toward outputs that directly contribute to the mission of advancing healthcare and health outcomes**.

Information Requested

We offer the following comments on the requested information:

Rigor and Reproducibility. We agree that reproducible and replicable biomedical approaches are foundational to rigorous research. We caution, however, that a distinction needs to be drawn between exploratory research aimed at hypothesis generation and confirmatory research aimed at hypothesis testing. The National Academy of Sciences describes the distinction thus:

> "*Confirmatory research is research that starts with a well-defined research question and a priori hypotheses before collecting data; confirmatory research can also be called hypothesis testing research. In contrast, researchers pursuing exploratory research collect data and then examine the data for potential variables of interest and relationships among variables, forming a posteriori hypotheses; as such, exploratory research can be considered hypothesis generating research."[23]*

One way to describe the distinction is that exploratory research is directed towards "unknown unknowns." Boguslav et al wrote: *"Research begins with a question. It progresses through accumulating knowledge such that a previously unexplored subject (an unknown unknown) becomes an active research area exploring the questions (known unknowns), until a body of established facts emerges (known knowns)." [57]*From a risk perspective, Derbyshire et al. describe unknown unknowns as *"…elements of knowledge that would be of importance to assessing the focal risk if they were known but which are presently unknown to anyone including those tasked with making the risk assessment, who, furthermore, lack awareness of not knowing them."[58]* Only when something is "known," can it be subject to rigorous, controlled hypothesis testing or considered for utility in applied or translational research.

From the perspective of reproducibility and replicability, exploratory research may yield observations that are, at least in part, irreproducible or irreplicable since it is impossible to control for unknown unknowns. In fact, failure to replicate observations in exploratory research can be useful in that it may expose the presence of unknown factors and prompt further exploratory studies aimed at their characterization. **We caution that arbitrary standards focused on reproducibility or replicability could be inherently ineffective if applied to exploratory research could compromise the process of exploration and create impediments to progress.**

We endorse efforts to improve the rigor of confirmatory research aimed at testing discrete hypotheses with potential applications in healthcare or industrial development.

We propose the following:

- **NIH should increase its funding for education in research methodologies**, particularly at the graduate level;
- **NIH should assure that investigators have adequate funding for their proposed research** and do not feel pressured to "cut corners" to satisfy budget constraints;
- **Free dissemination and critique of ideas without political interference or consequences is central to our communal responsibility to conduct rigorous research**;
- **Federal health agencies need to set an exemplary standard** in promulgating only applications of research that has been replicated, distinguishing early-stage findings that are only exploratory, and calling out information that is merely observational or conjecture.

Data, Software, and Model Sharing. We agree that "sharing scientific data and tools accelerates biomedical research discovery" and urge the NIH to continue efforts to make data sharing a

normative practice. We believe the NIH has long set a high standard for making data widely available for scientific research and applications. The sharing of publicly funded research, data collections, analytical tools, and even is increasingly normative. We would emphasize the following:

- The **confidentiality of personal, clinical data must remain an overriding principle** without compromise;
- **Data and research tools made available by the federal government must set a high standard and must represent a complete and accurate representation of data that is collected, or research results generated with public funds. These shared collections cannot be subject to ideological modification or censoring**. We note that these collections must be complete datasets and are likely to include content that is not uniformly considered to be accurate or useful. Only information that is formally retracted after scientific critique should be eliminated.

Training and Mentorship. We applaud NIH's continued commitment to training a professional and technical workforce. We call on the NIH to ensure adequate, continuous funding for these programs.

Collaboration. We applaud the NIH's commitment to collaboration. We recommend that the NIH consider collaborations beyond the traditional disciplines of medicine, biology, biochemistry, psychology, informatics, analytics, and physics and promote collaborations also between scientists and individuals expert in management, economics, finance, and public policy. Such collaborations could significantly help bridge the gap between laboratory science and applications in medicine.

Entrepreneurship and Translation. While most discussions of entrepreneurship and the transition from basic to applied or translational science focus on issues of finance**, we believe a critical, overlooked aspects of entrepreneurship and translation is technology readiness.**

> Evidence in engineering suggests that the maturation of technology has predictive value for subsequent development performance including both technical risk, cost, and timelines. The GAO, for example, finds that **entering product development with immature technologies leads to greater costs and schedule delays and have higher development costs than those based on mature technologies.** This has led agencies including DOD and NASA to incorporate technology readiness assessments in their determination of whether novel technologies are sufficiently mature to proceed into development and production. Our research suggests that pharmaceutical development is similarly dependent on the maturation of the underlying basic science, that few, if any targeted therapeutics are successfully developed until this science passes a passes a quantitatively-defined threshold of maturation, and that the timelines of clinical development are significantly longer if clinical development commences before this threshold is reached.[6-9, 22, 59] Our work further suggests that innovation can stall when private capital is invested in immature technologies and is, therefore, dissuaded from subsequent investments in the space as it matures. **Given the high risk, limited availability, high cost of investment capital for early-stage development, greater focus on characterizing the readiness of basic science for development could radically reshape the landscape of innovation and promote more rapid translation of the most promising technologies.**

Foundational Scientific Exploration Our research suggests that the primary contribution of NIH funding to pharmaceutical innovation is its support of basic science. **Basic research conducted with federal funding has an essential role in the innovation ecosystem as the primary driver of discovery and development and a source of economic stimulus.**

Basic science is defined as *"…experimental or theoretical work undertaken primarily to acquire knowledge….without any particular application or use in view*" [60], recognizing that NIH spending is inherently "*use-inspired*" [61]. Our work further suggests that virtually all of this work is performed in public institutions (as opposed to industry) and as noted above, the maturation of this body of research may be a critical determinant of product development success, timelines, and costs. [6-9, 22, 59] Initiating applied, translational, or development work based on an immature body of basic research increases the risk, timelines, and costs to industry.

We would argue that initiatives focused on specific diseases or conditions can only succeed if there is a commitment to establishing a mature body of basic research before considering translation or development. We have studied innovation following passage of the National Cancer Act of 1972. This Act was long considered to have been a failure due to the absence of new therapies or substantial improvement in outcomes for decades after the Act was passed. Our analysis suggest that the Act was, in fact, highly successful. The War on Cancer was launched at the outset of an era that saw the emergence of molecular biology, genomics, and biotechnology and the Act was instrumental in funding much of this foundational work. Our analysis shows that it was only when the foundational, basic research in these areas reached a threshold of maturation that new product began to appear. The last 20 years have seen an explosion of new therapies for cancer and dramatic advances in both treatments and outcomes. [22] Similarly, many early failures of candidate therapies for Alzheimer's involved candidate products that entered development before research on Amyloid was mature while the recent successes uniformly entered development only after that body of research was established. [62]

We would also call attention to the economic efficiencies created by public support for basic science. As described below, the timeline between basic research on drug targets and first approval of a drug associated with that target can be as much as 25 years. [21, 63, 64] This long timeline, coupled with the high cost of capital in industry (10.5%) makes investment in basic science untenable for industry. It is possible to estimate the cost savings provided to industry by NIH funding for basic science that is then transferred to industry through the public domain, workforce training or licensing. Industry typically calculates drug development costs as the sum of spending on the approved drug, an estimate of spending on failed candidate products, and a 10.5% cost of capital. Applying the same calculation to assess the value provided to industry by public funding of basic research, we estimate that the average cost saving to industry is as much as $2.9 billion, an amount roughly equivalent to reported development costs. [20, 21, 65]

Public funding for basic science also creates economic efficiencies through spillovers. Studies suggest that an average of three novel therapeutics are developed for each biological target. [19, 66] This is evident, for example, by the observation that most NIH-funded projects related to at least one drug approval are typically associated with multiple approved drugs. This means that the public investment in basic science is ultimately distributed across multiple approvals, reducing the effective public investment/drug and multiplying the returns from product development.

We believe that a **continued government commitment to funding basic science is essential for the future success of pharmaceutical innovation and commercialization. Initiatives that draw resources away from basic science toward applied or translational research in any specific therapeutic or disease area, no matter how well intentioned, represent a threat to future progress.**

Public Impact. NIH seeks approaches for assessing and recognizing the broader public impact of biomedical research, including contributions to health outcomes, economic growth, clinical practice, public trust, accessibility, and societal benefit.

Supporting research

Characterizing the NIH contribution to drug approvals

Recent criticisms of the National Institutes of Health (NIH) argue that NIH spending is overly focused on basic rather than applied or translational research and structured around investigator-initiated research rather than mission-driven strategies [67-69] to "enhance health, lengthen life, and reduce illness and disability" [37]. Ideological concerns aside, these arguments build on longstanding questions about the effectiveness of NIH's reliance on investigator-initiated research and an uncertain alignment between NIH research spending and the US burden of disease [41-46, 70-72]. **While we endorse a process of continuous assessment and evidence-based improvement of the scientific enterprise [73, 74], we believe recent criticisms of the NIH are unfounded.**

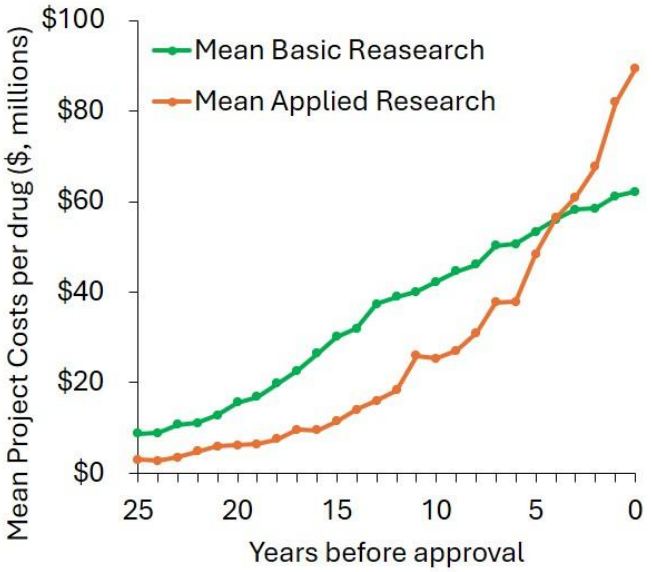
In a series of studies,[vi] we have assessed the NIH contribution to new drug approvals from 2010-2024 [9, 18, 19, 38, 75, 76]. While the biopharmaceutical industry is primarily responsible for discovery, development, and commercialization of novel medicines, we found that the NIH funded basic or applied research related to 99.2% (614/619) of the drugs approved from 2010-2024 before their first FDA approval.[vii] This research involved 386 thousand project years of funding with NIH costs of $254.8 billion including $208.0 billion (81.6%) for basic research and $46.8 billion (18.4%) for applied research. [38] In other studies, we have estimated that ~3-4% of this NIH funding contributes to clinical development [76] and that <1% leads to patents that provide market exclusivity. [75] **Overall, the average cost of NIH funded research associated with each approval has been estimated to be ~$1.5 billion/drug approval, an amount comparable to reported industry spending**. [19-21] **Calculated with a social discount rate of 3% (corresponding to the average cost of borrowing) [73, 77] this represent a public investment cost of $1.7 billion/drug approval.**

The role of NIH funded basic research

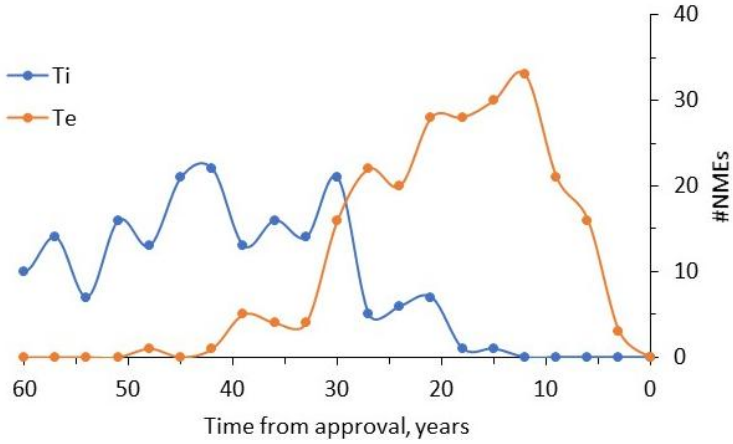
Federally funded basic research plays an essential role in pharmaceutical innovation. Our analyses demonstrate that few if any novel medicines are successfully developed until there is a mature body of basic research on the drug's target. [7-9, 22] Basic research is defined by as *"…experimental or theoretical work undertaken primarily to acquire knowledge….without any particular application or use in view*" [60], recognizing that NIH spending is inherently "*use-inspired*" [61] by the mission of improving health. Importantly, this includes not only salient research milestones, but also studies necessary to replicate, refute, or refine earlier research and establish the body of reliable evidence required for drug development and regulatory approval [7, 8, 18, 23]. In our studies, we identify publications describing basic research by searching PubMed for both the newly approved drug and its biological target. Publications (PMID) and selecting those with NIH funding identified in NIH RePORTER. PMID and the projects that provided funding are categorized as applied research if they were identified in a search for the drug. PMIDs and projects are categorized as applied if they are identified only in a search for the drug target.

Our studies also enabled an analysis of the timeline between public funding of basic and applied science and new drug approvals. These studies suggest that NIH funding supporting basic science on the drug target may support much as 25 years before first approval of a drug associated with that target. [38] This pattern is consistent with the observations of others. [7-9, 78-81]. In contrast we observed a sharp increase in applied science on specific drugs 5-10 years prior to first approval, consistent with timelines of preclinical and clinical development. [21, 63, 64] **The recognized lags between federal funding for basic**

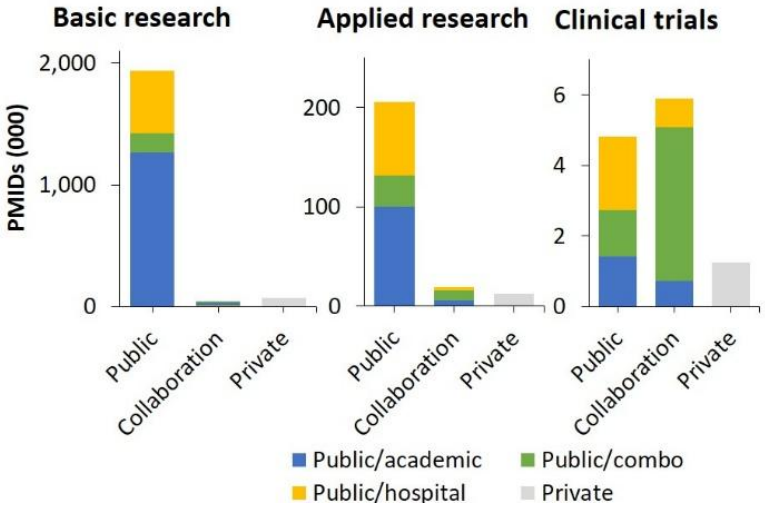
or applied research and drug approvals need to be incorporated in any understanding of the timelines and economics of pharmaceutical innovation



We analyze the maturation of basic science using the analytical TIME model,[6, 7] which models the accumulation of publications as an S-curve analogous to the functions used to model the advance of technologies in other fields (engineering). [6, 7, 9] We then calculate a numerically defined point of "initiation" (Ti) and an "established point" (Te) as the points of maximum acceleration and deceleration respectively. The adjacent figure shows the timeline between the Ti and Te points for 356 drugs approved 2010-2019. For most drugs, basic research on the drug target was initiated >25 years before drug approval and that this body of research had passed this established threshold, indicative of a mature body of basic science, before each successful drug approval. [9]



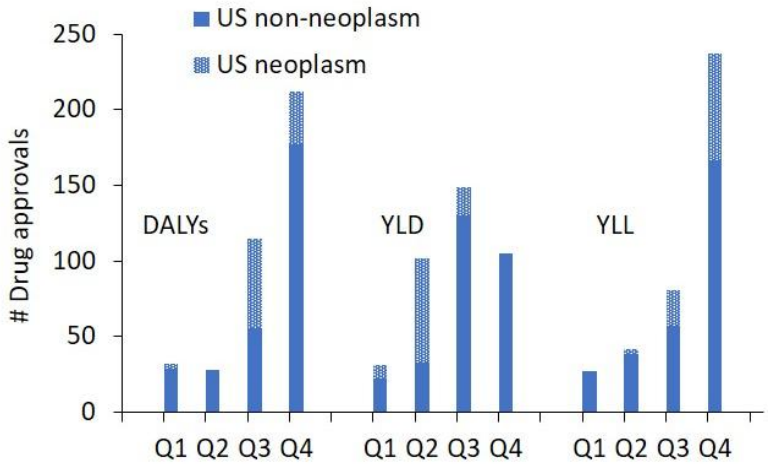
The large majority of basic research associated with drug approvals is conducted in academic centers with little involvement from industry. These studies identify the institutional affiliations of authors of publications describing basic or applied (including clinical) research related to 387 drugs approved from 2010-2019. Affiliations were designated as public (academic, hospital, or both) or private (for profit corporation). Publications could involve entirely public authors, private authors, or be a collaboration with both public and private authorship. These data show that virtually all basic research was published by academic institutions or public hospitals, a small increase in collaborative applied research, and broad collaboration in clinical development. [82] Recognizing that industry may be less likely to publish basic research, these data provide no evidence to support any significant industry contribution to the advance of basic science.



The effectiveness of investigator-initiated funding mechanisms

A large majority of the NIH funding related to drug approvals was awarded through investigator-initiated, rather than mission-driven, funding mechanisms. Excluding intramural research, 86.4% of project years and 77.5% of NIH costs were associated with funding opportunities that typically solicit investigator-initiated research proposals. This included a preponderance of funding awarded through investigator-initiated R (research) grants or P (Program Project and Center) grants. Mission-driven mechanisms including U (Collaborative Program) grants and institutional P grants provided the majority of funding for clinical trials, but only a small fraction of the total NIH funding associated with drug approvals. This analysis further showed that **investigator-initiated mechanisms were significantly more likely to support research projects that were associated with drug approvals than program-driven mechanisms.**

While program-driven funding mechanisms may be preferred for directing resources toward specified, short-term public needs, investigator-initiated, scientifically driven mechanisms are preferred for advancing long term innovation. These observations are consistent with theoretical arguments that central, strategic planning in science, as in a market economy, is inherently less efficient. Polanyi, for example, argues that neither scientists nor policymakers can reliably predict which lines of inquiry will produce commercially valuable discoveries. In contrast. Polanyi argues that scientific progress depends on the decentralized decisions of scientists who "freely [make] their own choice of problems." [83]. He argues that decentralized systems allow researchers to investigate problems that appear soluble and scientifically important, pursuing varied avenues of investigation, and continuously apply newly emerging knowledge to improve their direction or pursue opportunities with greater promise. **We call on the NIH to resist initiatives to reform the process of investigator-driven research at the NIH or to promote program driven initiatives over scientifically driven investigations.**



One criticism of investigator-driven mechanisms for research funding has been that there is an uncertain association between NIH funding and the US disease burden [41-46]. Our analysis of the NIH funding associated with drug approvals shows wide, highly asymmetrical variation in the level of NIH spending associated with individual drug approvals. Thus, while t**he number new drug approvals has been shown to be associated with the US burden of disease [47, 48]** this association may not be reflected in an association between NIH spending and disease burden. This suggests that meaningful assessment of the alignment between NIH funding and health priorities must be based on specific outputs from NIH funded research rather than simply levels of NIH spending. **These findings suggest that considerable caution must be exercised in assessing the relationship between NIH funding and the burden of disease in the United States and that any meaningful assessment should be based on health-related outputs rather than the number of grants or amount of funding awarded.**

Effectiveness of the NIH in directing research funding to productive outputs

Our studies have also estimated the fraction of all NIH research spending that is productively employed forward drug approvals.[viii] These studies suggest that 59.7% (range 53.7 to 63.6%) of reported NIH spending for basic research from 2010-2024, 20.5% (range 18.5% to 22.6%) of reported NIH spending for applied research, and 40.9% (range: 36.8% to 43.9%) of all NIH research spending was productively employed toward drug approvals.

Extensive evidence demonstrates that new drug approvals are significantly associated with increased life expectancy and reduced mortality, disability, and functional limitations [49-56]. In this context the observation that a substantial fraction of NIH funding is being productively employed towards outputs that directly contribute to the mission of advancing healthcare and health outcomes. We notes that these totals describe only NIH funding for research prior to first FDA approval and does not include NIH spending for research related to product line extensions, follow-on indications, comparative- or cost-effectiveness studies, or evidence-based guidelines that could influence the impact of these drugs on health. These totals also may not fully capture NIH contributions to candidate products that failed in clinical trials or were abandoned by industry for commercial reasons [21] and do not addres other outputs from NIH funded research that would be expected to have positive impacts on health such as novel devices, diagnostics, materials, instrumentation, or improvements in surgery, nursing, patient safety, nutrition, health equity, wellness management, and healthcare delivery. 78]

In sum, the evidence suggests that the NIH is highly effective in directing research funding to projects with the potential for positively impact the healthcare and health of Americans. These analyses do not support the need for a broad overhaul of NIH's operational model.

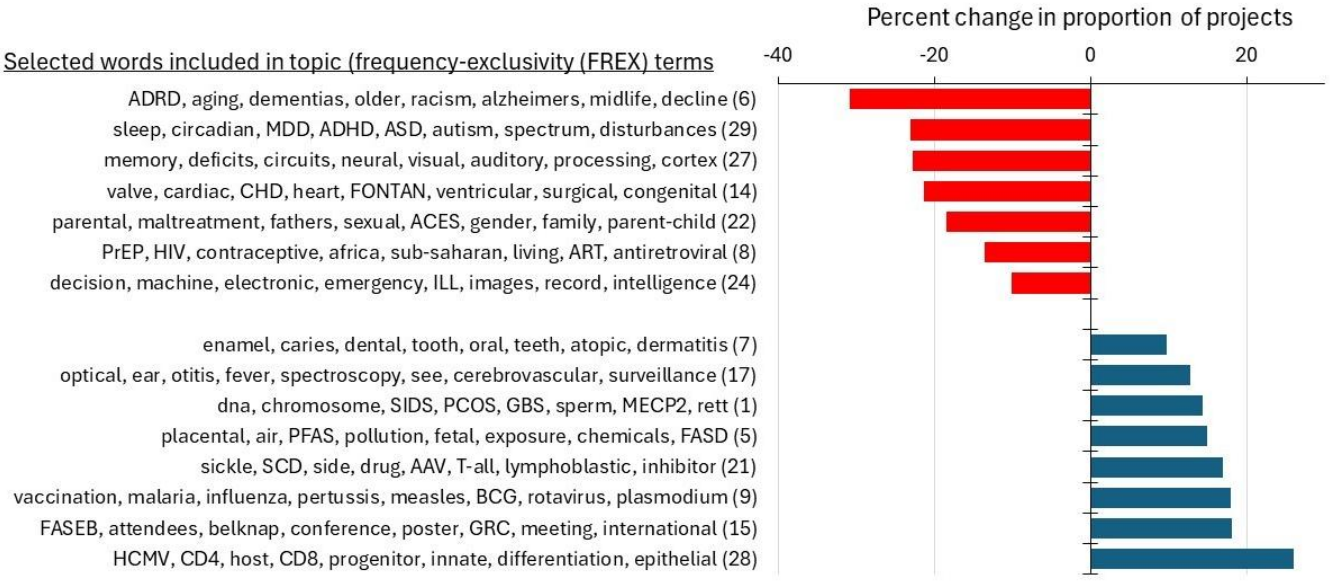
There is extensive evidence that federal funding for biomedical research stimulates private investment in innovation. Several criticisms of the Economic studies demonstrate that the stock of public basic research is significantly associated with increases in drug approvals [79] and that growth of disease specific research is significantly associated with the number of drugs entering clinical trials [84] and the number of private sector patents [85]. NIH funding has also been shown to stimulate private investment, with each dollar of public basic biomedical research stimulating $8.38 private spending after eight years and each dollar of public clinical research stimulating $2.35 private spending after three years. [81] There is little evidence for the assertion that federal funding for biomedical research "crowds out" private investment.

Sensitivity of NIH research portfolio to censoring

In a preliminary study we have examined NIH funding for pediatric research projects from 2020-2026 identified in NIH RePORTER. [40] As part of this analysis we performed statistical topic modeling that identifies latent topics (i.e., groups of words that co-occur) across the Public Health Relevance narratives of !15,000 funded research projects involving children or newborns and examined the share of each narrative devoted to each topic (word group). Through this analysis, the topical composition of funded projects can be measured and compared across years without manual coding. We then looked at the frequency of the 30 topics using terms that were commonly found within each topic and relatively exclusive to each topic and compared the prevalence of these terms in projects funded FY2020-FY2024 and those funded FY2025– FY2026.

The results illustrate substantial reductions in topics associated with terms that appeared in the 2025 NIH guidance on awards "…not aligned with the NIH's priorities."[39] For example, "racism" appeared among

FREX terms within a topic also characterized by the terms “Alzheimer’s” and “aging.” Similarly, “gender” appeared within a topic also characterized by “fathers” and “parent-child.” These findings show the broader scientific context in which words subject to the 2025 NIH guidance occur (i.e., flagged terms sit within topics that also contain vocabulary not governed by this guidance. This suggests that administrative screening based on individual terms could cause collateral damage to areas of research that continue to have high priority. The NIH needs to be attentive to collateral damage that can arise from any efforts to selectively restrict funding on specific topics based on simple text analysis.



References

1. Dennery, P., et al., *Strategies To Enhance NIH-Funded Pediatric Research.*
2. Sampat, B.N. and F.R. Lichtenberg, *What are the respective roles of the public and private sectors in pharmaceutical innovation?* Health Affairs, 2011. **30**(2): p. 332-339.
3. Chakravarthy, R., et al., *Public-and private-sector contributions to the research and development of the most transformational drugs in the past 25 years: from theory to therapy.* Therapeutic Innovation & Regulatory Science, 2016. **50**(6): p. 759-768.
4. Reichert, J.M. and C.-P. Milne, *Public and private sector contributions to the discovery and development of "impact" drugs.* American journal of therapeutics, 2002. **9**(6): p. 543-555.
5. Flier, J.S., *Academia and industry: allocating credit for discovery and development of new therapies.* The Journal of clinical investigation, 2019. **129**(6): p. 2172-2174.
6. McNamee, L.M. and F.D. Ledley, *Patterns of technological innovation in biotech.* Nature biotechnology, 2012. **30**(10): p. 937-943.
7. McNamee, L.M., M.J. Walsh, and F.D. Ledley, *Timelines of translational science: from technology initiation to FDA approval.* PLoS ONE, 2017. **12**(5): p. e0177371.
8. Beierlein, J.M., et al., *Landscape of innovation for cardiovascular pharmaceuticals: from basic science to new molecular entities.* Clinical therapeutics, 2017. **39**(7): p. 1409-1425. e20.
9. Cleary, E., M.J. Jackson, and F. Ledley, *Government as the First Investor in Biopharmaceutical Innovation: Evidence From New Drug Approvals 2010–2019.* Institute for New Economic Thinking Working Paper Series, 2020(133).
10. Stevens, A.J., et al., *The role of public-sector research in the discovery of drugs and vaccines.* New England Journal of Medicine, 2011. **364**(6): p. 535-541.
11. Kaitin, K.I., N.R. Bryant, and L. Lasagna, *The Role of the Research-Based Pharmaceutical Industry in Medical Progress in the United States.* The Journal of Clinical Pharmacology, 1993. **33**(5): p. 412-417.
12. DiMasi, J.A., R.W. Hansen, and H.G. Grabowski, *The price of innovation: new estimates of drug development costs.* Journal of Health Economics, 2003. **22**(2): p. 151-185.
13. Ledley, F., Galkina-Cleary, E., Zhang, K. , *NIH research funding for patents that provide market exclusivity for new drugs 2010–2019.* submitted.
14. Nayak, R.K., et al., *Public-sector contributions to novel biologic drugs.* JAMA Internal Medicine, 2021. **181**(11): p. 1522-1525.
15. Patridge, E.V., et al., *An analysis of original research contributions toward FDA-approved drugs.* Drug Discovery Today, 2015. **20**(10): p. 1182-1187.
16. Nayak, R.K., J. Avorn, and A.S. Kesselheim, *Public sector financial support for late stage discovery of new drugs in the United States: cohort study.* bmj, 2019. **367**.
17. Unger, J.M., et al., *Population, clinical, and scientific impact of National Cancer Institute's National Clinical Trials Network treatment studies.* Journal of Clinical Oncology, 2023. **41**(11): p. 2020-2028.
18. Cleary, E.G., et al., *Contribution of NIH funding to new drug approvals 2010–2016.* Proceedings of the National Academy of Sciences, 2018. **115**(10): p. 2329-2334.
19. Cleary, E., et al., *Comparison of Research Spending on New Drug Approvals by the National Institutes of Health vs the Pharmaceutical Industry, 2010-2019.* JAMA Health Forum, 2023. **4**(4): p. e230511-e230511.
20. Wouters, O.J., M. McKee, and J. Luyten, *Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018.* JAMA, 2020. **323**(9): p. 844-853.
21. DiMasi, J.A., H.G. Grabowski, and R.W. Hansen, *Innovation in the pharmaceutical industry: new estimates of R&D costs.* Journal of Health Economics, 2016. **47**: p. 20-33.
22. McNamee, L.M. and F.D. Ledley, *Modeling timelines for translational science in cancer; the impact of technological maturation.* PLoS ONE, 2017. **12**(3): p. e0174538.
23. NAS, *Reproducibility and replicability in science*, ed. N.A.o. Sciences. 2019: National Academies of

Sciences.
24. Ledley, F.D., et al., *Profitability of large pharmaceutical companies compared with other large public companies.* Jama, 2020. **323**(9): p. 834-843.
25. Cleary, E.G., et al., *Comparing long-term value creation after biotech and non-biotech IPOs, 1997–2016.* Plos one, 2021. **16**(1): p. e0243813.
26. Twomey, J., et al., *Contribution of revenue from drugs subject to price negotiation under the Inflation Reduction Act to the revenue, profit, and returns of pharmaceutical manufacturers.* Drug Discovery Today, 2025: p. 104585.
27. IOM, *Valuing health for regulatory cost-effectiveness analysis: Edited by Wilhelmine Miller, Lisa A. Robinson and Robert S. Lawrence.* . 2007, National Academy of Medicine.
28. OMB, *Circular A-4, Regulatory Analysis*. 2023: Office of Management and Budget.
29. EPA, *Guidelines for Preparing Economic Analyses, 3rd Edition*. 2024, Environmental Protection Agency.
30. Hutubessy, R., D. Chisholm, and T.T. Edejer, *Generalized cost-effectiveness analysis for national-level priority-setting in the health sector.* Cost Eff Resour Alloc, 2003. **1**(1): p. 8.
31. Zhou, E.W., et al., *Considering Returns on Federal Investment in the Negotiated'Maximum Fair Price'of Drugs Under the Inflation Reduction Act: an Analysis.* Institute for New Economic Thinking Working Paper Series, 2024(219).
32. Chaves da Silva, P.G. and F. Ledley, *Forget the Posturing – The Inflation Reduction Act May Work Better Than Many Expected*. 2024, Institute for New Economic Thinking Institute for New Economic Thinking.
33. Chaves da Silva, P., R.M. Conti, and F.D. Ledley, *Social and private value created through commercialization of Gilead Sciences' innovative medicines for Hepatitis C and HIV: a cross-sectional analysis.* medRxiv, 2025: p. 2025.06. 11.25329435.
34. Garrison Jr, L.P., et al., *Estimating the allocation of the economic value generated by utilization of all-oral direct-acting antivirals for hepatitis C in the United States, 2015 to 2019.* Value in Health, 2024. **27**(8): p. 1021-1029.
35. Philipson, T.J. and A.B. Jena. *Who benefits from new medical technologies? Estimates of consumer and producer surpluses for HIV/AIDS drugs*. in *Forum for Health Economics & Policy*. 2006. De Gruyter.
36. Lakdawalla, D.N., et al., *An economic evaluation of the war on cancer.* Journal of health economics, 2010. **29**(3): p. 333-346.
37. NIH. *Mission and Goals*. 2026; Available from: https://www.nih.gov/about-nih/mission-goals.
38. Zhou, E.W. and F.D. Ledley, *Fraction of National Institutes of Health research spending 2010-2024 productively employed toward new drug approvals.* (submitted).
39. NIH, *Staff Guidance - Reviewing Grants for Priority Alignment, Issue Date: December 20, 2025.* Posted by AAAS 2025.
40. Mozer, R. and F. ledley, *Trends in NIH funding for pediatric resaerch, FY2020-Fy2026.* submitted.
41. IOM, *Scientific opportunities and public needs: improving priority setting and public input at the National Institutes of Health*, C.o.t.N.R.P.-S. Process, Editor. 1998, Institute of Medicine: Institute of Medicine.
42. Gross, C.P., G.F. Anderson, and N.R. Powe, *The relation between funding by the National Institutes of Health and the burden of disease.* New England Journal of Medicine, 1999. **340**(24): p. 1881-1887.
43. Gillum, L.A., et al., *NIH disease funding levels and burden of disease.* PloS one, 2011. **6**(2): p. e16837.
44. Sampat, B.N., K. Buterbaugh, and M. Perl, *New evidence on the allocation of NIH funds across diseases.* The Milbank Quarterly, 2013. **91**(1): p. 163-185.
45. Carter, A.J. and M. Gevorkian, *The persistence of very low correlations between NIH research funding and disease burdens.* Public Health in Practice, 2025. **9**: p. 100580.
46. Gao, T., et al., *Allocation of the National Institutes of Health funding and burden of disease in 2008–2021 in the United States.* npj Health Systems, 2025. **2**(1): p. 41.
47. Ballreich, J.M., et al., *Allocation of National Institutes of Health funding by disease category in 2008 and 2019.* JAMA network open, 2021. **4**(1): p. e2034890-e2034890.

48. Jackson, M.J., G. Vaughan, and F.D. Ledley, *Association between expedited review designations and the US or global burden of disease for drugs approved by the US Food and Drug Administration, 2010–2019: a cross-sectional analysis.* BMJ open, 2024. **14**(3): p. e076542.
49. Lichtenberg, F.R., *The effect of pharmaceutical innovation on longevity: Evidence from the US and 26 high-income countries.* Economics & Human Biology, 2022. **46**: p. 101124.
50. Lichtenberg, F.R. *The Impact of Biopharmaceutical Innovation on Disability, Social Security Recipiency, and Use of Medical Care of US Community Residents, 1998–2015*. in *Forum for Health Economics and Policy*. 2022. De Gruyter.
51. Dubois, P. and M. Kyle, *The effects of pharmaceutical innovation on cancer mortality rates.* 2016.
52. Cutler, D.M. and M. McClellan, *Is technological change in medicine worth it?* Health affairs, 2001. **20**(5): p. 11-29.
53. Philipson, T. and A. Jena, *Who benefits from new medical technologies? Estimates of consumer and producer surpluses for HIV/AIDS drugs*. 2005, National Bureau of Economic Research Cambridge, Mass., USA.
54. Lichtenberg, F.R. *The effect of pharmaceutical innovation on longevity: patient level evidence from the 1996–2002 medical expenditure panel survey and linked mortality public-use files*. in *Forum for Health Economics and Policy*. 2013. De Gruyter.
55. Lichtenberg, F.R., *The impact of pharmaceutical innovation on disability days and the use of medical services in the United States, 1997–2010.* Journal of Human Capital, 2014. **8**(4): p. 432-480.
56. Lichtenberg, F.R., *The Impact of Biopharmaceutical Innovation on Disability, Social Security Recipiency, and Use of Medical Care of US Community Residents, 1998-2015.* 2021.
57. Boguslav, M.R., et al., *Creating an ignorance-base: Exploring known unknowns in the scientific literature.* Journal of biomedical informatics, 2023. **143**: p. 104405.
58. Derbyshire, J. and T. Aven, *Out of sight but still in mind: Developing an expectation for surprises by formalizing unknowledge in a contemporary risk-assessment framework.* Risk Analysis, 2025. **45**(6): p. 1199-1206.
59. Ledley, F., et al., *Why commercialization of gene therapy stalled; examining the life cycles of gene therapy technologies.* Gene therapy, 2014. **21**(2): p. 188-194.
60. NSF, *Definitions of Research and Development: An Annotated Compilation of Official Sources.* 2018.
61. Stokes, D.E., *Pasteur's quadrant: Basic science and technological innovation.* 2011: Brookings Institution Press.
62. Beierlein, J.M., et al., *Patterns of innovation in Alzheimer's disease drug development: a strategic assessment based on technological maturity.* Clinical Therapeutics, 2015. **37**(8): p. 1643-1651. e3.
63. Eder, J., R. Sedrani, and C. Wiesmann, *The discovery of first-in-class drugs: origins and evolution.* Nature Reviews Drug Discovery, 2014. **13**(8): p. 577-587.
64. Schuhmacher, A., O. Gassmann, and M. Hinder, *Changing R&D models in research-based pharmaceutical companies.* Journal of translational medicine, 2016. **14**(1): p. 105.
65. Rennane, S., L. Baker, and A. Mulcahy, *Estimating the Cost of Industry Investment in Drug Research and Development: A Review of Methods and Results.* INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 2021. **58**: p. 00469580211059731.
66. Santos, R., et al., *A comprehensive map of molecular drug targets.* Nature reviews Drug discovery, 2017. **16**(1): p. 19-34.
67. Early, J.F. and T. Kealey, *Mission Lost: How NIH Leaders Stole Its Promise to America*. 2025, CATO Institute.
68. Pitts, P.J., *It's time to rethink, recalibrate, and redistribute NIH Funding.* Journal of the Academy of Public Health, 2025.
69. Cassidy, B., *NIH in the 21st Century, ensuring transparency and American biomedical leadership*. 2024, US Senate Committee on Health, Education, Labor, & Pensions.
70. Azoulay, P., et al., *The Impact of Public R&D Investment on Private-Sector Patenting: Evidence from NIH Funding.* 2015.
71. Sampat, B.N. and P. Azoulay, *The impact of publicly funded biomedical and health research: a review.*

Department of Health Policy and Management, Columbia University, 2011.
72. Sampat, B.N., *Taking Stock: The Past and Future of Biomedical Research Funding.* NBER Chapters, 2025.
73. OMB, *Circular A-94 Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs*, O.o.M.a. Budget, Editor. 1992.
74. VanDuyse, S.A., M.J. Fulford, and M.G. Bartlett, *ICH Q10 Pharmaceutical Quality System Guidance: Understanding Its Impact on Pharmaceutical Quality: ICH Q10 Pharmaceutical Quality System Guidance.* The AAPS journal, 2021. **23**(6): p. 117.
75. Ledley, F. and E. Galkina-Cleary, *NIH funding for patents that contribute to market exclusivity of drugs approved 2010–2019 and the public interest protections of Bayh-Dole.* PLOS, 2023.
76. Zhou, E.W., M.J. Jackson, and F.D. Ledley. *Spending on Phased Clinical Development of Approved Drugs by the US National Institutes of Health Compared With Industry*. in *JAMA Health Forum*. 2023. American Medical Association.
77. OMB, *Circular A-4, Regulatory Review*, O.o.M.a. Budget, Editor. 2017.
78. Morris, Z.S., S. Wooding, and J. Grant, *The answer is 17 years, what is the question: understanding time lags in translational research.* Journal of the Royal Society of Medicine, 2011. **104**(12): p. 510-520.
79. Toole, A.A., *The impact of public basic research on industrial innovation: Evidence from the pharmaceutical industry.* Research Policy, 2012. **41**(1): p. 1-12.
80. Toole, A.A., *Does public scientific research complement industry R&D investment? The case of NIH supported basic and clinical research and pharmaceutical industry R&D*. 2005, ZEW Discussion Papers.
81. Toole, A.A., *Does public scientific research complement private investment in research and development in the pharmaceutical industry?* The Journal of Law and Economics, 2007. **50**(1): p. 81-104.
82. Jackson, *Public and private sector contributions and partnerships on publications related to drugs approved 2010-209* in preparation.
83. Polanyi, M., *The Republic of Science: Its Political and Economic Theory.* Knowing and Being, 1962.
84. Blume-Kohout, M.E., *Does targeted, disease-specific public research funding influence pharmaceutical innovation?* Journal of Policy Analysis and Management, 2012. **31**(3): p. 641-660.
85. Azoulay, P., et al., *Public R&D investments and private-sector patenting: evidence from NIH funding rules.* The Review of economic studies, 2019. **86**(1): p. 117-152.

Endnotes

[i] Dr. Ledley is a physician, scientist, entrepreneur, and researcher who is currently professor of Natural & Applied Science and Management at Bentley University and the founding director of the Center for Integration of Science and Industry (www.bentley.edu/sciindustry). He has an MD from Georgetown University, conducted graduate research as a part-time employee of the NIH (NIDDK), trained in pediatrics and genetics at the Boston Children's Hospital, and was an ACS post-doctoral fellow with Dr. David Baltimore at the Massachusetts Institute of Technology. He has served on the faculties of the Howard Hughes Medical Institute, Baylor College of Medicine, Texas Children's Hospital and Bentley University and was the founder of several biotechnology companies including GeneMedicine, Inc. (NASDAQ:GMED) and Variagenics, Inc (NASDAQ:VGEN). He has published >250 papers on topics ranging from medicine, human molecular genetics, and biotechnology to pharmaceutical development, corporate finance, bioethics, and public policy and is the inventor on 11 issued US patents. He has served on the Board of Directors of the Massachusetts Biotechnology Council and as an Advisor to Boston Children's Hospital and as a consultant to industry and government.

Dr. Ledley has received funding for research and training projects at the Baylor College of Medicine and Texas Children's Hospital and SBIR grants at GeneMedicine and Variagenics.

Dr. Ledley reports no conflicts of interest.

[ii] Note: data on pediatric outcomes is provided for illustrative purposes only and are not representative of the field of pediatrics and not intended to be research results. Data retrieved with ChatGPTPlus. All data were confirmed manually or through second sources. *accessed August 19. 2026*. Data are rounded to two significant figures from original source.

DATA SOURCES:

ROW	Historical reference	Contemporary reference
1	CDC/NCHS (1980)	Ely & Driscoll (2025)
2	National Center for Health Statistics; Vital statistics of the United States, 1975: Volume II—Mortality, Section 5, Life tables	National Center for Health Statistics; Vital statistics of the United States, 1975: Volume II—Mortality, Section 5, Life tables
3	Mallory (2022)	Mallory (2022)
4	National Cancer Institute (2024)	National Cancer Institute (2024)
5	Chen et al., (2019)	Chen et al., (2019)
6	McBennett, Davis, & Konstan (2022)	Cystic Fibrosis Foundation (2025)
7	Oster et al. (2013)	Oster et al. (2013)
8	Allen & Gibson (1961), citing Jervis	NICHD (2000)
9	Garnett et al (2021)	Garnett et al (2021)
10	Centers for Disease Control (1984a)	National Center for Health Statistics (2026a)
11	Engelhardt et al. (1980)	Mathis et al. (2025)
12	Meyer et al. (2000)	Marin et al. (2022)
13	Shay et al. (2001)	Hansen et al. (2022)
14	National Center for Health Statistics (1981)	National Center for Health Statistics (2026b)
15	National Cancer Institute, SEER Program (2020)	National Cancer Institute (2024)

SOURCES (Note: not included in numbered biography)

Allen, R. J., & Gibson, R. M. (1961). Phenylketonuria with normal intelligence. American Journal of Diseases of Children, 102(1), 115–122. Https://doi.org/10.1001/archpedi.1961.02080010117019; Centers for Disease Control and Prevention. (2025). Data and statistics: Congenital heart defects. https://www.cdc.gov/heart-defects/data/index.html; Centers for Disease Control and Prevention, National Center for Health Statistics. (1980). National infant mortality surveillance, 1980. https://www.cdc.gov/mmwr/preview/mmwrhtml/00001551.htm; Cystic Fibrosis Foundation. (2025). Patient Registry. https://www.cff.org/medical-professionals/patient-registry; Ely, D. M., & Driscoll, A. K. (2025). Infant mortality in the United States, 2023. National Vital Statistics Reports. https://www.cdc.gov/nchs/data/nvsr/nvsr74/nvsr74-07.pdf; Malloy, M. H. (2022). The limits of viability of extremely preterm infants. Baylor University Medical Center Proceedings, 35(5), 640–644. https://doi.org/10.1080/08998280.2022.2071073; McBennett, K. A., Davis, P. B., & Konstan, M. W. (2022). Increasing life expectancy in cystic fibrosis: Advances and challenges. Pediatric Pulmonology, 57(S1), S5–S12. https://doi.org/10.1002/ppul.25733; National Cancer Institute. (2024). Cancer in children and adolescents. https://www.cancer.gov/types/childhood-cancers/child-adolescent-cancers-fact-sheet; National Cancer Institute, Surveillance, Epidemiology, and End Results Program. (2026). Childhood brain and other nervous system cancer — Cancer Stat Facts. https://seer.cancer.gov/statfacts/html/childbrain.html; National Institute of Child Health and Human Development. (2000). Report of the NIH Consensus Development Conference on Phenylketonuria (PKU): Screening and management.

https://www.nichd.nih.gov/publications/pubs/pku/sub29; Weinreb, N. J., Deegan, P., Kacena, K. A., Mistry, P., Pastores, G. M., Velentgas, P., & vom Dahl, S. (2008). Life expectancy in Gaucher disease type 1. American Journal of Hematology, 83(12), 896–900. https://doi.org/10.1002/ajh.21305; Centers for Disease Control. (1984a). Premature mortality due to sudden infant death syndrome. Morbidity and Mortality Weekly Report. https://www.cdc.gov/mmwr/preview/mmwrhtml/00000704.htm
Centers for Disease Control. (1984b). Mumps—United States, 1980–1983. Morbidity and Mortality Weekly Report. https://www.cdc.gov/mmwr/preview/mmwrhtml/00000162.htm; Centers for Disease Control. (1984c). Rubella and congenital rubella syndrome surveillance, United States. CDC Surveillance Summaries, 33(4SS). https://stacks.cdc.gov/view/cdc/248509; Centers for Disease Control and Prevention. (2024). Chapter 9: Mumps. Manual for the Surveillance of Vaccine-Preventable Diseases. https://www.cdc.gov/surv-manual/php/table-of-contents/chapter-9-mumps.html; Centers for Disease Control and Prevention. (2025). Impact of U.S. MMR vaccination program: Rubella. https://www.cdc.gov/rubella/vaccine-impact/index.html; Engelhardt, S. J., Halsey, N. A., Eddins, D. L., & Hinman, A. R. (1980). Measles mortality in the United States 1971–1975. American Journal of Public Health, 70(11), 1166–1169. https://doi.org/10.2105/AJPH.70.11.1166; Hansen, C. L., Chaves, S. S., Demont, C., Viboud, C., & Mortensen, L. H. (2022). Mortality associated with influenza and respiratory syncytial virus in the US, 1999–2018. JAMA Network Open, 5(2), e220527. https://doi.org/10.1001/jamanetworkopen.2022.0527; Hoyert, D. L. (2026). Maternal mortality rates in the United States, 2024. National Center for Health Statistics. https://www.cdc.gov/nchs/data/hestat/hestat113.htm; Marin, M., Leung, J., Anderson, T. C., Lopez, A. S., & Schmid, D. S. (2022). Decline in severe varicella disease during the United States varicella vaccination program: Hospitalizations and deaths, 1990–2019. Journal of Infectious Diseases. https://stacks.cdc.gov/view/cdc/131793; Mathis, A. D., et al. (2025). Measles update—United States, January 1–April 17, 2025. Morbidity and Mortality Weekly Report, 74. https://www.cdc.gov/mmwr/volumes/74/wr/mm7414a1.htm; Meyer, P. A., Seward, J. F., Jumaan, A. O., & Wharton, M. (2000). Varicella mortality: Trends before vaccine licensure in the United States, 1970–1994. Journal of Infectious Diseases, 182(2), 383–390. https://doi.org/10.1086/315714; National Cancer Institute, Surveillance, Epidemiology, and End Results Program. (2020). SEER Cancer Statistics Review, 1975–2017, Table 28.3: Age-adjusted cancer U.S. death rates, childhood cancer. https://seer.cancer.gov/archive/csr/1975_2017/browse_csr.php?pageSEL=sect_28_table.03&sectionSEL=28; National Center for Health Statistics. (1981). Annual summary of births, deaths, marriages, and divorces: United States, 1980. Monthly Vital Statistics Report, 29(13). https://www.cdc.gov/nchs/data/mvsr/supp/mv29_13.pdf; National Center for Health Statistics. (1986). Health, United States, 1986 and Prevention Profile. U.S. Department of Health and Human Services. https://www.cdc.gov/nchs/data/hus/hus86acc.pdf; National Center for Health Statistics. (2026a). Mortality in the United States, 2024. https://www.cdc.gov/nchs/products/databriefs/db548.htm; National Center for Health Statistics. (2026b). FastStats: Influenza. https://www.cdc.gov/nchs/fastats/flu.htm; National Center for Health Statistics. (2026c). FastStats: Accidents or unintentional injuries. https://www.cdc.gov/nchs/fastats/accidental-injury.htm; National Highway Traffic Safety Administration. (2024). Traffic Safety Facts 2022 Data: Children (DOT HS 813 575). U.S. Department of Transportation. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813575; Shay, D. K., Holman, R. C., Roosevelt, G. E., Clarke, M. J., & Anderson, L. J. (2001). Bronchiolitis-associated mortality and estimates of respiratory syncytial virus-associated deaths among US children, 1979–1997. Journal of Infectious Diseases, 183(1), 16–22. https://doi.org/10.1086/317655; Oster, M. E., Lee, K. A., Honein, M. A., Riehle-Colarusso, T., Shin, M., & Correa, A. (2013). Temporal trends in survival among infants with critical congenital heart defects. Pediatrics, 131(5), e1502–e1508. https://doi.org/10.1542/peds.2012-3435; U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation. (2000). Trends in the well-being of America's children and youth, 2000. https://aspe.hhs.gov/reports/trends-well-being-americas-children-youth-2000-0; Chen, X., Pan, J., Wang, S., Hong, S., Hong, S., & He, S. (2019). The epidemiological trend of acute myeloid leukemia in childhood: A population-based analysis. Journal of Cancer, 10(20), 4824–4835. https://doi.org/10.7150/jca.32326.; Malloy, M. H. (2022). The limits of viability of extremely preterm infants. Baylor University Medical Center Proceedings, 35(5), 640–644. https://doi.org/10.1080/08998280.2022.2071073.

[iii] There is an extensive body of research characterizing the private and public sector contributions to pharmaceutical innovation 2. Sampat, B.N. and F.R. Lichtenberg, *What are the respective roles of the public and private sectors in pharmaceutical innovation?* Health Affairs, 2011. **30**(2): p. 332-339, 3. Chakravarthy, R., et al., *Public-and private-sector contributions to the research and development of the most transformational drugs in the past 25 years: from theory to therapy.* Therapeutic Innovation & Regulatory Science, 2016. **50**(6): p. 759-768, 4. Reichert, J.M. and C.-P. Milne, *Public and private sector contributions to the discovery and development of "impact" drugs.* American journal of therapeutics, 2002. **9**(6): p. 543-555, 5. Flier, J.S., *Academia and industry: allocating credit for discovery and development of new therapies.* The Journal of clinical investigation, 2019. **129**(6): p. 2172-2174. NIH funding for basic research contributes to establishing a mature body of basic research on drug targets, which is requisite for successful drug development 6. McNamee, L.M. and F.D. Ledley, *Patterns of technological innovation in biotech.* Nature biotechnology, 2012. **30**(10): p. 937-943, 7. McNamee, L.M., M.J. Walsh, and F.D. Ledley, *Timelines of translational science: from technology initiation to FDA approval.* PLoS ONE, 2017. **12**(5): p. e0177371, 8. Beierlein, J.M., et al., *Landscape of innovation for cardiovascular pharmaceuticals: from basic science to new molecular entities.* Clinical therapeutics, 2017. **39**(7): p. 1409-1425. e20, 9. Cleary, E., M.J. Jackson, and F. Ledley, *Government as the First Investor in Biopharmaceutical Innovation: Evidence From New Drug Approvals 2010–2019.* Institute for New Economic Thinking Working Paper Series, 2020(133).. Furthermore, 7-10% of new drugs have been shown to originate in the public sector 2. Sampat, B.N. and F.R. Lichtenberg, *What are the respective roles of the public and private sectors in pharmaceutical innovation?* Health Affairs, 2011. **30**(2): p. 332-339, 10. Stevens, A.J., et al., *The role of public-sector research in the discovery of drugs and vaccines.* New England Journal of Medicine, 2011. **364**(6): p. 535-541, 11. Kaitin, K.I., N.R. Bryant, and L. Lasagna, *The Role of the Research-Based Pharmaceutical Industry in Medical Progress in the United States.* The Journal of Clinical Pharmacology, 1993. **33**(5): p. 412-417, 12. DiMasi, J.A., R.W. Hansen, and H.G.

Grabowski, *The price of innovation: new estimates of drug development costs.* Journal of Health Economics, 2003. **22**(2): p. 151-185, 13. Ledley, F., Galkina-Cleary, E., Zhang, K. , *NIH research funding for patents that provide market exclusivity for new drugs 2010–2019.* submitted., including as many as 19% of first-in-class drugs 14. Nayak, R.K., et al., *Public-sector contributions to novel biologic drugs.* JAMA Internal Medicine, 2021. **181**(11): p. 1522-1525.. NIH funding also contributes to many clinical trials 3. Chakravarthy, R., et al., *Public-and private-sector contributions to the research and development of the most transformational drugs in the past 25 years: from theory to therapy.* Therapeutic Innovation & Regulatory Science, 2016. **50**(6): p. 759-768, 15. Patridge, E.V., et al., *An analysis of original research contributions toward FDA-approved drugs.* Drug Discovery Today, 2015. **20**(10): p. 1182-1187, 16. Nayak, R.K., J. Avorn, and A.S. Kesselheim, *Public sector financial support for late stage discovery of new drugs in the United States: cohort study.* bmj, 2019. **367**, 16. Ibid. most prominently through the National Cancer Institute's National Clinical Trials Network 17. Unger, J.M., et al., *Population, clinical, and scientific impact of National Cancer Institute's National Clinical Trials Network treatment studies.* Journal of Clinical Oncology, 2023. **41**(11): p. 2020-2028.

[iv] NIH funding related to new drug approvals from 2010-2024 was identified using a modification of previously described methods 9. Cleary, E., M.J. Jackson, and F. Ledley, *Government as the First Investor in Biopharmaceutical Innovation: Evidence From New Drug Approvals 2010–2019.* Institute for New Economic Thinking Working Paper Series, 2020(133), 18. Cleary, E.G., et al., *Contribution of NIH funding to new drug approvals 2010–2016.* Proceedings of the National Academy of Sciences, 2018. **115**(10): p. 2329-2334, 19. Cleary, E., et al., *Comparison of Research Spending on New Drug Approvals by the National Institutes of Health vs the Pharmaceutical Industry, 2010-2019.* JAMA Health Forum, 2023. **4**(4): p. e230511-e230511.. Details are described in the supplement to Cleary et al, 2023 19. Ibid.. PMID related to the drug or its target are identified through PubMed search. Those identified through a drug search are categorized as applied research. Those identified through target search (but not drug search) are categorized as basic research. NIH funded projects providing funding for these PMID are identified in NIH Reporter and one year of project funding (project year) is assigned for each PMID-PROJECT pair corresponding to the: (i) PMID publication year; (ii) project end year for PMID published 1-4 years after project end: or (iii) year of FDA approval for PMID published 1-4 years after first approval. Project years and costs are described after eliminating duplicates arising from PMID associated with multiple projects or projects producing multiple PMID.

[v] Our research assessed the total health value created through Medicare Part D payment for the drugs subject to price negotiation in 2024 and 2025. This work uses methods for valuing health impacts of public policy recommended by the Institute of Medicine 27. IOM, *Valuing health for regulatory cost-effectiveness analysis: Edited by Wilhelmine Miller, Lisa A. Robinson and Robert S. Lawrence.* . 2007, National Academy of Medicine., Office of Management and Budget 28. OMB, *Circular A-4, Regulatory Analysis*. 2023: Office of Management and Budget. and others 29.EPA, *Guidelines for Preparing Economic Analyses, 3rd Edition*. 2024, Environmental Protection Agency, 30. Hutubessy, R., D. Chisholm, and T.T. Edejer, *Generalized cost-effectiveness analysis for national-level priority-setting in the health sector.* Cost Eff Resour Alloc, 2003. **1**(1): p. 8.. These methods determine the "Health Benefit" created measured in QALYs; the assessed value of that benefit ("Health Value") based on survey data of American's "Willingness to Pay (WTP)" for one QALY of health ("WTP/QALY"), and the Residual Health Value (net Health Value) calculated as the Total Health Value minus the (gross) Medicare Part D Costs.31. Zhou, E.W., et al., *Considering Returns on Federal Investment in the Negotiated'Maximum Fair Price'of Drugs Under the Inflation Reduction Act: an Analysis.* Institute for New Economic Thinking Working Paper Series, 2024(219), 32. Chaves da Silva, P.G. and F. Ledley, *Forget the Posturing – The Inflation Reduction Act May Work Better Than Many Expected*. 2024, Institute for New Economic Thinking Institute for New Economic Thinking, 33. Chaves da Silva, P., R.M. Conti, and F.D. Ledley, *Social and private value created through commercialization of Gilead Sciences' innovative medicines for Hepatitis C and HIV: a cross-sectional analysis.* medRxiv, 2025: p. 2025.06. 11.25329435. These methods apply conventional accounting methods for assessing value, though there are also meaningful analogies to cost-effectiveness analysis or pharmacoeconomic studies 34. Garrison Jr, L.P., et al., *Estimating the allocation of the economic value generated by utilization of all-oral direct-acting antivirals for hepatitis C in the United States, 2015 to 2019.* Value in Health, 2024. **27**(8): p. 1021-1029, 35. Philipson, T.J. and A.B. Jena. *Who benefits from new medical technologies? Estimates of consumer and producer surpluses for HIV/AIDS drugs*. in *Forum for Health Economics & Policy*. 2006. De Gruyter, 36. Lakdawalla, D.N., et al., *An economic evaluation of the war on cancer.* Journal of health economics, 2010. **29**(3): p. 333-346.

[vi] NIH funding related to new drug approvals from 2010-2024 was identified using a modification of previously described methods 9. Cleary, E., M.J. Jackson, and F. Ledley, *Government as the First Investor in Biopharmaceutical Innovation: Evidence From New Drug Approvals 2010–2019.* Institute for New Economic Thinking Working Paper Series, 2020(133), 18. Cleary, E.G., et al., *Contribution of NIH funding to new drug approvals 2010–2016.* Proceedings of the National Academy of Sciences, 2018. **115**(10): p. 2329-2334, 19. Cleary, E., et al., *Comparison of Research Spending on New Drug Approvals by the National Institutes of Health vs the Pharmaceutical Industry, 2010-2019.* JAMA Health Forum, 2023. **4**(4): p. e230511-e230511.. Details are described in the supplement to Cleary et al, 2023 19. Ibid.. PMID related to the drug or its target are identified through PubMed search. Those identified through a drug search are categorized as applied research. Those identified through target search (but not drug search) are categorized as basic research. NIH funded projects providing funding for these PMID are identified in NIH Reporter and one year of project funding (project year) is assigned for each PMID-PROJECT pair corresponding to the: (i) PMID publication year; (ii) project end year for PMID published 1-4 years after project end: or (iii) year of FDA approval for PMID published 1-4 years after first approval. Project years and costs are described after eliminating duplicates arising from PMID associated with multiple projects or projects producing multiple PMID.

[vii] There is an extensive body of research characterizing the private and public sector contributions to pharmaceutical innovation 2. Sampat, B.N. and F.R. Lichtenberg, *What are the respective roles of the public and private sectors in pharmaceutical innovation?*

Health Affairs, 2011. **30**(2): p. 332-339, 3. Chakravarthy, R., et al., *Public-and private-sector contributions to the research and development of the most transformational drugs in the past 25 years: from theory to therapy.* Therapeutic Innovation & Regulatory Science, 2016. **50**(6): p. 759-768, 4. Reichert, J.M. and C.-P. Milne, *Public and private sector contributions to the discovery and development of "impact" drugs.* American journal of therapeutics, 2002. **9**(6): p. 543-555, 5. Flier, J.S., *Academia and industry: allocating credit for discovery and development of new therapies.* The Journal of clinical investigation, 2019. **129**(6): p. 2172-2174. NIH funding for basic research contributes to establishing a mature body of basic research on drug targets, which is requisite for successful drug development 6. McNamee, L.M. and F.D. Ledley, *Patterns of technological innovation in biotech.* Nature biotechnology, 2012. **30**(10): p. 937-943, 7. McNamee, L.M., M.J. Walsh, and F.D. Ledley, *Timelines of translational science: from technology initiation to FDA approval.* PLoS ONE, 2017. **12**(5): p. e0177371, 8. Beierlein, J.M., et al., *Landscape of innovation for cardiovascular pharmaceuticals: from basic science to new molecular entities.* Clinical therapeutics, 2017. **39**(7): p. 1409-1425. e20, 9.Cleary, E., M.J. Jackson, and F. Ledley, *Government as the First Investor in Biopharmaceutical Innovation: Evidence From New Drug Approvals 2010–2019.* Institute for New Economic Thinking Working Paper Series, 2020(133).. Furthermore, 7-10% of new drugs have been shown to originate in the public sector 2. Sampat, B.N. and F.R. Lichtenberg, *What are the respective roles of the public and private sectors in pharmaceutical innovation?* Health Affairs, 2011. **30**(2): p. 332-339, 10. Stevens, A.J., et al., *The role of public-sector research in the discovery of drugs and vaccines.* New England Journal of Medicine, 2011. **364**(6): p. 535-541, 11. Kaitin, K.I., N.R. Bryant, and L. Lasagna, *The Role of the Research-Based Pharmaceutical Industry in Medical Progress in the United States.* The Journal of Clinical Pharmacology, 1993. **33**(5): p. 412-417, 12. DiMasi, J.A., R.W. Hansen, and H.G. Grabowski, *The price of innovation: new estimates of drug development costs.* Journal of Health Economics, 2003. **22**(2): p. 151-185, 13.Ledley, F., Galkina-Cleary, E., Zhang, K. , *NIH research funding for patents that provide market exclusivity for new drugs 2010–2019.* submitted., including as many as 19% of first-in-class drugs 14. Nayak, R.K., et al., *Public-sector contributions to novel biologic drugs.* JAMA Internal Medicine, 2021. **181**(11): p. 1522-1525.. NIH funding also contributes to many clinical trials 3. Chakravarthy, R., et al., *Public-and private-sector contributions to the research and development of the most transformational drugs in the past 25 years: from theory to therapy.* Therapeutic Innovation & Regulatory Science, 2016. **50**(6): p. 759-768, 15. Patridge, E.V., et al., *An analysis of original research contributions toward FDA-approved drugs.* Drug Discovery Today, 2015. **20**(10): p. 1182-1187, 16. Nayak, R.K., J. Avorn, and A.S. Kesselheim, *Public sector financial support for late stage discovery of new drugs in the United States: cohort study.* bmj, 2019. **367**, 16. Ibid. most prominently through the National Cancer Institute's National Clinical Trials Network 17. Unger, J.M., et al., *Population, clinical, and scientific impact of National Cancer Institute's National Clinical Trials Network treatment studies.* Journal of Clinical Oncology, 2023. **41**(11): p. 2020-2028..

[viii] Total NIH spending from 2010-2024 contributing to drug approvals was estimated as the sum of spending related to new approvals over this period and inferred costs for a range of projected approvals after 2024. Spending related to future drug approvals was inferred from 2010-2024 (i) projected annual approvals and approvals associated with novel targets; (ii) observed time course of basic or applied research spending 25 years before approvals prior through 2024; (iii) average number of duplicated project years; (iv) mean, annual project costs with adjustment for COVID spending after 2020. Detailed methods are described in 38. Zhou, E.W. and F.D. Ledley, *Fraction of National Institutes of Health research spending 2010-2024 productively employed toward new drug approvals.* (submitted).

283. Silent Spring Institute

I am responding: On behalf of an organization

Name: Dr. Jennifer Liss Ohayon

Name of Organization: Silent Spring Institute

Type of Organization: Non-Profit Research Organization

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: While incentives and infrastructure are important for supporting the reproduction and replication of existing findings in certain key areas, this cannot detract from funding novel research and new techniques. One way to increase the level of confidence in existing studies is through implementing rigorous quality assurance/quality control (QA/QC) in all funded studies. Many studies do not adequately describe or interpret quality control (QC) data in publications, which decreases the ability of readers to evaluate data quality and appropriate uses of the data (such as valid comparison to other health studies or application to regulatory decision-making). Silent Spring Institute has published guidance on how to implement quality assurance/quality control (QA/QC) in environmental exposure measurement projects (including emphasizing the importance of incorporating QC samples in study design) and draw conclusions about data validity. NIH can strongly emphasize the importance of rigorous plans for data management, quality control, and quality assurance for all funded projects. These plans should pertain to all stages of the project, including collection, storage, shipment, and analysis of biological samples). Moreover, NIH can require adequate descriptions of QA/QC processes in publications resulting from funded research so that other researchers and government officials are able to assess the quality of the research.

2. Data, Software, and Model Sharing: We support efforts to share scientific data and tools. For example, as part of the Women Workers Biomonitoring Collaborative (WWBC), Silent Spring Institute collaborated with other scientists and impacted workers to develop the first women workers' biospecimen bank. This archive of biological samples is available to other researchers who are studying occupational exposures, including those that support the advancement of non-targeted biomonitoring approaches and those studying associations between exposures and upstream biomarkers of effect with potential relevance to cancer. NIH can fund research that invests in innovative sample archives that other studies can benefit from, including encouraging researchers to establish infrastructure as part of their funded projects. Moreover, NIH can allocate supplemental grants to maintaining infrastructure and catalyzing new collaborations based on existing sample archives or datasets.

Moreover, for funded NIH projects, researchers can be encouraged to publish scientific data and code in public repositories. Code that is archived in repositories such as GitHub, can be assigned DOIs allowing researchers to track its subsequent use, citation counts, and overall impact; these metrics can be shared with NIH. Initial data sharing plans can include trackable metrics for datasets.

3. Training and Mentorship: We propose the following indicators for assessing and quantifying the impact of training and mentorship:
• Number of one-on-one contact opportunities that a trainee has with their mentor.
• Mentor investment in the quality of trainee products, including acceptance of trainee authored articles in high-impact journals.
• Trainee evaluations of mentors.
• Tracking non-traditional outputs from trainees, including media coverage, community presentations, and engagement with K-12 schools.

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: We are encouraged by NIH efforts to issue guidance on aggregate-level report-back, and we call on NIH to promptly formalize requirements for both aggregate- and individual-level sharing of research results in all funded research. Research on report-back in environmental exposure assessment and biomonitoring research has indicated that participants, including those from different socioeconomic backgrounds, overwhelmingly want to know their personal exposure results if given the choice and report-back can motivate participants to consider both personal and collective strategies to reduce toxics in their environment. Moreover, reporting back to individuals can allow new research questions to emerge based on consulting with participants, for example on outlier results.

NIH can also require lay summaries of all research results and require researchers to include in grant reports any activities they undertook to reach larger audiences such as publishing op-eds and giving media interviews, translating research results in social
media posts, and giving presentations to non-academic audiences (e.g., water and wastewater professionals, clinicians, and state public health officials). As an example of ensuring research reaches broader audiences, Silent Spring Institute created the PFAS Exchange, an online resource center about PFAS contaminant in drinking water, to help impacted communities, water utilities, and clinicians understand exposures and take action to protect health (pfas-exchange.org). The website includes a resource hub with fact sheets about PFAS and medical screening guidance for both impacted communities and clinicians, a map of sites across the U.S. with known or suspected contamination, and an exposure tool to help individuals interpret their water or blood results. As other examples, Silent Spring collaborated with social media influencers to communicate research findings regarding endocrine disrupting chemicals to a broader public and created a consumer app, Detox Me, to share evidence-based tips for exposure reduction. We also created a digital tool, DERBI, to facilitate reporting back of individual research results. NIH can require researchers to report on these types of non-traditional outreach and products and also allocate more research funding to researchers to incorporate these types of activities into funded studies.
Finally, funding opportunities can encourage researchers to pick impactful questions that people care about by funding more researcher-community collaborations.
Thank you for your consideration of our comments. We would be happy to schedule a time to discuss any of them further.

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/Silent_Spring_NIH_Measure_Reward_Scientific_Research_to_upload.pdf

SILENT SPRING INSTITUTE
Researching the Environment and Women's Health

320 Nevada Street, Suite 302, Newton, MA 02460 / 617.332.4288 / www.silentspring.org

August 19, 20026

Re: Request for Information on Measuring and Rewarding Scientific Impact, NOT-OD-26-087

We thank you for the opportunity to comment on NIH's Request for Information on Measuring and Rewarding Scientific Impact. Silent Spring Institute is an independent research organization that investigates links between the environment and women's health. Silent Spring has defined the set of chemicals that are expected to increase breast cancer risk based on their activity in human and toxicological studies, and our Mammary Carcinogens Review Database was featured in Science as "The Breast Cancer List." Silent Spring continues to update the list of breast cancer-relevant chemicals to include those that alter development of the mammary gland, act as endocrine disruptors and promote tumor growth, or have epidemiological evidence. Silent Spring's research program also includes numerous biomonitoring studies to understand personal exposures to endocrine disrupting chemicals (EDCs) and carcinogens, as well as studies of chemical exposures from consumer products, indoor and outdoor air, and house dust. Our Household Exposure Study was described by Environmental Science & Technology as the "most comprehensive analysis" of exposures in homes. We are leaders in developing best methods and tools for reporting back individual-level research results and have contributed to over two dozen publications evaluating impacts of report-back.

Below we respond to the following topics that NIH is soliciting input on.

Rigor and Reproducibility.

While incentives and infrastructure are important for supporting the reproduction and replication of existing findings in certain key areas, this cannot detract from funding novel research and new techniques. One way to increase the level of confidence in existing studies is through implementing rigorous quality assurance/quality control (QA/QC) in all funded studies. Many studies do not adequately describe or interpret quality control (QC) data in publications, which decreases the ability of readers to evaluate data quality and appropriate uses of the data (such as valid comparison to other health studies or application to regulatory decision-making). Silent Spring Institute has published guidance on how to implement quality assurance/quality control (QA/QC) in environmental exposure measurement projects (including emphasizing the importance of incorporating QC samples in study design) and draw conclusions about data validity.[1] NIH can strongly emphasize the importance of rigorous plans for data management, quality control, and quality assurance for all funded projects. These plans should pertain to all stages of the project, including collection, storage, shipment, and analysis of biological samples). Moreover, NIH can require adequate descriptions of QA/QC processes in publications resulting from funded research so that other researchers and government officials are able to assess the quality of the research.

Data, Software, and Model Sharing

We support efforts to share scientific data and tools. For example, as part of the Women Workers Biomonitoring Collaborative (WWBC), Silent Spring Institute collaborated with other scientists and impacted workers to develop the first women workers' biospecimen bank. This archive of biological samples is available to other researchers who are studying occupational exposures, including those that support the advancement of non-targeted biomonitoring approaches and those studying associations between exposures and upstream biomarkers of effect with potential relevance to cancer.[2] NIH can fund research that invests in innovative sample archives that other studies can benefit from, including encouraging researchers to establish infrastructure as part of their funded projects. Moreover, NIH can allocate supplemental grants to maintaining infrastructure and catalyzing new collaborations based on existing sample archives or datasets.

Moreover, for funded NIH projects, researchers can be encouraged to publish scientific data and code in public repositories. Code that is archived in repositories such as GitHub, can be assigned DOIs allowing researchers to track its subsequent use, citation counts, and overall impact; these metrics can be shared with NIH. Initial data sharing plans can include trackable metrics for datasets.

Training and Mentorship.

We propose the following indicators for assessing and quantifying the impact of training and mentorship:

- Number of one-on-one contact opportunities that a trainee has with their mentor.
- Mentor investment in the quality of trainee products, including acceptance of trainee authored articles in high-impact journals.
- Trainee evaluations of mentors.
- Tracking non-traditional outputs from trainees, including media coverage, community presentations, and engagement with K-12 schools.

Public Impact

We are encouraged by NIH efforts to issue guidance on aggregate-level report-back, and we call on NIH to promptly formalize requirements for both aggregate- and individual-level sharing of research results in all funded research. Research on report-back in environmental exposure assessment and biomonitoring research has indicated that participants, including those from different socioeconomic backgrounds, overwhelmingly want to know their personal exposure results if given the choice and report-back can motivate participants to consider both personal and collective strategies to reduce toxics in their environment.[3-7] Moreover, reporting back to individuals can allow new research questions to emerge based on consulting with participants, for example on outlier results.

NIH can also require lay summaries of all research results and require researchers to include in grant reports any activities they undertook to reach larger audiences such as publishing op-eds and giving media interviews, translating research results in social

media posts, and giving presentations to non-academic audiences (e.g., water and wastewater professionals, clinicians, and state public health officials). As an example of ensuring research reaches broader audiences, Silent Spring Institute created the PFAS Exchange, an online resource center about PFAS contaminant in drinking water, to help impacted communities, water utilities, and clinicians understand exposures and take action to protect health (*pfas-exchange.org*). The website includes a resource hub with fact sheets about PFAS and medical screening guidance for both impacted communities and clinicians, a map of sites across the U.S. with known or suspected contamination, and an exposure tool to help individuals interpret their water or blood results. As other examples, Silent Spring collaborated with social media influencers to communicate research findings regarding endocrine disrupting chemicals to a broader public[8] and created a consumer app, Detox Me, to share evidence-based tips for exposure reduction. We also created a digital tool, DERBI, to facilitate reporting back of individual research results.[9] NIH can require researchers to report on these types of non-traditional outreach and products and also allocate more research funding to researchers to incorporate these types of activities into funded studies.

Finally, funding opportunities can encourage researchers to pick impactful questions that people care about by funding more researcher-community collaborations.

Thank you for your consideration of our comments. We would be happy to schedule a time to discuss any of them further.



Jennifer Liss Ohayon
Research Scientist

1. Udesky JO, Dodson RE, Perovich LJ, Rudel RA. Wrangling environmental exposure data: guidance for getting the best information from your laboratory measurements. *Environmental Health*. 2019;18(1):99.
2. Ohayon JL, Rasanayagam S, Rudel RA, et al. Translating community-based participatory research into broadscale sociopolitical change: insights from a coalition of women firefighters, scientists, and environmental health advocates. *Environmental Health*. 2023;22(1):60.
3. Perovich LJ, Ohayon JL, Cousins EM, et al. Reporting to parents on children's exposures to asthma triggers in low-income and public housing, an interview-based case study of ethics, environmental literacy, individual action, and public health benefits. *Environmental Health*. 2018;17(1):48.
4. Brody JG, Cirillo PM, Boronow KE, et al. Outcomes from returning individual versus only study-wide biomonitoring results in an environmental exposure study using

the Digital Exposure Report-Back Interface (DERBI). *Environmental Health Perspectives*. 2021;129(11):117005.
5. Adams C, Brown P, Morello-Frosch R, et al. Disentangling the exposure experience: the roles of community context and report-back of environmental exposure data. *Journal of health and social behavior*. 2011;52(2):180-196.
6. Brody JG, Dunagan SC, Morello-Frosch R, Brown P, Patton S, Rudel RA. Reporting individual results for biomonitoring and environmental exposures: lessons learned from environmental communication case studies. *Environmental Health*. 2014;13(1):40.
7. Ohayon JL, Cousins E, Brown P, Morello-Frosch R, Brody JG. Researcher and institutional review board perspectives on the benefits and challenges of reporting back biomonitoring and environmental exposure results. *Environmental Research*. 2017;153:140-149.
8. Franklin ET, Boronow KE, Liss Ohayon J, et al. The power of strategic social media influencer communication to improve Black women's knowledge and awareness of environmental endocrine-disrupting chemicals: surveys of Instagram users. *Journal of medical Internet research*. 2025;27:e66128.
9. Boronow KE, Susmann HP, Gajos KZ, et al. DERBI: a digital method to help researchers offer "right-to-know" personal exposure results. *Environmental health perspectives*. 2017;125(2):A27.

284. Mark H Histed

I am responding: On behalf of myself

Name: Mark H Histed

Name of Organization: N/A

Type of Organization: Not Applicable

Role: Investigator/Researcher

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: My comments apply to this category; they are listed in Other Notes below and attached as a PDF.

2. Data, Software, and Model Sharing: My comments apply to this category; they are listed in Other Notes below and attached as a PDF.

3. Training and Mentorship: My comments apply to this category; they are listed in Other Notes below and attached as a PDF.

4. Collaboration: My comments apply to this category; they are listed in Other Notes below and attached as a PDF.

5. Entrepreneurship and Translation: My comments apply to this category; they are listed in Other Notes below and attached as a PDF.

6. Foundational Scientific Exploration: My comments apply to this category; they are listed in Other Notes below and attached as a PDF.

7. Public Impact: My comments apply to this category; they are listed in Other Notes below and attached as a PDF.

8. Other Areas: Comment on RFI on Measuring And Rewarding Scientific Impact

Mark H. Histed, PhD, commenting in personal capacity and not on behalf of any organization

"NIH seeks public input on the defining features of modern 21st-century science that should be measured, recognized, and incentivized. "

Executive summary: Quantitative measurements of science, and other creative fields, inherently reduce innovation and limit big ideas. New metrics and measurement can in some cases add information, but NIH should strive to de-emphasize metrics in any evaluation or incentive structures in science. Instead of attempting to reduce the process of science to inherently limited data and measurement, NIH should rely on the gold-standard procedure of evaluating science that has been used to great success in the US for decades: peer review from the scientific community. This peer review should be applied to materials and research papers developed as a product of the creative and intellectual work of scientists.

Committees of scientists that read papers will do better at assessing the value of new ideas than any quantitative measure or metric, which will always be backward-looking..

*** The science ecosystem is a complex and nonlinear system impossible to characterize fully with data and measurement
Science is a highly complex social system which relies on frontier thinking from tens of thousands of scientists. Any empirical model or data will have difficulty assessing the quality of a complex system like science, because empirical models are backward-looking --- data themselves are observations of past events. Thus, quantitative evaluation will never be a complete way to capture the complex system that is science. NIH should therefore be incredibly cautious of bringing new metrics to bear to evaluate innovation.

To illustrate how empirical models alone are limited for understanding the world, without the accompanying development of theoretical models and a back-and-forth process that adjust models (or metrics) as new observations arrive, let me give an example. Neuroscience, my field, shows how data and quantitative measurements are useful — but are not a full description of a complex system. In the cerebral cortex, excitatory and inhibitory neural networks are tightly coupled to each other. This produces a phenomenon called the balanced state mode of network operation, which increases sensitivity to input, and decouples the activity of nearby neurons to maximize the ability to read out information carried by their responses. Cortical networks in the balanced state have coupled local excitation, and thus exhibit the inhibition-stabilized network (ISN) property. That is a property in which self-amplified activity that can boost certain input patterns is tracked and kept stable by local inhibition.

I give this background to note that ISN property would likely never have been found by pure experimental measurement or data collection. In 1987, Misha Tsodyks and a group of American and international collaborators (https://pubmed.ncbi.nlm.nih.gov/9151754/) showed that ISNs produce 'paradoxical responses,' where boosting inhibition unexpectedly causes excitatory cells to increase, not decrease, their activity. In a two-variable continuous-time model they demonstrated that linear perturbations in the model phase space produce unintuitive effects, curved paths in the phase space that change direction from the marginal or initial incremental direction. Many years later, focused experiments to test this nonlinear model by the Miller, Ferster, and my own group, amongst others, showed this property holds in the cortex.

The ISN model is merely a two-variable system and yet produces complex dynamics. And most empirical measurements --- applying regressions to large-scale data, for example --- measure only the initial marginal effect and missed these curved paths in the state space. Even adding nonlinear terms to a regression does not do a good job inferring curved/non-marginal dynamics, not least because more complex non-linearities become increasingly data-hungry to estimate. That is, most measurements that can be collected in finite time rely on a local linearization of a bigger complex system.

To quantitatively describe real-world complex systems, whether brain networks or the incredibly complex sociological system that is a science and innovation economy, we must not just develop metrics based on data. We must move beyond that to apply the simple models imposed by data to the construction of mathematical models that capture important nonlinear phenomena. This is difficult and

takes a long time. We see this in other fields too -- the dynamic general static equilibrium (DGSE) models of the macroeconomy have been built on many, many economics studies, and are still not perfect predictors of our economic future. Developing metrics to evaluate science, for all these reasons, is likely to fail. At best it will distort science and constrain genuinely new ideas. If metrics are applied in a backward-looking way to evaluate which ideas are good -- well, we already have measures of success in science, that involve persuading the scientific community and the public, or assessing commercial or medical impact. And adding new metrics that will necessarily be simplifications of what we want to measure merely creates a temptation to game the metrics.

*** Measurements and metrics will be gamed by humans, and thus support the status quo and suppress innovation

Perhaps the most dangerous feature of new metrics is described by Goodhart's Law, coined by British economist Charles Goodhart in 1975: "When a measure becomes a target, it ceases to be a good measure." Metrics go bad because humans react to incentives and find ways to target and game metrics. Andrew Likierman wrote an article called "The Five Traps of Performance Metrics" in the Harvard Business Review (https://hbr.org/2009/10/the-five-traps-of-performance-measurement). One of Likierman's traps is "Gaming your Metrics." He suggests using a diversity of metrics and changing them frequently so they cannot be gamed. But any metric created by NIH is likely to not change fast enough --- given the many stakeholders and democratic accountability of NIH, including to Congress --- to satisfy Likierman's advice. Science requires more nimbleness, requires investigators empowered to quickly follow the data where it leads.

Moreover, metrics tend to support the status quo. To avoid gaming, metrics must change over time. But that means that any metric is a trailing measure of important factors in a complex system. These factors cut against innovation.

It should also be noted that wealthy interests invested in the status quo have often funded quantitative work in social systems, and those efforts have benefited the wealthy by reinforcing the status quo. One example is the "Law and Economics" effort in the law. That moved antitrust law in a direction that benefited entrenched and powerful interests. (See e.g. Heyer K, /Consumer Welfare and the Legacy of Robert Bork/, 2015, the Journal of Law and Economics: "Bork's antitrust analysis drew largely on basic microeconomic theory"... Khan, L /Amazon's Antitrust Paradox/, "unequipped to capture the architecture of market power in the modern economy...").

Aspects of the quantitative-first "Metascience" effort show signs of this capture by wealthy interests. While the vast majority of metascience researchers are engaged in a good-faith pursuit of truth, several thinktanks that support this work are funded by libertarian billionaires or have leadership or board members with libertarian training or current ties. We should be careful that metascience is not used as a way to dull American public innovation — to benefit a few wealthy donors — by reinforcing the status quo.

*** Peer review of qualitative written and oral materials prepared by creative scientists is the gold standard for evaluating science

If metrics are not the way forward to evaluate science, what is? We have decades of experience in the

United States of gold-standard methods for evaluating science, and that relies on qualitative peer review, where rotating panels of experts think carefully about past and proposed work and evaluate it.
The most obvious example of peer review of funding is the NIH extramural grant review system. Peer review panel membership changes over time, which limits the possibility of capture by a few. When peer review is working well, twenty to thirty percent of applications are funded, and reviewers read proposals carefully. The proposals are developed by scientists with significant investment of time, and again in an ideal situation scientists write no more than one grant application every year or two.

But review by peers is also the way that scientific impact is evaluated. Persuading members of the scientific community that ideas are good, to create scientific consensus, is the way impact has been successfully measured.

Two other examples of successful peer review come from the NIH intramural review process and the related HHMI review process. In both cases the scientific research programs to be reviewed are described by the principal investigator in written documents and then the expert panel convened for review hears an oral presentation by the scientist. It can be an advantage to do this kind of investigator, not project-based, review --- which can weight past track record heavily and trust the investigator to do future high-risk work. NIH might consider more widespread evaluation of scientific impact via expert panels of that sort.

There are multiple options for ways to conduct peer review and NIH should experiment with different approaches. However, the feature at the core of each of these methods is that the scientific judgement of many experts is harnessed.

Reducing scientific evaluation to metrics is a recipe for disaster. We should have humility about what it is possible to capture in a complex system like science with quantitative measures alone. Data and measurement can supplement expert decision-making but should never play a primary or even major role in choice of the best science. Science is a creative, forward-looking process and we should treat it as such.

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/HISTED-response-to-RFI-on-Measuring-And-Rewarding-Scientific-Impact-260819.pdf

Description: Response -- applies to all categories listed here.

Comment on RFI on Measuring And Rewarding Scientific Impact

Mark H. Histed, PhD, commenting in personal capacity and not on behalf of any organization

"NIH seeks public input on the defining features of modern 21st-century science that should be measured, recognized, and incentivized. "

Executive summary: Quantitative measurements of science, and other creative fields, inherently reduce innovation and limit big ideas. New metrics and measurement can in some cases add information, but NIH should strive to de-emphasize metrics in any evaluation or incentive structures in science. Instead of attempting to reduce the process of science to inherently limited data and measurement, NIH should rely on the gold-standard procedure of evaluating science that has been used to great success in the US for decades: peer review from the scientific community. This peer review should be applied to materials and research papers developed as a product of the creative and intellectual work of scientists. Committees of scientists that read papers will do better at assessing the value of new ideas than any quantitative measure or metric, which will always be backward-looking..

The science ecosystem is a complex and nonlinear system impossible to characterize fully with data and measurement

Science is a highly complex social system which relies on frontier thinking from tens of thousands of scientists. Any empirical model or data will have difficulty assessing the quality of a complex system like science, because empirical models are backward-looking --- data themselves are observations of past events. Thus, quantitative evaluation will ***never*** be a complete way to capture the complex system that is science. NIH should therefore be incredibly cautious of bringing new metrics to bear to evaluate innovation.

To illustrate how empirical models alone are limited for understanding the world, without the accompanying development of theoretical models and a back-and-forth process that adjust models (or metrics) as new observations arrive, let me give an example. Neuroscience, my field, shows how data and quantitative measurements are useful — but are not a full description of a complex system. In the cerebral cortex, excitatory and inhibitory neural networks are tightly coupled to each other. This produces a phenomenon called the balanced state mode of network operation, which increases sensitivity to input, and decouples the activity of nearby neurons to maximize the ability to read out information carried by their responses. Cortical networks in the balanced state have coupled local excitation, and thus exhibit the inhibition-stabilized network (ISN) property. That is a property in which self-amplified activity that can boost certain input patterns is tracked and kept stable by local inhibition.

I give this background to note that ISN property would likely never have been found by pure experimental measurement or data collection. In 1987, Misha Tsodyks and a group of American and international collaborators (https://pubmed.ncbi.nlm.nih.gov/9151754/) showed that ISNs produce

'paradoxical responses,' where boosting inhibition unexpectedly causes excitatory cells to increase, not decrease, their activity. In a two-variable continuous-time model they demonstrated that linear perturbations in the model phase space produce unintuitive effects, curved paths in the phase space that change direction from the marginal or initial incremental direction. Many years later, focused experiments to test this nonlinear model by the Miller, Ferster, and my own group, amongst others, showed this property holds in the cortex.

The ISN model is merely a two-variable system and yet produces complex dynamics. And most empirical measurements --- applying regressions to large-scale data, for example --- measure only the initial marginal effect and missed these curved paths in the state space. Even adding nonlinear terms to a regression does not do a good job inferring curved/non-marginal dynamics, not least because more complex non-linearities become increasingly data-hungry to estimate. That is, most measurements that can be collected in finite time rely on a local linearization of a bigger complex system.

To quantitatively describe real-world complex systems, whether brain networks or the incredibly complex sociological system that is a science and innovation economy, we must not just develop metrics based on data. We must move beyond that to apply the simple models imposed by data to the construction of mathematical models that capture important nonlinear phenomena. This is difficult and takes a long time. We see this in other fields too -- the dynamic general static equilibrium (DGSE) models of the macroeconomy have been built on many, many economics studies, and are still not perfect predictors of our economic future. Developing metrics to evaluate science, for all these reasons, is likely to fail. At best it will distort science and constrain genuinely new ideas. If metrics are applied in a backward-looking way to evaluate which ideas are good -- well, we already have measures of success in science, that involve persuading the scientific community and the public, or assessing commercial or medical impact. And adding new metrics that will necessarily be simplifications of what we want to measure merely creates a temptation to game the metrics.

Measurements and metrics will be gamed by humans, and thus support the status quo and suppress innovation

Perhaps the most dangerous feature of new metrics is described by Goodhart's Law, coined by British economist Charles Goodhart in 1975: "When a measure becomes a target, it ceases to be a good measure." Metrics go bad because humans react to incentives and find ways to target and game metrics. Andrew Likierman wrote an article called "The Five Traps of Performance Metrics" in the Harvard Business Review (https://hbr.org/2009/10/the-five-traps-of-performance-measurement). One of Likierman's traps is "Gaming your Metrics." He suggests using a diversity of metrics and changing them frequently so they cannot be gamed. But any metric created by NIH is likely to not change fast enough --- given the many stakeholders and democratic accountability of NIH, including to Congress --- to satisfy Likierman’s advice. Science requires more nimbleness, requires investigators empowered to quickly follow the data where it leads.

Moreover, metrics tend to support the status quo. To avoid gaming, metrics must change over time. But that means that any metric is a trailing measure of important factors in a complex system. These factors cut against innovation.

It should also be noted that wealthy interests invested in the status quo have often funded quantitative work in social systems, and those efforts have benefited the wealthy by reinforcing the status quo. One example is the "Law and Economics" effort in the law. That moved antitrust law in a direction that benefited entrenched and powerful interests. (See e.g. Heyer K, /Consumer Welfare and the Legacy of Robert Bork/, 2015, the Journal of Law and Economics: “Bork’s antitrust analysis drew largely on basic microeconomic theory”... Khan, L /Amazon’s Antitrust Paradox/, “unequipped to capture the architecture of market power in the modern economy...”).

Aspects of the quantitative-first "Metascience" effort show signs of this capture by wealthy interests. While the vast majority of metascience researchers are engaged in a good-faith pursuit of truth, several thinktanks that support this work are funded by libertarian billionaires or have leadership or board members with libertarian training or current ties. We should be careful that metascience is not used as a way to dull American public innovation — to benefit a few wealthy donors — by reinforcing the status quo.

Peer review of qualitative written and oral materials prepared by creative scientists is the gold standard for evaluating science

If metrics are not the way forward to evaluate science, what is? We have decades of experience in the United States of gold-standard methods for evaluating science, and that relies on qualitative peer review, where rotating panels of experts think carefully about past and proposed work and evaluate it.

The most obvious example of peer review of funding is the NIH extramural grant review system. Peer review panel membership changes over time, which limits the possibility of capture by a few. When peer review is working well, twenty to thirty percent of applications are funded, and reviewers read proposals carefully. The proposals are developed by scientists with significant investment of time, and again in an ideal situation scientists write no more than one grant application every year or two.

But review by peers is also the way that scientific impact is evaluated. Persuading members of the scientific community that ideas are good, to create scientific consensus, is the way impact has been successfully measured.

Two other examples of successful peer review come from the NIH intramural review process and the related HHMI review process. In both cases the scientific research programs to be reviewed are described by the principal investigator in written documents and then the expert panel convened for review hears an oral presentation by the scientist. It can be an advantage to do this kind of investigator, not project-based, review --- which can weight past track record heavily and trust the investigator to do

future high-risk work. NIH might consider more widespread evaluation of scientific impact via expert panels of that sort.

There are multiple options for ways to conduct peer review and NIH should experiment with different approaches. However, the feature at the core of each of these methods is that the scientific judgement of many experts is harnessed.

Reducing scientific evaluation to metrics is a recipe for disaster. We should have humility about what it is possible to capture in a complex system like science with quantitative measures alone. Data and measurement can supplement expert decision-making but should never play a primary or even major role in choice of the best science. Science is a creative, forward-looking process and we should treat it as such.

285. NIAID Office of Data Science and Emerging Technologies

I am responding: On behalf of an organization

Name: Rachel Waymack

Name of Organization: NIAID Office of Data Science and Emerging Technologies

Type of Organization: Government Agency

Role: Government Official

Date Submitted: 8/19/2026

1. Rigor and Reproducibility: NIH should emphasize that sharing data, software, and computational models, specifically in such a way that an outside researcher can understand and reuse them, is required for rigorous and reproducible science. NIH should provide incentives for and clear guidance on best sharing practices.

As biomedical research increasingly creates and depends on “big data” and computational methods, rigor and reproducibility in this field increasingly depend on researchers adequately sharing and describing all of the data, software, and computational models used in their studies. We found that reproducibility often fails not necessarily due to the underlying science, but due to data, analysis details, or other aspects not being shared or sufficiently described.

In our proof-of-concept exercise, the most common roadblocks to reproducibility that we encountered were:
- Data and/or code not shared
- Code shared, but with insufficient documentation for installation or use
- Data shared, but without a data dictionary or other field-level description
- Raw data shared, but without sufficient description of processing and analysis methods used to generate the published results

We suggest that NIH emphasize sharing that enables reproducibility. If outside researchers do not have enough information to be able to use shared data or code to reproduce findings, this is not sufficient sharing. NIH should therefore incentivize and provide guidance on sharing data, software, and computational models in such a way that enables others to easily reuse these materials.

Indicators NIH could consider in this area include:
- Whether all data needed to evaluate the main findings of a publication are deposited in a repository with the desirable characteristics laid out by NIH (e.g., assigning unique persistent identifiers, requiring metadata, etc.)
- Whether said data is shared with relevant metadata and/or data dictionaries
- Whether relevant code is deposited in public repositories with version control
- Whether said code is shared with sufficient documentation for another researcher to install and successfully use the code

- Whether sufficient detail of preprocessing steps is provided to move from raw data to analytic datasets
- Whether workflows are shared in ways that support reuse such as conda packages or containers
- Whether relevant community standards and interoperable formats are used when possible
- Whether persistent identifiers are assigned to datasets, software, models, and other research outputs

One implementation approach would be for NIH to require that funded studies deposit the minimum set of materials (i.e. data, code, computational workflows, etc.) needed to reproduce key published findings in appropriate repositories, except where legal or privacy constraints prevent full public release. NIH should provide clear expectations for adequate sharing for computational reproducibility, such as minimum standards for data and workflow descriptions. These expectations would need to be coupled with NIH funding to support training and related efforts such as curation, annotation, and repository maintenance to ensure the requirements are not overly burdensome.

2. Data, Software, and Model Sharing: NIH should emphasize that sharing data, software, and computational models, specifically in such a way that an outside researcher can understand and reuse them, is required for rigorous and reproducible science. NIH should provide incentives for and clear guidance on best sharing practices.

As biomedical research increasingly creates and depends on “big data” and computational methods, rigor and reproducibility in this field increasingly depend on researchers adequately sharing and describing all of the data, software, and computational models used in their studies. We found that reproducibility often fails not necessarily due to the underlying science, but due to data, analysis details, or other aspects not being shared or sufficiently described.

In our proof-of-concept exercise, the most common roadblocks to reproducibility that we encountered were:
- Data and/or code not shared
- Code shared, but with insufficient documentation for installation or use
- Data shared, but without a data dictionary or other field-level description
- Raw data shared, but without sufficient description of processing and analysis methods used to generate the published results

We suggest that NIH emphasize sharing that enables reproducibility. If outside researchers do not have enough information to be able to use shared data or code to reproduce findings, this is not sufficient sharing. NIH should therefore incentivize and provide guidance on sharing data, software, and computational models in such a way that enables others to easily reuse these materials.

Indicators NIH could consider in this area include:
- Whether all data needed to evaluate the main findings of a publication are deposited in a repository with the desirable characteristics laid out by NIH (e.g., assigning unique persistent identifiers, requiring metadata, etc.)
- Whether said data is shared with relevant metadata and/or data dictionaries
- Whether relevant code is deposited in public repositories with version control
- Whether said code is shared with sufficient documentation for another researcher to install and successfully use the code

- Whether sufficient detail of preprocessing steps is provided to move from raw data to analytic datasets
- Whether workflows are shared in ways that support reuse such as conda packages or containers
- Whether relevant community standards and interoperable formats are used when possible
- Whether persistent identifiers are assigned to datasets, software, models, and other research outputs

One implementation approach would be for NIH to require that funded studies deposit the minimum set of materials (i.e. data, code, computational workflows, etc.) needed to reproduce key published findings in appropriate repositories, except where legal or privacy constraints prevent full public release. NIH should provide clear expectations for adequate sharing for computational reproducibility, such as minimum standards for data and workflow descriptions. These expectations would need to be coupled with NIH funding to support training and related efforts such as curation, annotation, and repository maintenance to ensure the requirements are not overly burdensome.

3. Training and Mentorship: NIH training programs should include education and mentorship in FAIR principles and reproducibility best practices.
At the end of the day, any science being performed will only be as good and as rigorous as the scientists performing it. For this reason, any efforts to strengthen rigor and reproducibility in biomedical research should emphasize initiatives to train researchers in data management and sharing best practices. This would likely address many of the barriers to reproducibility we encountered in our exercise.
We therefore recommend that NIH prioritize training and mentorship efforts that instill in upcoming and current researchers the best practices for ensuring reproducibility. Potential indicators NIH could consider in this area include:
- Training programs that cover data management, metadata creation and curation, code sharing, and workflow documentation
- Requiring FAIR principles and other reproducibility practices be incorporated in trainee education programs across specialties

4. Collaboration: NIH should expand policies to assess and recognize collaborative contributions to science, beyond first and senior authorship.
We appreciate that NIH acknowledges science is increasingly team-based and is seeking input on how best to recognize such collaborative contributions. From our perspective as data scientists, we recognize that teams with computational and data-sharing expertise are often better positioned to produce transparent and reproducible research outputs.
Potential metrics for NIH to consider for assessing and recognizing contributions to team-based science include:
- Contributions to data curation, software development, and software/workflow maintenance
- Use of shared standards, repositories, and workflows
- Development of resources (data repositories, software tools, etc.) that enable reuse by others
NIH should encourage more explicit documentation of team roles and recognize contributions beyond first or senior authorship through the development of better metrics to track data and software sharing (for example, the S-index), events focused on highlighting exceptional examples of reproducible science, or other avenues.

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/ResponsetoRFIonRigor.docx

Description: This submission reflects ODSET's perspective on approaches to measuring and rewarding scientific impact but does not represent the official NIAID position.

Request for Information on Measuring and Rewarding Scientific Impact

Response to: NIH Guide Notice NOT-OD-26-087

Prepared by:

Rachel Waymack and members of the NIAID Office of Data Science and Emerging Technologies: Reed Shabman, Sydney Foote, Meghan Hartwick, Lisa Mayer, Dasha Pokutnaya.

Disclaimer: The response below represents perspectives and points to consider from NIAID subject matter experts on this topic and does not necessarily reflect the official NIAID position. As such, the response uses 'we' in reference to the contributors of this document, not 'NIAID'

A. Executive Summary

The NIAID Office of Data Science and Emerging Technologies (ODSET) includes a diverse group of data scientists. ODSET partners with NIAID Divisions and the extramural community to advance data science and computational modeling in infectious, allergic, and immunological diseases. ODSET supports methods and approaches that enable biomedical data, computational methods, and software tools to be findable, accessible, interoperable, and reusable (FAIR). We appreciate NIH's efforts to best measure and reward scientific impact and believe these efforts should be grounded in the FAIR principles.

Recently, ODSET conducted a proof-of concept exercise in which we sampled a limited number of NIH-funded publications and attempted to reproduce select figures or findings using the data and methods made available by the authors. In most cases we were unable to do so, often due to data or code not being shared or shared with inadequate documentation. While not a systematic study, this small exercise reinforced the point that reproducibility depends not just on sharing data or code, but on sharing in ways that make them usable by others.

From this perspective, the main points to consider in support of rigorous and reproducible science are:

1. NIH should incentivize and structurally support sharing research data ***and corresponding methods (e.g. computational workflows)*** in such a way that an outside researcher can understand and use them (expanded upon in section B.1).
2. NIH training programs should include education and mentorship in FAIR principles and reproducibility best practices (expanded upon in section B.2).
3. NIH should expand policies to assess and recognize collaborative contributions to science, beyond first and senior authorship, including but not limited to data managers, computational modelers, and software developers (expanded upon in section B.3).

We welcome further discussion with NIH colleagues, such as the Office of Research Innovation, Validation, and Application, and can be contacted at datascience@niaid.nih.gov or rachel.waymack@nih.gov.

We offer additional detailed comments below, which are broken down by category and numbered in alignment with the main points above:

B. Detailed response

B.1. Adequate sharing of data and computational workflows fosters rigor and reproducibility.

Response to sections: Rigor and Reproducibility; Data, Software, and Model Sharing

NIH should emphasize that sharing data, software, and computational models, specifically in such a way that an outside researcher can understand and reuse them, is required for rigorous and reproducible science. NIH should provide incentives for and clear guidance on best sharing practices.

As biomedical research increasingly creates and depends on "big data" and computational methods, rigor and reproducibility in this field increasingly depend on researchers adequately sharing and describing all of the data, software, and computational models used in their studies. We found that reproducibility often fails not necessarily due to the underlying science, but due to data, analysis details, or other aspects not being shared or sufficiently described.

In our proof-of-concept exercise, the most common roadblocks to reproducibility that we encountered were:

- Data and/or code not shared

- Code shared, but with insufficient documentation for installation or use
- Data shared, but without a data dictionary or other field-level description
- Raw data shared, but without sufficient description of processing and analysis methods used to generate the published results

We suggest that NIH emphasize sharing that enables reproducibility. If outside researchers do not have enough information to be able to use shared data or code to reproduce findings, this is not sufficient sharing. NIH should therefore incentivize and provide guidance on sharing data, software, and computational models in such a way that enables others to easily reuse these materials.

Indicators NIH could consider in this area include:

- Whether all data needed to evaluate the main findings of a publication are deposited in a repository with the desirable characteristics laid out by NIH (e.g., assigning unique persistent identifiers, requiring metadata, etc.)
- Whether said data is shared with relevant metadata and/or data dictionaries
- Whether relevant code is deposited in public repositories with version control
- Whether said code is shared with sufficient documentation for another researcher to install and successfully use the code
- Whether sufficient detail of preprocessing steps is provided to move from raw data to analytic datasets
- Whether workflows are shared in ways that support reuse such as conda packages or containers
- Whether relevant community standards and interoperable formats are used when possible
- Whether persistent identifiers are assigned to datasets, software, models, and other research outputs

One implementation approach would be for NIH to require that funded studies deposit the minimum set of materials (i.e. data, code, computational workflows, etc.) needed to reproduce key published findings in appropriate repositories, except where legal or privacy constraints prevent full public release. NIH should provide clear expectations for adequate sharing for computational reproducibility, such as minimum standards for data and workflow descriptions. These expectations would need to be coupled with NIH funding to support training and related efforts such as curation, annotation, and repository maintenance to ensure the requirements are not overly burdensome.

B.2. NIH training programs should include education and mentorship in FAIR principles and reproducibility best practices.

Response to section: Training and Mentorship

NIH training programs should include education and mentorship in FAIR principles and reproducibility best practices.

At the end of the day, any science being performed will only be as good and as rigorous as the scientists performing it. For this reason, any efforts to strengthen rigor and reproducibility in biomedical research should emphasize initiatives to train researchers in data management and sharing best practices. This would likely address many of the barriers to reproducibility we encountered in our exercise.

We therefore recommend that NIH prioritize training and mentorship efforts that instill in upcoming and current researchers the best practices for ensuring reproducibility. Potential indicators NIH could consider in this area include:

- Training programs that cover data management, metadata creation and curation, code sharing, and workflow documentation
- Requiring FAIR principles and other reproducibility practices be incorporated in trainee education programs across specialties

B.3. NIH should expand policies to assess and recognize collaborative contributions to science.

Response to section: Collaboration

NIH should expand policies to assess and recognize collaborative contributions to science, beyond first and senior authorship.

We appreciate that NIH acknowledges science is increasingly team-based and is seeking input on how best to recognize such collaborative contributions. From our perspective as data scientists, we recognize that teams with computational and data-sharing expertise are often better positioned to produce transparent and reproducible research outputs.

Potential metrics for NIH to consider for assessing and recognizing contributions to team-based science include:

- Contributions to data curation, software development, and software/workflow maintenance
- Use of shared standards, repositories, and workflows

- Development of resources (data repositories, software tools, etc.) that enable reuse by others

NIH should encourage more explicit documentation of team roles and recognize contributions beyond first or senior authorship through the development of better metrics to track data and software sharing (for example, the S-index), events focused on highlighting exceptional examples of reproducible science, or other avenues.

286. BioCrowd, Inc.

I am responding: On behalf of an organization

Name: Mark Stoner

Name of Organization: BioCrowd, Inc.

Type of Organization: Commercial Entity (Biotech/Device/Pharmaceutical Company/Small Business)

Role: Other

Role - Other: We're building BioCrowd, a capital deployment platform for clinical research, beginning with space biology. Microgravity accelerates biological changes associated with aging—including muscle loss, immune dysfunction, neuroinflammation, and cardiovascular decline—making space a powerful environment for studying age-related disease affecting longevity. BioCrowd helps universities and research institutions fund these programs and translate their scientific milestones into investable opportunities. https://www.biocrowd.com/

Date Submitted: 8/20/2026

1. Rigor and Reproducibility: Measurable indicators. Preregistration or registered analysis plans; availability of protocols, materials, code, and statistical specifications; validation of key resources; reporting of sample-size logic and sensitivity analyses; documented deviations; reporting of null or negative findings; independent reproduction or replication attempts; and whether later evidence strengthens, narrows, or refutes the original claim. For causal research, records should identify the causal question, estimand, assumptions, identification strategy, comparator, prespecified falsification or sensitivity tests, and limits on generalizability.

Potential incentives. Dedicated replication awards, administrative supplements, bridge funding for confirmatory work, recognition in progress reviews and subsequent applications, and formal credit for high-quality negative or inconclusive results.

Safeguard. Do not reward a positive replication rate. Reward transparent, rigorous execution and complete reporting regardless of result; otherwise, the system may encourage selective targets or suppression of discordant findings.

2. Data, Software, and Model Sharing: Measurable indicators. Persistent identifiers; metadata completeness; community standards and common data elements; time from primary result to availability; documented access conditions; FAIR characteristics; versioned and citable software or models; executable environments, tests, model cards, and provenance; maintenance status; and evidence of responsible reuse.

Potential incentives. Treat datasets, software, models, protocols, and curation as first-class research products in biosketches and review; fund stewardship and maintenance; and provide supplemental support when reuse demonstrates unusual public value.

Safeguard. Downloads alone are not impact, and unrestricted release is not always appropriate. Metrics should not penalize privacy-preserving controlled access, community governance, licensing obligations, cybersecurity, biosecurity, or legitimate protections needed for translation.

3. Training and Mentorship: Measurable indicators. Documented mentoring plans; protected mentoring time; trainee-defined goals; completion of rigor, quantitative, data, and responsible-conduct training; verified trainee contributions; career progression across academic, clinical, government, nonprofit, and industry pathways; and confidential mentorship-quality feedback.

Potential incentives. Recognize substantive mentoring in grant review and progress reporting, support mentor development, protect salary effort for mentoring, and reward teams that produce durable skills and career development across diverse pathways.

Safeguard. Raw trainee counts and prestige of placements can disadvantage small laboratories and nontraditional careers. Evaluate quality, opportunity, and value added relative to context rather than pedigree.

4. Collaboration: Measurable indicators. Verified contributorship; cross-disciplinary and cross-institutional roles; timely execution of data-use, material-transfer, and reliance agreements; shared protocol and data standards; patient and community participation where relevant; resolution of documented bottlenecks; and continuity of shared resources after the initial award.

Potential incentives. Permit structured team-science contribution statements, give reviewers explicit guidance to credit enabling roles, recognize shared research operations, and support modest coordination supplements tied to transparent deliverables.

Safeguard. Counting collaborators, institutions, or disciplines rewards superficial network size. Assess necessity, contribution quality, governance, conflict management, and delivered value.

DeSci. BioCrowd is building decentralized science (DeSci) into our long-term platform engagement models between R1, R1 institutions and individual scientists. This DeSci can be very helpful for accelerating science and reducing silos internal to institutions, between institutions in the US but internationally also. Yet still be able to protect developed intellectual property (IP) rights for investors and the universities that support them. DeSci can be very helpful for biodefense, pandemic response or rapid deployment research needs. This Ethereum website details how DeSci works and the benefits. https://ethereum.org/desci/

5. Entrepreneurship and Translation: Measurable indicators. Completion of need validation; reproducibility and technical-readiness evidence; invention disclosures and patents where appropriate; regulatory, manufacturing, reimbursement, implementation, or clinical-development milestones; licenses, responsible startups, sponsored research, and follow-on capital; time to the next decision-quality milestone; and evidence of accessibility, adoption, or patient benefit.

Potential incentives. Translation supplements, milestone-based awards, leave or mobility mechanisms, entrepreneur-in-residence access, regulatory and commercialization assistance, and recognition of validated handoffs to capable partners.

Safeguard. Company formation, valuation, patents, and dollars raised are incomplete proxies. NIH should reward credible de-risking and public value, require conflict disclosure, and keep public or charitable research decisions separate from private investment decisions.

6. Foundational Scientific Exploration: Measurable indicators. Clarity and testability of the question; novelty relative to the field; potential to create a new method, dataset, instrument, model, or conceptual framework; informative value of failure; challenge to an important assumption; future option value; and retrospective evidence that the work opened productive lines of inquiry.

Potential incentives. Staged high-risk awards with light initial burden, opportunities to continue or pivot based on learning, dedicated support for early-career and overlooked investigators, and portfolio-level evaluation that accepts well-designed failures in exchange for rare breakthroughs.
Safeguard. Rigid milestones can prematurely narrow exploratory work. Milestones should emphasize learning, methods, evidence quality, and explicit go, pivot, or stop decisions - not predetermined favorable findings.

7. Public Impact: Measurable indicators. Links to clinical trials, guidelines, standards, public-health programs, diagnostics, therapeutics, devices, implementation studies, and policy; changes in practice; reach to intended populations; time to access; health outcomes where attribution is supportable; reduction in burden or cost; domestic capability and workforce; public trust; and distribution of benefits across regions and populations.

Potential incentives. Follow-on implementation and dissemination awards, recognition for contributions to guidelines and standards, longitudinal linkage between awards and downstream records, and public-impact narratives supported by verifiable evidence.
Safeguard. Health impact often occurs years later and usually has multiple causes. Use contribution analysis and uncertainty ranges, avoid claims of sole attribution, and prevent near-term outcomes from crowding out basic science and rare-disease research.

8. Other Areas: 5.3 Incentive Design Lessons from the World Intellectual Property Organization (WIPO)

WIPO's 2024 guide on incentives in technology transfer offers a useful framework for an NIH DeSci pilot. It groups incentives into non-financial recognition, career advancement, and financial support; emphasizes clear, transparent and achievable criteria; recommends tailoring incentives to people, institutions and stages; and warns that a program can hit a target, such as producing more patents, while missing the underlying goal of durable economic or societal value. WIPO also advises rewarding technology-transfer effort and enabling contributions, not only downstream outcomes, and monitoring programs for unintended effects. [7]
For NIH, this supports a layered incentive portfolio in which DeSci infrastructure supplies verifiable evidence of contribution but does not determine scientific merit. The portfolio should recognize investigators, technology-transfer professionals, research software engineers, data stewards, regulatory staff, technicians, trainees, patients, and community partners according to their actual roles.
• Non-financial recognition: verified contributor records, public acknowledgment, reusable-output badges, leadership recognition, and visibility for rigorous negative, null, replication, stewardship, and translation work.

• Career advancement: structured credit in biosketches, progress reports, promotion dossiers, and team-science statements for reproducibility, data and software stewardship, mentorship, regulatory execution, and technology-transfer enabling work.
• Financial and resource support: administrative supplements, proof-of-concept or translational funds, protected staff time, training, conference support, and small fixed-dollar awards for bounded deliverables.
• Stage-appropriate timing: use near-term recognition and resources at protocol, evidence, validation, and handoff milestones rather than relying only on distant licensing income or commercialization events.
• Goal-aligned measurement: pair quantitative indicators with qualitative evidence and test whether incentives advance trustworthy knowledge and public benefit, not merely counts of patents, companies, publications, or transactions.
• Safeguards: preserve the right to publish subject to limited confidentiality and IP-protection review; protect foundational research; provide choice where feasible; and regularly audit for gaming, inequity, administrative burden, and displacement of intrinsic motivation.

Description: BIOCROWD, INC. Response to NIH Request for Information Measuring and Rewarding Scientific Impact | NOT-OD-26-087 Illustrated public submission draft | WIPO incentive and bounded-bounty update | August 19, 2026

287. SciClin LLC

I am responding: On behalf of an organization

Name:

Name of Organization: SciClin LLC

Type of Organization: Research Participant/Patient Advocacy Organization/Patient Advocate/Caregiver

Role: Investigator/Researcher

Date Submitted: 8/20/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: N/A

4. Collaboration: N/A

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: N/A

7. Public Impact: N/A

8. Other Areas: N/A

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/RFI_NIH-Report.pdf

Josly Pierre-Louis Odoom, PhD MS
Eva Davis, PhD
Annie Hinson, PhD
Christie Pierre-Louis, BS

Response to NIH RFI: Measuring and Rewarding Scientific Impact

Submitted by: SciClin LLC

SciClin is a patient-centered translational science consultancy

About This Response

SciClin works with academic research teams and institutions to build translational strategy into biomedical research at the design stage before data collection, not after. Our perspective on this RFI is shaped by direct, applied experience: we regularly observe where studies with strong internal validity nonetheless fail to reproduce cleanly or translate into clinical practice, and the root cause is almost always the same: the research question was never anchored to the patient population or clinical context it was ultimately meant to serve. In preparation for this response, we conducted a pilot qualitative study comprising six semi-structured, de-identified interviews with basic researchers and clinicians across multiple disease areas in the translational pipeline. We draw on that dataset throughout this response as applied evidence, alongside our broader workshop and advisory experience, and would welcome the opportunity to share the full coded dataset directly with NIH staff. Our team and professional network span institutions including Virginia Commonwealth University, University of Southern California, Indiana University, University of Rochester, and Longwood University. We have also previously discussed related priorities directly with NIH program staff.

Framing Comment

Current metrics of scientific success: publication counts, citation indices, and grant dollars largely measure activity within a single stage of the research pipeline. They do not capture whether a finding was ever positioned to survive contact with the patients, clinicians, and health systems it was meant to serve. The NIH's Replication and Reproducibility Initiative is an important corrective. Still, we would encourage the agency to resist measuring"rigor" and "impact" as sequential checkpoints independent of each other: rigor at the bench, impact at the bedside. In our work advising research teams on translational design, we consistently find that studies which reproduce reliably and translate successfully share a common feature: patient and clinical context was built into the research question and study design from the outset,

not retrofitted after results were generated. Reproducibility, in other words, is not only a property of a mechanistic finding in isolation, but it is also a property of how well that finding was designed to hold up as it moves toward real-world application. We believe any indicator framework should measure both dimensions together, and reward research designs that account for the second at the moment the first is being planned, rather than treating translational relevance as something assessed only once a discovery is already mature.

This has a direct bearing on how NIH defines "causal research," which the RFI specifically raises. A causal claim generated from a model system with undocumented relevance to the target patient population is, in practice, a claim about that model system alone; its causal status in the intended clinical context remains untested. We would encourage NIH to treat clinical/population context documentation as a prerequisite for evaluating the strength of a causal claim, not a separate translational concern.

On Rigor and Reproducibility

We recommend NIH consider indicators that assess reproducibility readiness at the design stage, not only replication outcomes after the fact. Concretely, this could take the form of a scorable field in grant applications and progress reports assessing whether a study's model system, patient-derived materials, or clinical endpoints are documented with enough context: disease stage, population characteristics, comorbidities, and care setting that a downstream investigator or clinical partner could evaluate the finding's applicability without needing to contact the original authors.

In our applied work with research teams, this single gap in the absence of patient/clinical context in method reporting is one of the most common and correctable barriers we encounter to both replication and translation. It is also, notably, inexpensive to fix relative to most rigor interventions: it requires no new wet-lab infrastructure, no additional data collection, and no extension of study timelines, only a documentation habit built into protocol design. A simple present/partial/absent scoring rubric could be piloted within existing progress-report or Research Performance Progress Report (RPPR) structures at low administrative cost, and could be validated retrospectively against existing replication failure data before being formally adopted. We would also note that this indicator is measurable by a third party

without requiring the original investigator's self-assessment, which addresses a common critique of rigor metrics that they rely on investigators grading their own work.

This pattern was directly reflected in our pilot interviews: each study participant (across both basic-research and clinical roles) independently raised structural or institutional barriers to integrating patient/clinical context into their work, and the same group named funding and career-incentive structures as compounding factors. Participants also drew a sharp, consistent line between patient/provider input that was merely "acknowledged" after the fact and input incorporated early enough to actually change experimental design, endpoints, or population definitions; the latter was the only version participants could point to that changed a real research decision. This maps directly onto NIH's reproducibility question where design-stage context improves rigor by ensuring relevance remains at the center of scientific discovery. We are positioned to make this dataset available to NIH staff for review.

On Data, Software, and Model Sharing

We would encourage NIH to extend its sharing indicators beyond raw data and code to include the *contextual metadata* that makes shared data usable for translational purposes, specifically, structured documentation of the population, clinical setting, and stage of disease represented in shared datasets and model systems. A dataset or cell line shared without this context can still be technically "open" while remaining functionally unusable to a downstream team trying to assess its relevance to a specific patient population. We would suggest NIH consider a minimum contextual-metadata standard as a condition of data/model sharing credit, similar in spirit to existing FAIR data principles, but specifically oriented toward translational reuse rather than technical reproducibility alone.

On Training and Mentorship

We would encourage NIH to recognize and fund training that builds translational design literacy directly into graduate and postdoctoral training, not as a supplemental commercialization or entrepreneurship elective late in a trainee's career, but as a core research-design skill taught alongside statistics and experimental design. In our own workshop-based work with research teams, we have found that translational thinking is most effective, and most durable, when it is taught as part of how a hypothesis is

framed, rather than added afterward as a business or dissemination skill. An indicator here might track institutional investment in this kind of design-stage training (distinct from tech-transfer or startup-formation training) as a component of training grant (T32, etc.) review. We are currently developing asynchronous self-paced e-learning courses for research faculty and learners to understand how to build strategic partnerships and design translational research to improve translation from bench to bedside.

Current grant pipelines reward mechanistic breakthroughs and do not fund the necessary activities (strategic partnership building, regulator readiness, industry-academic collaboration, community advisory board, provider/patient input) that bolster translational research. Thus, many scientific discoveries halt at publication and never evolve to therapeutic innovation in clinical practice. Also, translational research is much more time-intensive than curiosity-driven research. For impact to occur sustainably, there needs to be funding available distinct from the traditional R01 grant system for bench researchers who seek to identify mechanisms at the bench and translate these ideas into impact for patients.

On Collaboration

Cross-disciplinary collaboration is often assessed by counting co-authorship across departments or institutions, which can reward administrative breadth without capturing whether a collaboration meaningfully improved a study's design. We would suggest NIH consider indicators that specifically credit collaborations bridging basic science and clinical/patient-facing expertise; for example, whether a clinician, patient advocate, or clinical operations expert was involved during protocol design (not only during a later clinical trial phase). This is a narrower, more targeted version of team science credit that maps directly onto the translational goals this RFI is oriented around. Having the principal investigator list both a provider/clinician and a community advisory stakeholder as part of the research program in grant review will significantly bolster impact. We have worked with researchers who successfully translated scientific discovery into human applications, and of not was through early partnerships with providers and community partners early in the design process. This helped in research approaches, clinically relevant models, and findings that more broadly applied to various populations. Also, having an expert who is well-versed in regulatory compliance, especially for protocol development regarding toxicity in humans as part of the research program early on in the experimental design, was a significant positive indicator for

rewarding scientific impact. When we worked with bench scientists to understand the fundamental principles of scientific impact and translational practices, researchers were able to significantly improve the bridge between curiosity-driven research and therapeutic innovation. Lastly, to see the fruits of thoughtful collaboration toward impact, researchers must be incentivized to do so through the promotional structure or the grant funding; otherwise, there is no protection or guarantee that their efforts to promote scientific impact for patients will be able to sustain long-term.

On Foundational Scientific Exploration

We want to name a risk directly that we also flag later in this response: translational or reproducibility-readiness indicators, applied indiscriminately, could pressure investigators doing foundational, high-risk basic research to justify translational relevance prematurely or artificially. We do not believe that is NIH's intent, and we would recommend the agency state so explicitly, for example, by scoping any new translational-fidelity indicator to mechanisms and award types where translational or clinical intent is already a stated goal (translational R01s, U-series awards, SBIR/STTR), and explicitly exempting R01s, R21s, and other awards designated as basic/exploratory from that scoring. Foundational science should continue to be evaluated on its own terms: novelty, technical rigor, and paradigm-shifting potential without being asked to retroactively demonstrate a translational pathway it was never designed to have. Protecting this distinction is, in our view, what allows translational indicators elsewhere in the portfolio to be adopted with less resistance from the research community, since investigators will trust that the new measures are targeted rather than universal.

On Entrepreneurship and Translation

We would encourage NIH to fund and recognize translational design capacity-building training, workshops, and advisory support that help research teams build translational thinking into a study before data collection begins as distinct from, and complementary to, downstream commercialization support (tech transfer, licensing, startup formation), which is where most current "translation" incentives are concentrated. The two interventions address different points of failure: commercialization support helps a mature discovery reach the market; design-stage translational support helps ensure the discovery was

shaped to be relevant to patients in the first place. Both are necessary, and we would caution against measuring "translation" activity in a way that only captures the former.

Our own consultancy's proprietary frameworks (including our Patient-Centered Discovery Framework™) were built specifically to address this earlier, design-stage gap, and we have applied them directly with academic research teams. Bench researchers with high preclinical expertise note a lack of education in clinical research when framing their preclinical work to patients or in the context of translating their work into humans. Our pilot interviews reinforce why this gap persists: researchers described independently discovering they needed regulatory, IP, or manufacturing expertise, then separately locating who could provide it, then separately finding a way to fund it three sequential, avoidable delays for every team that encountered them. Participants also directly named the absence of translational milestones in tenure and promotion review as a structural disincentive, independent of any individual investigator's intent. We would welcome the opportunity to share implementation data and lessons learned with NIH staff as the agency refines its approach to this indicator category.

On Public Impact

We recommend a composite indicator that pairs a measure of scientific rigor with a measure of translational fidelity: the degree to which a study's design anticipated the population, setting, and constraints of its eventual real-world application. Public impact, in our experience, is best predicted not at the point of publication but at the point of study design: research designed with translational fidelity from the outset is disproportionately more likely to eventually demonstrate real-world health, clinical, or economic impact, simply because it was built to be usable by the systems meant to adopt it.

We have developed and piloted one such composite measure internally (our Patient-Centered Translation Index™, or PCTI); we mention it here as evidence that operational, scorable versions of this kind of indicator are feasible today, not merely aspirational, and are glad to discuss our methodology and findings with NIH directly outside this public comment process.

Our pilot interviews also surfaced a public-impact risk that current metrics do not capture at all: participants across both research and clinical roles described cases where an efficacious, translationally

sound finding still failed to reach patients not for scientific reasons, but because an industry partner judged it insufficiently profitable to manufacture, or because contract and IP terms left the originating academic or clinical team without downstream control. A public-impact indicator built only around whether research could translate, without accounting for whether the surrounding funding and contracting structures allow it to actually reach patients, will systematically overstate impact potential relative to realized impact. We would encourage NIH to consider this distinction explicitly in how "public impact" is defined and measured.

We would also note that SciClin occupies a distinct position relative to other organizations NIH is likely to hear from in response to RFI CTSA hubs, AAMC, and institutional tech-transfer offices, among others. Those entities are largely oriented toward institutional infrastructure and downstream commercialization. Our focus is narrower and earlier: design-stage translational thinking, delivered directly to research teams through workshops, translational audits, and applied consulting, independent of any single institution's infrastructure. We see this as complementary to, not competitive with, the institutional layer, and would welcome the chance to discuss how independent, cross-institutional translational support of this kind could factor into NIH's broader strategy.

Applied Evidence and a Proposed Pilot

The pilot dataset referenced throughout this response, with basic researchers and clinicians across multiple disease areas, thematically coded into seven primary barrier categories with ten linked recommendations, was produced independent of NIH funding, using our own methodology. We reference it here not as a completed study but as a proof of concept: a design-stage translational-fidelity indicator, and a reproducibility-readiness rubric of the kind described above, can be piloted quickly and inexpensively using qualitative and document-review methods already validated in this dataset.

We would welcome the opportunity to propose a small-scale pilot to NIH's Office of Science Policy directly, for example, applying a reproducibility-readiness rubric to a defined set of active R01 or translational-mechanism protocols across a handful of partner institutions, and reporting de-identified, aggregate findings back to OSP within a defined window as a concrete next step ahead of any broader

adoption decision. Alternatively, we are positioned to provide in-person workshops and share our online educational resources centered on translational readiness.

Feasibility and Unintended Consequences

Any new indicator in this space should be piloted with investigator and reviewer burden explicitly in mind; an additional checklist or index item is only useful if it takes minutes, not hours, to complete, and if it is scored by someone other than the original investigator to avoid self-assessment bias. Where possible, new indicators should be layered onto existing reporting mechanisms (RPPRs, progress reports, renewal applications) rather than introduced as standalone instruments requiring new infrastructure or review burden.

We would also flag two risks worth naming directly:

1. **Without care, "translation-readiness" metrics could inadvertently penalize foundational, high-risk basic research that has no near-term clinical application by design.**
    a. We would recommend any composite scoring approach apply only where translational intent is already claimed by the investigator (e.g., in R01s or U-series translational mechanisms), rather than universally across all NIH-funded science. If to be applied to all science for impact, the agency would best benefit from requiring more discussion on how the experimental design was derived from clinical observation or may lead to the betterment of patients, not as a background premise of study but as the backbone for experimental design accounting for sex differences, biological diversity, disease burden, technical replication, etc.
2. **Metrics tied to funding decisions create strong incentives to game documentation rather than genuinely improve design.**
    a. We would recommend the NIH pilot any new indicator as a reported and tracked metric before tying it directly to funding or renewal decisions, so that its validity can be assessed against downstream reproducibility and impact outcomes first.

Josly Pierre-Louis Odoom, PhD MS
Eva Davis, PhD
Annie Hinson, PhD
Christie Pierre-Louis, BS

Disclosure of AI use

In preparing this response, AI-based tools (Anthropic's Claude) were used to support two aspects of our work: (1) drafting and editorial assistance, including organizing and refining language for clarity and conciseness, and (2) analytical support in coding and synthesizing qualitative data from our pilot thematic analysis of patient/clinical integration barriers in translational science. All AI-assisted outputs were reviewed, verified, and edited by the author(s), who take full responsibility for the accuracy, interpretation, and conclusions presented in this response. The underlying research design. Interview conduct, thematic framework development, and substantive recommendations reflect the professional judgment and expertise of the author(s); AI tools served as a productivity aid rather than a substitute for that judgment.

SciClin LLC would welcome further engagement with NIH's Office of Science Policy on any of the above, including a walkthrough of our applied frameworks and full pilot dataset, and stands ready to scope the reproducibility-readiness pilot proposed above, outside of this public comment record.



Josly Pierre-Louis Odoom, PhD
CEO & Founder SciClin LLC-----
sciclinlabs.com
joslyplo@sciclinlabs.com

SciClin
Patients to Science. Science to Patients.

288. Angela Sy

I am responding: On behalf of myself

Name: Angela Sy

Name of Organization: University of Hawaii

Type of Organization: Academic Institution

Role: Investigator/Researcher

Date Submitted: 8/20/2026

1. Rigor and Reproducibility: N/A

2. Data, Software, and Model Sharing: N/A

3. Training and Mentorship: • Indicators of network, organizational and investigator capacity building, such as
o Development of investigators and clinician fellows
o Improved access to scientific research infrastructure
o Subsequent submission of HRPs after completing the RPP'
o Subsequent submission of peer-reviewed grant proposals (for example, K-awards, RPGs)
o Subsequent external independent funding obtained by participants in the program
o Subsequent implementation of research findings across the organization, partners, or the state to inform evidence-based practices and policy.

4. Collaboration: • Indicators of research or research network development and collaboration, such as:
o Number and characteristics of partner organizations
o Resources shared among partner organizations
o Research collaborations established across lead organizations and with other IDeA organizations (CTR-D) or network organizations and with non-network IDeA organizations (CTR-N)
o Engagement with PBRNs or community clinics to expand resources, expertise, and funding opportunities
o Number of HRP submissions by PBRN or community clinic investigators

5. Entrepreneurship and Translation: N/A

6. Foundational Scientific Exploration: • Indicators of scientific productivity and impact among program or network organizations, such as:
o Peer-reviewed research publications and citations
o Conference presentations
o A number of clinical trials and complex observational studies conducted
o Utilization of the I-CRRC:
- Tracking the number of Clinical Trial participation and the number of participants in the Clinical Research Coordinator Certificate(s)

- Real-World Data-enabled activities attributable to IDeA Clinical Research Resource Center (I-CRRC) engagement, such as relevant grant submissions, funded awards, and dissemination efforts

7. Public Impact: Funding for 5 year NIH grants, especially in the first cycle are not able to evaluate broader public impacts. Furthermore broader impacts are the contribution of many initiatives. This level of impact evaluation requires a separate initiative that specifically assesses the long term impacts, which only result from multiple initiatives working together. NIH could have NOFO specifically on impact evaluation or some type of supplemental funding for existing projects to conduct this resource intensive evaluation.;

8. Other Areas: NIH funded research centers—CTSA, U54—have been considering and have even identified common metrics addressing many of the areas in this RFI. More importantly in addition to identifying measures, a concerted planned effort is required to ensure that measures are feasible and have utility (https://www.eval.org/About/Competencies-Standards/Program-Evaluation-Standards). Some NIH funded research center consortia have discussed and implemented common metrics in the evaluation of it research impacts. The administration and implementation of common metrics requires a NIH wide strategy so that measures that have been identified (as a result of this and other efforts) have a clear purpose on how it will be used. For example CTSA outlines its purpose for common metrics: "The Common Metrics Initiative was designed to assess and optimize the CTSA Program's overall impact on the nation's health. Establishing a set of standard evaluation measures across CTSA Program institutions helps to focus program activities, streamline data collection, and demonstrate measurable progress toward program goals, including improvements in research translation and workforce development." (https://ncats.nih.gov/research/research-activities/ctsa/projects-and-initiatives). The NIMHD U54 Research Centers in Minority Institutions recommends identifying common metrics that allow for valid consortium-wide evaluations and that these metrics are used to facilitate research impacts for the community (https://www.mdpi.com/1660-4601/17/22/8373). The NHGRI U54 Centers for Genomic Research prospectively developed a Consortium evaluation plan based on common metrics across the funded centers to facilitate the ongoing evaluation, reporting, and adjustment of research centers to increase the genomics workforce
(https://www.sciencedirect.com/science/article/pii/S0002929725001016).
Efforts within the NIH community to identify common metrics for the research enterprise exist, and this initiative including the RFI process should work in alignment so as not to "reinvent the wheel." Common metrics sound like a good idea; its implementation and strategies to make such initiative useful and feasible require careful and thoughtful planning (references).
The next step is to have a clear agenda on the purpose of the common metrics, how it will be used, and the extent it will be collected, and how (by NIH). For example to what extent will the common metrics be aggregated, and if so, how will the aggregated results be used?
I would like to participate in future efforts regarding measuring scientific impact. Thank you for the opportunity to comment Angela Sy, DrPH sya@hawaii.edu

File Upload: https://osp.od.nih.gov/wp-content/uploads/ninja-forms/77/gansky-et-al-2025-common-data-element-cde-success-from-community-engagement-not-mandates-if-you-build-it-data-Copy.pdf

Description: See additional article in common metrics (common data elements) coordination and implementation: Kim H, Liang S, Lanier K, Helman S, Ameli N, Banava S, Cheng NF, Boyce J, Hswen Y, Oldenburg CE, Rhoads KF, Charlebois ED, Gansky SA, Brown Iii W. Bidirectional data harmonization across

the multiple chronic disease disparities research consortium: A pipeline description of opportunities and challenges. J Biomed Inform. 2026 May;177:105004. doi: 10.1016/j.jbi.2026.105004. Epub 2026 Feb 28. PMID: 41771392; PMCID: PMC13101454., Gansky SA, Cheng AC, Mendhe D, Chavez A, Yeh HC, Lanier K, Brown W 3rd, Hersch DE, Huang ES, Enders FT, Phelan SM. Common Data Element (CDE) Success from Community-Engagement Not Mandates: If You Build It [Data Warehouse], Will They [Participants] Come? Am J Public Health. 2025 Jul;115(S2):S110-S115. doi: 10.2105/AJPH.2025.308114. PMID: 40561407; PMCID: PMC12199679. Sancheznieto F, Pribbenow CM, Pfund C. Tools for coordinating measurement use in science education and training. J Microbiol Biol Educ. 2025 Aug 21;26(2):e0006025. doi: 10.1128/jmbe.00060-25. Epub 2025 May 7. PMID: 40331603; PMCID: PMC12369332.

Common Data Element (CDE) Success from Community-Engagement Not Mandates: If You Build It [Data Warehouse], Will They [Participants] Come?

Stuart A. Gansky, DrPH, Alex C. Cheng, PhD, Dinesh Mendhe, MS, Aryanna Chavez, MPH, Hsin-Chieh Yeh, PhD, Kathy Lanier, MPH, William Brown III, DrPH, PhD , Derek E. Hersch, MPH, Elbert S. Huang, MD, Felicity T. Enders, PhD, MPH, and Sean M. Phelan, PhD, MPH

ABOUT THE AUTHORS

Stuart A. Gansky, Kathy Lanier, and William Brown III are with the Research Coordinating Center to Reduce Disparities in Multiple Chronic Diseases, University of California, San Francisco, CA. Alex C. Cheng is with the Southeast Collaborative for Innovative and Equitable Solutions to Chronic Disease Disparities, Vanderbilt University Medical Center, Nashville, TN. Dinesh Mendhe is with the Rutgers–New York University Center for Asian Health Promotion and Equity, Institute for Health, Health Care Policy and Aging Research, Rutgers University, New Brunswick, NJ. Aryanna Chavez is with the University of California, Los Angeles–University of California, Irvine Center for Eliminating Cardio-Metabolic Disparities in Multi-Ethnic Populations, University of California - Irvine, Irvine, CA. Hsin-Chieh Yeh is with the Mid-Atlantic Center for Cardiometabolic Health Equity, Johns Hopkins University, Baltimore, MD. Derek E. Hersch, Felicity T. Enders, and Sean M. Phelan are with the Center for Chronic Disease Reduction and Equity Promotion Across Minnesota; DEH is at the University of Minnesota, Minneapolis, MN; FTE and SMP are with Mayo Clinic, Rochester, MN. Elbert S. Huang is with the Chicago Chronic Condition Equity Network, University of Chicago, Chicago, IL.

In the 1989 film *Field of Dreams*, a voice states, "If you build it, they will come." But if you build a common data element data warehouse, will they (community research participants) come?

Common data elements (CDEs)—having the same wording and response options for questionnaire items across studies—are important to advance health equity research because they provide standardized data that can be systematically collected, shared, and analyzed across different studies. Adopting uniform measures as CDEs enables researchers to pool data from diverse populations, thereby enhancing power and generalizability of research findings. This is particularly crucial for health equity research, which aims to identify and address disparities in health outcomes among different, often underrepresented demographic groups. CDEs ensure that key social determinants, or drivers, of health (SDOH) and comorbidities are consistently captured, facilitating cross-study comparisons and meta-analyses that are critical for understanding health inequities on a larger scale. Most publications report using researchers and experts for developing CDEs.[1,6]

The National Institutes of Health (NIH) and other funding agencies have recognized CDEs' importance, have implemented policies to promote their use, and have required their use in some recently established cross-cutting initiatives, specifically to foster synergistic evidence building supporting health equity. The NIH Scientific Data Council established a working group to recommend "a minimum core set of CDEs" for "NIH clinical research/trials" and sought public comment (https://bit.ly/43bhUpl). A systematic scoping review of 22 CDEs of SDOH development studies found that 73% included only research experts.[7] Such top-down approaches can lead to suboptimal results, but community-based approaches can be more successful.[8] We describe the process used to develop de-identified CDEs in the community-engaged research of the Health Equity Action Network (HEAN), discuss lessons learned, and make recommendations for future consortia.

However, developing and implementing CDEs require careful consideration to avoid potential drawbacks that could undermine their effectiveness in promoting health equity. Without active community involvement, particularly from marginalized and underrepresented populations, CDEs may lack cultural relevance or fail to address community-specific needs. Reduced participation, incomplete responses, and distrust in the research process can result. For example, inadequately

AJPH Supplement 2, 2025, Vol. 115, No. S2

tested or culturally insensitive items may alienate potential participants, compromising data quality and the overall research validity. To mitigate these risks, study investigators must collaborate with community partners early in CDE development. Thus, researchers can ensure that measures are contextually appropriate and ethically sound, increasing their acceptance and likelihood of meaningful engagement. Moreover, community collaboration enhances research credibility, fosters trust, and strengthens findings' applicability to populations with the most health disparities. This participatory approach not only improves data relevance and quality but also helps fulfill the broader goal of advancing health equity through inclusive and community-centered research practices.

HEALTH EQUITY ACTION NETWORK'S COMMON DATA ELEMENTS PROCESS

NIH's National Institute of Minority Health and Health Disparities (NIMHD) issued joint requests for applications as funding opportunity announcements for both the P50 regional research centers and the research coordinating center (RCC) for a new initiative. NIMHD required P50s to work closely with community partners and for the RCC to collect the CDEs of SDOH from P50 main research components in a de-identified data warehouse. However, NIMHD did not detail which CDEs would be mandatory. The NIH strongly encouraged using existing CDE registries, such as the 52-item PhenX Toolkit SDOH Collections (core, individual, and structural along with prerequisites) to "facilitate downstream harmonization and meta-analysis"[9(p258)] and "[foster] a new era of cooperation and collaboration and facilitating cross-study, transdisciplinary, and translational research."[9(p259),10] The RCC proposed including comorbidities and quality of life measures because the initiative focused on multiple chronic diseases with various health outcome measures.

The RCC evaluated the CDEs of SDOH using screening tools tabulated by the Social Interventions Research and Evaluation Network,[11,12] including the 26-item Centers for Medicaid and Medicare Services Accountable Health Communities Health-Related Social Needs Screener,[13] the 83-item All of Us SDOH module,[14,15] and the 32-item Kaiser Permanente Your Current Life Situation questionnaire.[16] The RCC evaluated resources (e.g., screeners, toolkits, and modules) using clinical and translational social and environmental nomenclature,[17] seeking to include at least one item from each domain, while keeping the overall list parsimonious. Using that nomenclature, 16 ectome (external or outside the body) domains are composed of five environmental (aerome: air, hydrome: water, terrome: soil, nutriome: consumed food, and biome: flora, fauna, and microbiota), seven social (philome: family and social support network, allolome: people not in the philome, legome: educational achievement and opportunities, econome: economic circumstances, ethnome: culture including race/ethnicity, actome: physical activity, and home: living environment and neighborhood), and four health care (iatrome: personal health care experience, therapome: therapeutic care, preventome: preventive care, and pharmacome: pharmaceuticals) factors. Ectome domains and CDEs of the HEAN items are depicted mapped onto the NIMHD research framework in Figure 1).[18]

Consortium options for CDEs were to adopt an existing resource (screener, toolkit, or module), entirely or partially, to map and collapse items from P50 projects' planned instruments, or to adapt existing resources and harmonize across P50 projects. Adoption had the advantage of selecting reliable and validated instruments, but it lacked HEAN community partner engagement. Map and collapse had the advantage of using measures P50s adopted with extensive community engagement, but initial measure reviews showed items mappable in limited domains and only after extensive collapsing, resulting in significant loss of information. Thus, we selected the last—adapt and harmonize—as a pragmatic middle ground balancing the pros and cons by parsimoniously including all domains and harmonizing across projects, while also including dialogue with community partners.

To facilitate harmonizing, the RCC included plans to use natural language processing but pivoted to large language models such as ChatGPT (OpenAI, San Francisco, CA) to increase efficiency. Certified translators translated the CDEs of the HEAN into Spanish and Vietnamese, as multiple P50s planned to conduct research with communities who spoke those languages. P50s translated CDEs into other center-specific languages (e.g. Chinese, Korean, and Somali).

Centers, research projects, and some community partners (including those from faith-based organizations, health care systems, community-based organizations, educational institutions, and community gathering places) evaluated the proposed 50-item CDE list of SDOH. There was notable friction between the P50s and the RCC regarding the scope and plans for the CDEs of SDOH; specifically, objections related to

a

		Levels of influence			
		Individual	**Interpersonal**	**Community**	**Societal**
Domains of Influence (Over the Life Course)	**Biological**	Sex at Birth Disability Comorbidities	Caregiver-Child Interaction Family Microbiome	Community Illness Exposure Herd Immunity	Sanitation Immunization Pathogen Exposure
	Behavioral	Pharmacome	Ethonome	Community Functioning	Policies and Laws
	Physical/Built Environment	Biome	Home	Aerome Hydrome Terrome Nutriome	Societal Structure
	Sociocultural Environment	Legome Econome	Philome Allolome	Actome	Social Norms Societal Structural Discrimination
	Health Care System	Iatome	Discrimination in Health Care	Therapome Preventome	Quality of Care Health Care Policies
Health Outcomes		**Individual Health**	**Family/Organizational Health**	**Community Health**	**Population Health**

b

		Levels of influence			
		Individual	**Interpersonal**	**Community**	**Societal**
Domains of Influence (Over the Life Course)	**Biological**	Sex at Birth Disability Comorbidities	Caregiver-Child Interaction Family Microbiome	Community Illness Exposure Herd Immunity	Sanitation Immunization Pathogen Exposure
	Behavioral	Tobacco Alcohol	Household Language	Community Functioning	Policies and Laws
	Physical/Built Environment	Lack of Transportation	Household Structure Household Resources	Food Access Environmental Exposures	Societal Structure
	Sociocultural Environment	Race/Ethnicity Gender Education Income Limited English	Relationship Status Social Isolation Social Support	Physical Activity	Social Norms Societal Structural Discrimination
	Health Care System	Insurance Utilization Health Literacy	Discrimination in Health Care	Rx Availability	Quality of Care Health Care Policies
Health Outcomes		Self-Rated Health	**Family/Organizational Health**	**Community Health**	**Population Health**

FIGURE 1— Domains and Items Mapped Onto the National Institute of Minority Health and Health Disparities (NIMHD) Research Framework (a) Clinical and Translational Nomenclature Domains, and (b) Health Equity Action Network (HEAN) Common Data Element Items

CDEs considered not central to individual study hypotheses, changes needed in pragmatic designs using extant data, some specific CDEs on safety-related or sensitive topics, increased participant burden, and lack of planned cross-center analyses. The RCC suggested some cross-center analyses; however, the requests for applications required creating a data warehouse. Some consortium members noted that if no local or study resources were available to address a SDOH need, then the screening item was potentially unethical. Negative affect and depressive symptom

assessment was debatable depending on local resources available. Others note that CDEs specific to adults should be excluded for studies focused on children. During discussions, other consortium members requested that data that could be self-incriminating or have legal implications be excluded.

Thus, the RCC removed the following items based on consortium community feedback: country of origin, recreational or illegal drug use, and physical or financial abuse. Also based on consortium input, a disability item was added. Qualitative research projects were strongly encouraged to add the CDEs of SDOH to permit mixed methods research. Two projects that did not obtain informed consent of any individuals—one implementation science study with process measures and one evaluation study using only aggregate data from an honest broker—were exempted from collecting any CDEs. Three studies focused on children were granted exceptions for specific CDEs focused on adults.

Census-based geolocation enables more precise and equitable SDOH and health outcome analyses compared with zip codes or zip code tabulation areas. Unlike zip codes, which are designed primarily for mail delivery and often cover large, heterogeneous areas, census-based geolocation offers finer granularity with defining specific geographic boundaries (i.e., census tracts and blocks). Using zip codes (or zip code tabulation areas) in health research presents several limitations. Zip code boundaries can encompass highly diverse populations, including both high and low levels in important variables such as socioeconomic status, health care access, and environmental factors. Relying on zip code as geolocation may result in misleading conclusions, such as masked significant health disparities. By contrast, census-based geolocation enables researchers to better pinpoint specific areas of need and tailor interventions to better address inequities, potentially enhancing targeted, data-driven health equity initiatives.

To reduce response burden on individual participants, we are obtaining neighborhood-level data based on each participant's geocoded residential address. However, because the pooled consortium CDE data will be completely de-identified, we will coarsen linked neighborhood characteristics and retain them for the data warehouse, but we will remove actual geolocation information (e.g., Census tract and address including zip code) before transferring to the RCC.

The final 42 CDEs of SDOH—plus one of two comorbidity CDE versions and one quality of life CDE, along with 12 high-value optional CDEs of SDOH, summarized in a codebook and REDCap (Research Electronic Data Capture, Vanderbilt University, Nashville, TN)[19] data dictionary, along with a data use agreement template—are publicly available on the consortium Web site (https://health-equity-action.org).

PROGRESS AND ACHIEVEMENTS TO DATE

Of the 28 nonexempt initial R01-level projects, 100% are collecting CDEs. Across those projects, the mean percentage of required CDEs being collected is 98%; 27 of 28 nonexempt projects are collecting at least 90% of the required CDEs. All codebooks for the 28 nonexempt projects have been received. Most (85%) data use agreements from institutions requiring them to provide data to the University of California, San Francisco as the recipient institution have been fully executed, and four are in progress. To date, 20 R01-level projects have transferred CDE data to the RCC.

LESSONS LEARNED AND RECOMMENDATIONS FOR THE FUTURE

Although developing and implementing CDEs in HEAN was successful, we learned many lessons throughout the process, which we offer here as suggestions for similar future endeavors. Initially, start the CDE selection process as early as possible in the project life cycle. One key element of successful CDEs of the HEAN collection was the requests for applications that included the expectation and requirement for CDEs, which allowed individual center and study primary investigators to plan for CDE collection from the beginning. Beginning the CDE selection process before funding participating centers or projects has three important benefits. First, providing candidate CDE measures to study primary investigators at the outset of their studies enables them to conduct meaningful engagement with the community members and their partner organizations to review items, discuss rationale for CDEs, and provide recommendations and feedback to refine and improve the process. For example, community partners may identify items that create discomfort or difficulty for people in their communities to answer and suggest alternatives that might lead to better quality data. Involving communities early in the process also shows respect for their active participation in study design, consistent with the values that underlie community-engaged research. Second, early study primary investigator and team

involvement in CDE development helps primary investigators develop their own survey instruments, as they will know which data elements will be required and can avoid redesigning surveys to accommodate new measures. Third, beginning the process early helps ensure that study primary investigators have time to discuss and negotiate exemptions that might be necessary to accommodate study designs that make CDE collection difficult or impossible.

Another recommendation for successful CDE implementation is to clearly and frequently communicate with center and study primary investigators about the importance of CDE collection and planned future data uses. Communicating, as specifically as possible, the ways data will be reused (e.g., a cross-center meta-analysis of a common intervention type) provides two important benefits. First, understanding the effort's purpose is likely to improve buy-in and effort from researchers if the planned outcomes are consistent with their goals and values. Second, communicating plans allows primary investigators and other study leaders, including community partners, to contribute their own ideas for novel data reuse, allowing more opportunities for engagement in the work of achieving those goals, benefiting the end results, and encouraging shared responsibility for efforts and benefits (e.g., authorship).

PRINCIPLES FOR THE FUTURE

Tensions between bottom-up grassroots and top-down funder-driven CDE development approaches present a critical challenge in community-engaged data collection. Grassroots efforts, based on the specific needs and contexts of local communities, ensure relevance and community empowerment but often lack the standardization and comprehensiveness necessary for broader data comparison. Conversely, top-down approaches prioritize uniformity and scalability but may impose external priorities that do not fully align with each unique community's needs. To resolve this, a balanced approach is recommended—one that is community centered yet allows local adaptation to extract meaningful SDOH information in standardized frameworks. Achieving this balance requires working collaboratively with community partners and advisory boards from project initiation and CDE development inception and is characterized by genuine dialogue and cocreation. This approach ultimately would combine the strengths of both methods and thus allow scientific research questions to be investigated rigorously while being deeply responsive to community contexts.

Based on our experiences with multiple efforts to develop CDEs in previous consortia across multiple awards funded at different institutions,[20,21] we recommend that sponsors build in three or four months in the first year for CDE development and pretesting across projects in the consortium. No matter how much a set of CDEs has been developed and tested with other community partners, each implementation with new communities or new community members needs to include ample time for their involvement as equals to modify proposed CDEs to foster developing and maintaining authentic partnerships in community-engaged research. Thus, funder-mandated, required CDEs developed without engaging the specific communities in these studies are likely to reduce enrollment as well as data completeness, validity, or both. Skipping community engagement with expert-developed CDEs may seem expeditious, but it almost always involves delays from backtracking to involve communities after the fact, which is less ideal and can take just as long or longer. AJPH

CORRESPONDENCE

Correspondence should be sent to Stuart A. Gansky, UCSF, Box #1361, San Francisco, CA 94143 (e-mail: stuart.gansky@ucsf.edu). Reprints can be ordered at https://www.ajph.org by clicking the "Reprints" link.

PUBLICATION INFORMATION

Full Citation: Gansky SA, Cheng AC, Mendhe D, et al. Common data element (CDE) success from community-engagement not mandates: if you build it [data warehouse], will they [participants] come? *Am J Public Health*. 2025;115(S2):S110–S115.

Acceptance Date: March 10, 2025.

DOI: https://doi.org/10.2105/AJPH.2025.308114

ORCID iDs:

Stuart A. Gansky https://orcid.org/0000-0003-3147-0884

Alex C. Cheng https://orcid.org/0000-0002-1787-691X

Dinesh Mendhe https://orcid.org/0000-0003-1158-3259

Aryanna Chavez https://orcid.org/0009-0003-7127-2273

Hsin-Chieh Yeh https://orcid.org/0000-0002-5738-0652

William Brown III https://orcid.org/0000-0003-4322-4859

Derek E. Hersch https://orcid.org/0000-0003-4867-1156

Elbert S. Huang https://orcid.org/0000-0002-4628-2061

Felicity T. Enders https://orcid.org/0000-0002-3261-5400

Sean M. Phelan https://orcid.org/0000-0003-2091-6297

CONTRIBUTORS

S. A. Gansky conceptualized the study and developed the methodology. S. A. Gansky, A. C. Cheng, D. Mendhe, A. Chavez, H. C. Yeh, and D. E. Hersch wrote the original draft of the article. S. A. Gansky, A. C. Cheng, D. Mendhe, A. Chavez, H.-C. Yeh, D. E. Hersch, E. S. Huang, F. T. Enders, and S. M. Phelan contributed to writing and editing the article. K. Lanier and W. Brown III contributed to developing the methodology. F. T. Enders visualized the

figure. All authors reviewed and approved the final version of the article.

ACKNOWLEDGMENTS

Research reported in this article was supported by the National Institute on Minority Health and Health Disparities, National Institutes of Health (NIH; awards U24MD017250, P50MD017319, P50MD017338, P50MD017341, P50MD017342, P50MD017344, P50MD017347, P50MD017348, P50MD017349, P50MD017351, P50MD017356, and P50MD017366).

The authors are part of the Health Equity Action Network's Common Data Element and Data Harmonization Working Group. Other members include Juan Espinoza, Payal Shah, Katherine Crew, Jeanine Genkinger, Rita Kukafka, John Jackson, Lisa Lanza, Aparna Kalbag, James Selig, Ruofei Du, Dina Jones, April Agne, April Carson, Emily Mire, Tapan Mehta, Ivan Azarov, Alejandra Casillas, Monica Ayala-Rivera, Krystal Griffith, Natalie Robles, Megan Huisingh-Scheetz, Brad Appelhans, Tongtan Chantarat, Julia Stantoznik, Isaac Schlotterbeck, Robert Brook, Michael Twiner, Jackie Parker, and Brian Reed.

Note. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to disclose.

REFERENCES

1. Grinnon ST, Miller K, Marler JR, et al. National Institute of Neurological Disorders and Stroke Common Data Element Project—approach and methods. *Clin Trials.* 2012;9(3):322–329. https://doi.org/10.1177/1740774512438980
2. Saver JL, Klerman EB, Buchan AM; Leducq Consortium International pour la Recherche Circadienne sur l'AVC Effects in Stroke, et al. Consensus recommendations for standardized data elements, scales, and time segmentations in studies of human circadian/diurnal biology and stroke. *Stroke.* 2023;54(7):1943–1949. https://doi.org/10.1161/STROKEAHA.122.041394
3. Rosenfield K, Bowers TR, Barnett CF; Pulmonary Embolism Research Collaborative Attendees, et al. Standardized data elements for patients with acute pulmonary embolism: a consensus report from the Pulmonary Embolism Research Collaborative. *Circulation.* 2024;150(14):1140–1150. https://doi.org/10.1161/CIRCULATIONAHA.124.067482
4. Adams MCB, Hassett AL, Clauw DJ, Hurley RW. The NIH pain common data elements (CDE): a great start but a long way to the finish line. *Pain Med.* 2025;26(3):146–155. https://doi.org/10.1093/pm/pnae110
5. Kirkpatrick L, Rao CK, Beatty CW, et al. Common data elements regarding social determinants of health in pediatric epilepsy research: a concept mapping study. *Pediatr Neurol.* 2025;165:87–95. https://doi.org/10.1016/j.pediatrneurol.2025.01.025
6. Murphy DJ, Anderson W, Heavner SH, et al. Development of a core critical care data dictionary with common data elements to characterize critical illness and injuries using a modified Delphi method. *Crit Care Med.* [Epub ahead of print February 21, 2025]. https://doi.org/10.1097/CCM.0000000000006595
7. Levites Strekalova YA, Nelson JD, Weber HM, Wang X, Midence SM. Application of the Delphi method to the development of common data elements for social drivers of health: a systematic scoping review. *Transl Behav Med.* 2024;14(7):426–433. https://doi.org/10.1093/tbm/ibae020
8. Behar-Horenstein LS, Suiter S, Snyder F, Laurila K. Consensus building to inform common evaluation metrics for the Comprehensive Partnerships to Advance Cancer Health Equity (CPACPHE) Program. *J Cancer Educ.* 2023;38(1):231–239. https://doi.org/10.1007/s13187-021-02103-1
9. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX Toolkit: get the most from your measures. *Am J Epidemiol.* 2011;174(3):253–260. https://doi.org/10.1093/aje/kwr193
10. PhenX Toolkit. Social determinants of health collections. December 18, 2024. Available at: https://www.phenxtoolkit.org/collections/view/6. Accessed April 25, 2025.
11. Henrikson NB, Blasi PR, Dorsey CN, et al. Psychometric and pragmatic properties of social risk screening tools: a systematic review. *Am J Prev Med.* 2019;57(6 suppl 1):S13–S24. https://doi.org/10.1016/j.amepre.2019.07.012
12. Cartier Y, Fichtenberg F, Gottlieb L. Social needs screening tool comparison table. 2019. Available at: https://sirenetwork.ucsf.edu/tools-resources/resources/screening-tools-comparison. Accessed December 14, 2021.
13. Accountable Health Communities. A guide to using the Accountable Health Communities Health-Related Social Needs Screening Tool: promising practices and key insights. December 2023. Available at: https://www.cms.gov/priorities/innovation/media/document/ahcm-screeningtool-companion. Accessed September 23 2024.
14. All of Us Research Program. Social Determinants of Health (SDOH). November 1, 2021. Available at: https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/surveys/SDOH_Survey_English.pdf. Accessed December 14, 2021.
15. Tesfaye S, Cronin RM, Lopez-Class M, et al. Measuring social determinants of health in the All of Us Research Program. *Sci Rep.* 2024;14(1):8815. https://doi.org/10.1038/s41598-024-57410-6
16. LaForge K, Gold R, Cottrell E, et al. How 6 organizations developed tools and processes for social determinants of health screening in primary care: an overview. *J Ambul Care Manage.* 2018;41(1):2–14. https://doi.org/10.1097/JAC.0000000000000221
17. Davis MM, Shanley TP. The missing -omes: proposing social and environmental nomenclature in precision medicine. *Clin Transl Sci.* 2017;10(2):64–66. https://doi.org/10.1111/cts.12453
18. National Institute on Minority Health and Health Disparities. NIMHD research framework. 2017. Available at: https://nimhd.nih.gov/researchFramework. Accessed April 14, 2021.
19. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42(2):377–381. https://doi.org/10.1016/j.jbi.2008.08.010
20. Albino J, Tiwari T, Gansky SA; Early Childhood Caries Collaborating Centers, et al. The basic research factors questionnaire for studying early childhood caries. *BMC Oral Health*. 2017;17(1):83. https://doi.org/10.1186/s12903-017-0374-5
21. Ahmadi N, Zoch M, Kelbert P, et al. Methods used in the development of common data models for health data: scoping review. *JMIR Med Inform.* 2023;11:e45116. https://doi.org/10.2196/45116