

Compiled Public Comments on a DRAFT
NIH Policy for Data Management and
Sharing and Supplemental DRAFT
Guidance

Guide Notice Number: NOT-OD-20-013

November 06, 2019 – January 10, 2020

Table of Contents

1. [Stacy Stuart, University of Arkansas](#)
2. [Joel Voss, Northwestern University](#)
3. [Andrew Vickers, Memorial Sloan Kettering Cancer Center](#)
4. [Anonymous](#)
5. [Jesse Forrest, Immune Tolerance Network](#)
6. [George McNamara, Johns Hopkins University](#)
7. [Alik Widge, University of Minnesota](#)
8. [Nancy Janitz](#)
9. [elly lee, gachon university](#)
10. [Emily Scott, University of Michigan](#)
11. [Samarendra Mohanty, Nanoscope Technologies, LLC](#)
12. [Mona Hicks, One Mind](#)
13. [Karen Sherman, Kaiser Permanente Washington Health Research Institute](#)
14. [Melinda Marino, Arizona State University](#)
15. [Deirdre Joy, NIAID/NIH](#)
16. [Andreas Mueller, Columbia University](#)
17. [Michele Diaz, Pennsylvania State University](#)
18. [Ellen M. Wijsman, University of Wahsington](#)
19. [Richard A Kahn, Emory University](#)
20. [Brian C Trainor, UC Davis](#)
21. [David Cormode, University of Pennsylvania](#)
22. [Evan Mayo-Wilson](#)
23. [Mara Mather, USC](#)
24. [Hannah V. Carey, PhD, FASEB President, Federation of American Societies for Experimental Biology \(FASEB\)](#)
25. [Betsy L Humphreys](#)
26. [Rochel Gelman, Rutgers, Center for Cognitive Science and Psychology](#)
27. [Christine Morrison, CDC](#)
28. [Rhoda Au, Boston University, School of Medicine/Public Health](#)
29. [Keri Hornbuckle, University of Iowa](#)
30. [Hunter Moseley, University of Kentucky](#)
31. [Mary Janevic, University of Michigan](#)
32. [Julie Lima, Vincent Mor, Faye Dvorchak, Roeie Gutman, Brown University](#)
33. [Toni Harbaugh, NCI/Frederick National Laboratory for Cancer Research](#)
34. [Jennifer DeBerg, University of Iowa](#)
35. [Everett Carpenter, HHS-NIH-NCI-DCCPS](#)
36. [Ho Jung Yoo, University of California San Diego](#)
37. [Lucia Peixoto, Washington State University](#)
38. [Fred Oswald, Rice University](#)

39. [Christian Murray, Murtek Systems](#)
40. [Pam Dixon, World Privacy Forum](#)
41. [Michael Hoffman, Princess Margaret Cancer Centre](#)
42. [Susanna-Assunta Sansone, FAIRsharing](#)
43. [Bruce Stillman, Cold Spring Harbor Laboratory](#)
44. [Data Services Team, NYU Health Sciences Library](#)
45. [Michael McDonell, Washington State University](#)
46. [Casey Greene, University of Pennsylvania](#)
47. [Joshua Batson, Chan Zuckerberg Biohub](#)
48. [Charles Warden, City of Hope National Medical Center](#)
49. [Anna Greene, Alex's Lemonade Stand Foundation](#)
50. [Emma Grace, The Chicago School of Professional Psychology - Washington, DC](#)
51. [Boris Barbour, The PubPeer Foundation](#)
52. [Chris Brown, Atlas Research](#)
53. [Stephan Bour, Digital Infuzion, Inc.](#)
54. [Greg Raschke, NC State University Libraries](#)
55. [Stephanie Fox-Rawlings, National Center for Health Research](#)
56. [Scott Kahn, Helmsley Charitable Trust](#)
57. [Douglas P. Kiel, MD, MPH, Marcus Institute for Aging Research, Hebrew SeniorLife](#)
58. [REBECCA LI, Vivli](#)
59. [Steve Pieper, isomics, Inc.](#)
60. [Sean McGurn, Triple Point Security, NIH Extramural Data Security Team \(EDST\)](#)
61. [Chris Bourg, Massachusetts Institute of Technology - MIT Libraries](#)
62. [American Society of Bone and Mineral Research, American Society of Bone and Mineral Research](#)
63. [Harry W. Orf, PhD, Massachusetts General Hospital](#)
64. [Kerry Ressler, MD, PhD, McLean Hospital](#)
65. [Paul Anderson, BWH](#)
66. [Ravi Thadhani, M.D., MPH, Partners HealthCare](#)
67. [Lauren Gross, The American Association of Immunologists](#)
68. [Carol Pulver, Frontier Science Foundation](#)
69. [Meghan Faherty, Jean Mayer USDA HNRCA at Tufts University](#)
70. [Meriel Patrick, on behalf of Research Data Oxford, University of Oxford](#)
71. [Rebecca Osthus, American Physiological Society](#)
72. [Benjamin Haibe-Kains, University Health Network](#)
73. [Lynda Marie Emel, Fred Hutchinson Cancer Research Center](#)
74. [John Noel, Sleep Research Society](#)
75. [Robert M Cook-Deegan, Arizona State University](#)
76. [Tom Cheever, NIAMS/NIH](#)
77. [James H Jose MD, Children's Healthcare of Atlanta](#)
78. [Tobin Magle, Research Data Access and Preservation Association](#)

79. [Richard Platt, Adrian Hernandez, Lesley Curtis, Kevin Weinfurt \(see Purpose for full list\), Harvard Pilgrim Health Care Institute and Harvard Medical School; Duke University School of Medicine](#)
80. [Brett Harnett, University of Cincinnati](#)
81. [Mary Ellen K. Davis, Executive Director, Association of College & Research Libraries \(ACRL\)](#)
82. [Erik Deumens, University of Florida](#)
83. [Jason Hilton, Stanford University](#)
84. [Suzie Allard, University of Tennessee](#)
85. [Mary Lee Kennedy, Executive Director, Association of Research Libraries](#)
86. [Sue Miller](#)
87. [Tonia M. Masson, Society of Toxicology \(SOT\)](#)
88. [Mary Langman, Medical Library Association & Association of Academic Health Sciences Libraries](#)
89. [Jorge Contreras and Tammy Frisby, University of Utah](#)
90. [Mara Blake, on behalf of JHU Data Services, Data Services, Sheridan Libraries, Johns Hopkins University](#)
91. [Hae Kyung Im, The University Chicago](#)
92. [Scott Edmunds, GigaScience](#)
93. [Anshul Kundaje, Stanford University](#)
94. [Anonymous, International Society for Biological and Environmental Repositories \(ISBER\)](#)
95. [Jo Anne Goodnight, The Jackson Laboratory](#)
96. [Dylan Roskams-Edris, Canadian Open Neuroscience Platform and the Tanenbaum Open Science Institute](#)
97. [Keith Webster, Carnegie Mellon University](#)
98. [Timothy J. Triche, Jr., Van Andel Institute](#)
99. [Robert Allaway, Sage Bionetworks](#)
100. [Anthony Gitter, University of Wisconsin-Madison](#)
101. [Henry Chang, M.D.](#)
102. [Abeed Sarker, Emory University](#)
103. [Holly Murray, F1000](#)
104. [Sarah Damaske, Incoming Associate Director, Population Research Institute, The Pennsylvania State University](#)
105. [Barbara Stranger, Northwestern University Feinberg School of Medicine](#)
106. [Duke University Libraries Research Data Working Group, Duke University](#)
107. [John Wilbanks, Sage Bionetworks](#)
108. [Jeffrey Kidd, University of Michigan](#)
109. [Stuart Buck, Arnold Ventures](#)
110. [Janel Fedler, University of Iowa](#)
111. [Ian Moss, International Association of Scientific, Technical, and Medical Publishers \(STM\)](#)
112. [Jennifer Doty, Emory University](#)

113. [Jerry Blancato, EPA/ORD](#)
114. [Jeffery Smith, AMIA](#)
115. [M. Saiful Huq, PhD, President ,AAPM, American Association of Physicists in Medicine \(AAPM\)](#)
116. [Jaclyn Lucas, Beckman Research Institution of the City of Hope](#)
117. [Amazon Web Services, Amazon Web Services](#)
118. [Lisa Arafune, Coalition for Academic Scientific Computation \(CASC\)](#)
119. [Anurupa Dev, Association of American Medical Colleges](#)
120. [David Carr, Wellcome Trust](#)
121. [Andrew Smith, ELIXIR](#)
122. [Sarah Greene, Health Care Systems Research Network](#)
123. [Barbara E. Bierer MD, Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard \(MRCT Center\)](#)
124. [Stephanie J. Lee, MD, MPH, American Society of Hematology](#)
125. [Jennifer Graff, National Pharmaceutical Council](#)
126. [Chuck Cook, Global Biodata Coalition](#)
127. [Janis Geary, Arizona State University](#)
128. [Christopher Austin, Johns Hopkins University](#)
129. [Sarah Wright](#)
130. [Nicole Capdarest-Arest, University of California, Davis](#)
131. [Kevin McGhee, New York Genome Center](#)
132. [Joanna Groden, University of Illinois at Chicago](#)
133. [Patrick Dunn, Emma Afferton, John Campbell, Henry Schaefer, Elizabeth Thomson, ImmPort \(www.immport.org\)](#)
134. [Elisa Hurley, Public Responsibility in Medicine and Research \(PRIM&R\)](#)
135. [Ibraheem Ali, University of California, Los Angeles](#)
136. [Research Triangle Institute, Research Triangle Insititute](#)
137. [Jennifer A Doherty and Cornelia Ulrich, Huntsman Cancer Institute](#)
138. [Christopher Carr, RSNA](#)
139. [Salvatore La Rosa, Children's Tumor Foundation](#)
140. [Houri K. Vorperian, University of Wisconsin-Madison](#)
141. [Sally Gore, Lamar Soutter Library, University of Massachusetts Medical School](#)
142. [Erin F. Hering, Association for Research in Vision and Ophthalmology](#)
143. [Briana Ezray and Cynthia Hudson-Vitale, The Pennsylvania State University](#)
144. [Holly J. Falk-Krzesinski, PhD, Elsevier](#)
145. [Mary Jo Hoeksema, The Population Association of America/Association of Population Centers](#)
146. [Gerald J. Perry, Assoc Dean University Libraries/Lori Schultz, Sr Dir Research, Innovation & Impact, University of Arizona](#)
147. [Heather Joseph, SPARC \(The Scholarly Publishing and Academic Resources Coalition](#)
148. [Maryrose Franko, Health Research Alliance](#)

149. [Twila Reighley, Michigan State University](#)
150. [Collaborative Study on the Genetics of Alcoholism \(COGA\), COGA](#)
151. [Juliane Baron, Federation of Associations in Behavioral & Brain Sciences](#)
152. [Dr. Lisa Simpson, AcademyHealth](#)
153. [Tina Koplinski, Versiti Wisconsin](#)
154. [Susan Meyn, Association of Biomolecular Resource Facilities \(ABRF\)](#)
155. [Heidi Imker, University of Illinois at Urbana Champaign](#)
156. [Robert R. Montgomery, Blood Research Institute](#)
157. [Idan Gabdank, Stanford](#)
158. [Cole Allick \(Turtle Mountain Band of Chippewa Indians\), MHA, Washington State University, Institute for Research and Education to Advance Community Health \(IREACH\) and Partnerships for Native Health \(P4NH\)](#)
159. [Diane Lehman Wilson, University of Michigan Medical School](#)
160. [Dee Dee Aubourg, Acumen, LLC](#)
161. [Alessia Daniele, Cornell University and Weill Cornell Medicine](#)
162. [Brian Scarpelli & Alexandra McLeod, Connected Health Initiative](#)
163. [Emily Harris, Not Applicable](#)
164. [Seun Ajiboye, American Association for Dental Research](#)
165. [Council on Governmental Relations, Council on Governmental Relations](#)
166. [John Watts, Texas A&M University Libraries](#)
167. [Dennis Dean, Seven Bridges](#)
168. [Duke Office of Scientific Integrity, Duke University](#)
169. [Anthony Carvalloza, The Rockefeller University](#)
170. [Molly Timko, Hugo W. Moser Research Institute at Kennedy Krieger, Inc.](#)
171. [Damien Croteau-Chonka, Brigham and Women's Hospital / Harvard Medical School](#)
172. [Amonida Zadissa, European Bioinformatics Institute \(EMBL-EBI\)](#)
173. [Sarah Nelson, on behalf of the UW Genetic Analysis Center, University of Washington Genetic Analysis Center](#)
174. [Abigail Goben, University of Illinois at Chicago University Library](#)
175. [Katie Steen, Association of American Universities](#)
176. [Jacob Carlson, University of Michigan Library](#)
177. [Glenn Dillon, American Heart Association](#)
178. [Pamela Webb and Lisa Johnston, University of Minnesota](#)
179. [Heidi Rehm, Massachusetts General Hospital](#)
180. [Jessica Chong, University of Washington](#)
181. [Elizabeth A. McGlynn, Kaiser Permanente](#)
182. [Megan Potterbusch, George Washington University](#)
183. [Greg Janée, University of California at Santa Barbara](#)
184. [Ellen O'Meara, Kaiser Permanente Washington Health Research Institute](#)
185. [Agnes, University of California](#)
186. [Rajni Samevedam, MPH, Principal/Director, Booz Allen Hamilton](#)

187. [Amanda Gentzel, University of Massachusetts Amherst](#)
188. [Katherine Boronow and Julia Brody, Silent Spring Institute](#)
189. [Kenneth J. Ottenbacher, University of Texas Medical Branch, Galveston, TX](#)
190. [Maureen McArthur Hart, Global Genes](#)
191. [Kimberly Sabelko, Susan G Komen](#)
192. [Sarah Nusser, Iowa State University](#)
193. [Peter Sorger, Harvard Program in Therapeutic Science](#)
194. [Moffitt Cancer Center, H. Lee Moffitt Cancer Center & Research Institute, Inc.](#)
195. [Merce Crosas, Harvard University](#)
196. [Jennifer K. Wagner and Michelle N. Meyer, Geisinger](#)
197. [Ruth O'Hara, PhD, Stanford University](#)
198. [Felice J Levine, American Educational Research Association](#)
199. [James Love, KEI](#)
200. [Melissa Haendel and Julie McMurry, Monarch Initiative](#)
201. [Kristi Holmes, CTSA Program Center for Data to Health \(CD2H\)](#)
202. [Peter Sorger, Laura Maliszewski, Catherine Luria, Harvard Medical School](#)
203. [Kirk Francis, Kitcki A. Carroll, The United South and Eastern Tribes Sovereignty Protection Fund \(USET SPF\)](#)

Submission ID: 1223

Date: 11/8/2019

Name: Stacy Stuart

Name of Organization: University of Arkansas

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Infectious Diseases

Submission ID: 1224

Date: 11/9/2019

Name: Joel Voss

Name of Organization: Northwestern University

Type of Data of Primary Interest: Imaging

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Cognitive Neuroscience

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The issue I see with this policy is that without raising the budget cap on grants, you'll be asking us again to do more with the same amount of funding. For research with human subjects at least, the regulatory requirements have already become so cumbersome due to the new NIH policies that, at least in my experience, one full-time project manager must be devoted to this aspect of the project, which cuts into the budget allocated to the actual research and scientific goals. Although I value data sharing and appreciate the NIH acknowledgment that this practice is costly, it is not tenable to require it unless either the funding amount is raised or this is considered as an additional budget item that does not count towards the budget limit.

Submission ID: 1227

Date: 11/11/2019

Name: Andrew Vickers

Name of Organization: Memorial Sloan Kettering Cancer Center

Type of Data of Primary Interest: Clinical

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Cancer

Attachment:

Vickers data sharing comment.docx

From Andrew Vickers, Memorial Sloan Kettering Cancer Center
vickersa@mskcc.org

My name is Andrew Vickers and I have written about data sharing for many years, including papers in the literature (e.g. <https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-7-15> and <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007078>) and the lay press (e.g. <https://www.nytimes.com/2008/01/22/health/views/22essa.html>). I strongly support the NIH initiatives to promote data sharing.

I have one main comment. The policy allows researchers to determine whether data will be made available on an unrestricted basis or whether there will be “restricted access (made available only after the requestor has received approval to use the requested scientific data).” I would like to point out that “restricted access” is what we have had in science from year dot, and is what has been causing problems in the first place. What I detail in my prior work in the area is that researchers often cannot be reached (e.g. they have moved institutions) or that they simply refuse to share data (“we’ve evaluated your request and have decided that it is of insufficient scientific interest / we have similar analyses that are planned” etc. etc.). The policy compounds these problems by stating only that researchers need do no more than “consider describing the general terms of access for the data”.

I don’t see how this will avoid us getting into the situation where researchers can’t be reached, or they routinely refuse to share data and do so on grounds that are not related to the good of science as a whole. There needs to be a more explicit mechanism in the procedure to protect and promote data sharing such that data can be tracked down and that requests are only denied when there are compelling reasons to believe that this is not in the best interests of science. For instance, a restricted access mechanism might need to include the following steps:

- 1) Methods to contact investigators to request data must be robust to issues such as researchers changing institutions or retiring.
- 2) Anyone requesting the data must be informed (e.g. by providing a weblink) of the NIH policy that data must be shared except in defined circumstances.
- 3) Any refusals to share data must be accompanied by a memo giving detailed and compelling reasons. This memo would need to be sent to the NIH and would be published, alongside the original data request.
- 4) An ombudsman should be created to mediate disputes about data sharing.

Submission ID: 1237

Date: 11/12/2019

Name:

Section VII: Compliance and Enforcement:

The data for any clinical trial will not likely be ready for final archiving in a repository or other site until it has been analyzed. Therefore, to state compliance with archiving reviewed at regular reporting intervals is a moot point, as is the language of the extramural grant bullet. It will not be known until after the analysis if the PI has complied with archiving. What will a threat of termination of award do? The likelihood there is any funding by the archive time is highly unlikely. Therefore, there is no strength behind the requirement. Might there be value in holding back a percentage of funding with its release after archiving has been completed?

Additionally, can NIH annually post a list of noncompliant PIs on the public-facing website?

Submission ID: 1238

Date: 11/12/2019

Name: Jesse Forrest

Name of Organization: Immune Tolerance Network

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All data.

Type of Organization: Nonprofit Research Organization

Role: Other

Role - Other: Principal Data Architect

Domain of Research Most Important to You or Your Organization:

Immunology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section I clearly states the intended goals of Data Management and Sharing policy. I'm very happy to see the FAIR principles called out here -- I believe the FAIR principles should be adopted as widely as possible in the scientific community as they promote ease of access to data while preserving code in a reusable way.

Section II: Definitions:

Clear and easily understood definitions.

Section III: Scope:

The scope is clear and concise. I'm very excited to see that this policy will apply to all research funded by the NIH. I think this is a big win for the scientific community.

Section IV: Effective Date(s):

The effective dates are clear and easy to follow.

Section V: Requirements:

The requirements look good. A little sparse on specifics but this is understandable as most requirements will be in the NIH-ICO-approved plan (this is my understanding).

Section VI: Data Management and Sharing Plans:

This section is also clear and concise. I'm glad to see the statement "NIH recognizes that certain factors (e.g., legal, ethical, technical) may limit the ability to preserve and share data." --It is good to see that this understanding is baked into the draft policy. Though, researchers may not be aware of some of the databases the NIH provides to submit sensitive data to, for example: dbGaP for GWAS and WGS data.

Section VII: Compliance and Enforcement:

This section seems standard, straightforward.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

It is good to see this supplemental guidance doc. Often, smaller research organizations are flaky with their implementation of data security standards or common archival practices because the cost can be limiting. I think this guidance document will go a long way in helping potential grantees understand what resources are available to them.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

I'm quite thrilled to see this guidance document -- Smaller research organizations will have difficulty producing a Data Management and Sharing plan simply due to a lack of technical expertise on the subject. This guidance document lays out easy to follow descriptions of each section without using overly complex terminology.

There are a few small typos in this documents though: Mostly missing spaces. Some examples include:

Section 1: Data Type - Bullet point 1:

"(e.g., exome sequences of 20 to 30 gene variants..."

"Section 5: Data Sharing Agreements, Licenses, and Other Use Limitations:"

Other Considerations Relevant to this DRAFT Policy Proposal:

Overall I'm ecstatic to see this draft policy as it touches on the spirit of open and easily accessible research data while closely adhering to the goals of the NIH.

I would encourage the NIH to include either the FAIR principles paper in full as a supplemental guidance doc, or a summary of the FAIR principles.

Feel free to contact me directly for an example FAIR principles summary document.
(jforrest@immunetolerance.org)

Submission ID: 1239

Date: 11/12/2019

Name: George McNamara

Name of Organization: Johns Hopkins University

Type of Data of Primary Interest: Imaging

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Healthcare

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

NIH provides NO JUSTIFICATION for its statement (in supplemental DRAFT Guidance pdf):
"Note, NIH does not expect researchers to share all scientific data generated in a study."

I suggest instead: "All data from research funded partially or entirely by NIH is expected to be publicly available. the only exceptions are with respect to HIPAA compliance, similar protected personal information, and U.S. national security.

Copyright: The DRAFT Guidance and Supplemental Draft Guidance fail to mention copyright. The NIH funded scientific community is already stuck with for profit companies charging access for much of the publications funded by NIH. NIH's failure to address copyright simply invites publishers -- and researchers -- hiding their data behind a paywall. I strongly urge NIH to update its Guidance with text consistent with the following logic:

Data are facts. Facts cannot be copyrighted.

Attribution is not a problem: I do note that attribution of data source (i.e. original author's publication, with complete methods can be handled by the simple expectation: "give credit where credit is due".

I published in 2006 (McNamara et al 2006 Cytometry A 69A: 863-871, open access <https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.20304>)

The principle that data is not subject to copyright provides a framework in which all scientific data should be made freely accessible.

In addition to obtaining numerical data on request, the U.S. Supreme Court's observation that data is not subject to copyright (14) provided the rationale to digitize spectra from published graphs.

Data Is Not Copyrightable

During the course of developing this data, one of us had an epiphany while reading in Lessig (18) about a U.S. Supreme Court decision: data is not subject to copyright (14). Text and commentary about Feist can be found on many legal web sites by doing a Google search. Indeed, the broad availability of the text of Supreme Court decisions is because they are not subject to copyright. The Feist decision reaffirmed the U.S. Copyright act of 1976 that "there can be no copyright in facts". The basis for the Feist decision can be found in the U.S. Constitution.

14. Feist Publications, Inc. v. Rural Tel. Serv. Co. 1991;499 U.S. 340.

18. Lessig L. The Future of Ideas. New York: Random House; 2001. p 368.

Note: The current PubSpectra data download web site is

<https://works.bepress.com/gmcnamara/9> (the PubSpectra data has been re-used in several fluorescence spectra graphing web sites).

Submission ID: 1241

Date: 11/13/2019

Name: Alik Widge

Name of Organization: University of Minnesota

Type of Data of Primary Interest: Imaging

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

cognitive neuroscience, particularly large-scale electrophysiology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

These appear reasonable.

Section II: Definitions:

These appear reasonable.

Section III: Scope:

These appear reasonable.

Section IV: Effective Date(s):

I do not believe this would take very long to comply with. It adds maybe an hour to total prep time for an R01-equivalent. Make it happen soon.

Section V: Requirements:

These appear reasonable.

Section VI: Data Management and Sharing Plans:

These appear reasonable.

Section VII: Compliance and Enforcement:

These appear reasonable. I would not mind if it were a scored review criterion.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

One thing that is almost, but not quite, covered here: can we use this guidance to cover pre-payment of a fee to make data available long term/in perpetuity? There have been difficulties in some institutions where project funds were not permitted to be used to buy, e.g, the next 10 years of data storage up front (to buy services to be rendered outside the project period).

I think some of the language in here about fees would make that allowable, but I would like you to be more explicit. The biggest issue we are facing right now is that so many repositories charge per year, and we can't find a way to pay those fees continuously!

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

These appear reasonable.

Other Considerations Relevant to this DRAFT Policy Proposal:

Broadly: this is a timely and much needed initiative. It is a reasonable policy. We are a fair-sized lab (three R-equivalent grants, six doctoral-level researchers) and I would not consider it overly burdensome to comply with this policy. I *would* estimate that my total costs of compliance will be around \$25,000 per grant, maybe even \$50,000 if repository fees get high. This is something to consider, because compliance would be difficult for a modular-budget award.

Submission ID: 1244

Date: 11/15/19

Name: Nancy Janitz

Type of Data of Primary Interest: Qualitative

Type of Organization: Government Agency

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Brain Initiative Research and the FDA Clinical Trials and the SDOH On Developing Brain of Children and young adults

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I am not Going To Worry about my privacy if it is going to help the Research and the children and future generations

Section II: Definitions:

I have provided the information that I have experienced and have observed in the past and it is very important to the Research Community To do the best Research to prevent future generations of going through what I have experienced and seen in other people. Both Mental Illness and Dual Addiction .

Section IV: Effective Date(s):

I submitted my information on May 2016: and Then I found out that the FDA was Conducting Advisory Hearings regarding my medications prescribed since 2001!?!

Section V: Requirements:

I am giving my permission to use my personal data for NIH And Affiliate Agencies for Furthering Research On Brain Initiative and the Children.

Section VI: Data Management and Sharing Plans:

I am providing my permission for the use of the Information In NIH Research Funding Grants and I have provided it voluntarily without any payment or other Recognition; I simply want to be able to use my life experiences and observations to help the children and future generations.

Submission ID: 1245

Date: 11/16/19

Name: elly lee

Name of Organization: gachon university

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

cardiovascular

Submission ID: 1246

Date: 11/17/19

Name: Emily Scott

Name of Organization: University of Michigan

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Structural Biology

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

If there are ongoing costs for data storage beyond the lifetime of the NIH funding, how will this be managed?

Submission ID: 1249

Date: 11/20/19

Name: Samarendra Mohanty

Name of Organization: Nanoscope Technologies, LLC

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: Biotech/Pharmaceutical Company

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Neuroscience, Ophthalmology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Small businesses should not be forced to share data with public as it will put them in competitive disadvantage.

Section III: Scope:

Small businesses should be excluded

Section VI: Data Management and Sharing Plans:

Small businesses should be excluded from public data sharing at least 5 yrs of data being generated

Submission ID: 1253

Date: 11/26/19

Name: Mona Hicks

Name of Organization: One Mind

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All types of data are of interest.

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

All data are of interest because we value the importance of understanding typical and atypical mechanisms of brain function in prevention and management of neurological and psychiatric disorders.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The policy is clear and concise. It has the potential to radically improve biomedical research as we know it because data sharing will likely 1) encourage investigators to invest more time and resources for obtaining high quality, well-curated data and 2) enable the re-analysis and use of data for new research questions.

Section III: Scope:

Including all funded projects in the policy is big, bold and transformative.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Acknowledging that additional resources will be needed is a strength of the policy. The big question is how much this will cost and whether efficiencies can be developed over time to reduce the costs.

Other Considerations Relevant to this DRAFT Policy Proposal:

It will be important to have a 5 and 10 year plan for cost benefit analysis of the data sharing policy that can be broadly disseminated. Admittedly, changing the research enterprise will take time, that's why 5 - 10 years is recommended.

Submission ID: 1255

Date: 11/29/19

Name: Karen Sherman

Name of Organization: Kaiser Permanente Washington Health Research Institute

Type of Data of Primary Interest: Clinical

Type of Organization: Health Care Delivery Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

clinical trials and observational data, some of which can come from electronic health records and others from questionnaires

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I think that data sharing at some level is useful and appropriate for furthering scientific research. I do, however, have concerns about data sharing in extremely small studies.

Section II: Definitions:

Metadata: The first use of data is confusing. Suggest: A document that provides additional information describing the data to be shared. For example, USE WHAT YOU ALREADY HAVE).

Section III: Scope:

I think this could be difficult for small grants, such as R03 and R21 unless additional funds are provided for actually sharing data, especially if they need to be in a repository or something like that.

Section IV: Effective Date(s):

Agree should not be retroactive.

Section V: Requirements:

This is fine. It's important to note that sharing individual level data may not be appropriate for every study.

Section VI: Data Management and Sharing Plans:

The requirement for just-in-time seems a bit onerous as there is not funding for doing this. While some of this may be generic, others may not be and would require a bit more work.

Section VII: Compliance and Enforcement:

I suggest reminders related to data sharing at appropriate intervals because the need to alter the plan is real.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

I am concerned that smaller awards are already tight in their budget so I believe the option for a supplement for data sharing or extra costs for that are appropriate .

Submission ID: 1256

Date: 12/02/19

Name: Melinda Marino

Name of Organization: Arizona State University

Type of Data of Primary Interest: Qualitative

Type of Organization: University

Role: Other

Role - Other: Research Administrator

DRAFT NIH Policy for Data Management and Sharing

Section II: Definitions:

Metadata could and probably should be defined in more detail. It would be good to see that explicitly articulated, e.g.

Metadata: Data describing scientific data that provide additional information in sufficient detail to reduce the risk of misinterpretation and to make such scientific data more understandable (e.g., date, independent sample and variable description, unit definitions, outcome measures, and any intermediate, descriptive, or phenotypic observational variables).

Submission ID: 1257

Date: 12/03/19

Name: Deirdre Joy

Name of Organization: NIAID/NIH

Type of Data of Primary Interest: Genomic

Type of Organization: Government Agency

Role: Government Official

Domain of Research Most Important to You or Your Organization:

Infectious Diseases

DRAFT NIH Policy for Data Management and Sharing

Section VII: Compliance and Enforcement:

There should be a section of the RPPR specific to compliance with the data sharing policy and progress to-date in implementing the Data Sharing Plan in which the PI has to address data they have generated and steps they have taken to make it public. A question on the Program Officer checklist without a corresponding Data Sharing section in the RPPR will make the policy nearly unenforceable.

Submission ID: 1258

Date: 12/03/19

Name: Andreas Mueller

Name of Organization: Columbia University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Data Analysis

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Machine Learning, Data Science

DRAFT NIH Policy for Data Management and Sharing

Section II: Definitions:

The definitions do not clarify if software is considered data. I would highly encourage the NIH to explicitly include software artifacts in the research data. Nearly all research that requires data also requires custom software artifacts to process that data. However, the words "software" or "code" are not mentioned in the draft. While software is mentioned in the "Elements" document, this section is about which software is needed, not about providing the software.

I would strongly argue that the software itself is an artifact that needs to be documented, preserved and licensed.

Submission ID: 1259

Date: 12/03/19

Name: Michele Diaz

Name of Organization: Pennsylvania State University

Type of Data of Primary Interest: Imaging

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

cognitive neuroscience

DRAFT NIH Policy for Data Management and Sharing

Section VI: Data Management and Sharing Plans:

this seems entirely reasonable, particularly given that there could be allowable costs. That seems to be one of the biggest barriers to sharing (mainly the time cost (i.e., personnel cost) in making data available. There are also concerns about releasing data prematurely that would allow others to "scoop" findings. But the wording of the proposal seems to imply that there would be flexibility. The requirement simply seems to be making PIs come up with a plan for data sharing, which they should be thinking about anyway.

Review panels are starting to look for this more and more (even when it's not required), and I think overall, there is greater benefit, than cost to data sharing.

Submission ID: 1260

Date: 12/03/19

Name: Ellen M. Wijsman

Name of Organization: University of Wahsington

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

genetic epidemiology

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

It will be very expensive to do this properly. Yes, the initiative allows investigators to budget in costs, but those funds are going to have to come from somewhere. Data managers, even just out of school, cost a lot more than most junior biologists, if you can hire them at all. Is spending so much money on data sharing really a good idea? The money will have to come from somewhere. That will put even more stress on selecting projects to fund, and on getting the research done. And by not allowing normal costs for facilities and management fees, it will be even harder. Where are we going to put such people if we aren't allowed to charge for their space? How will we pay for their internet charges, their use of resources, etc.? If the data are to be stored on a university-supplied site, who is going to pay for that facility in perpetuity? NIH has already failed in its early mission of trying to capture all genomic data: once we hit sequence data, the funding to dbGaP to receive and store those data (especially given how few people successfully downloaded the data), dbGaP expelled the big sequencing projects because of resource use.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

While I understand the *ideas* being proposed, I expect that if the data management & sharing document is approved, there are going to be big unanticipated negative outcomes that may hurt our scientific progress more than not having access to everyone's data. The first problem is that the document assumes that we can just hire data managers at will. Nothing could be farther from the truth. We simply do not have enough competent data management people in this country with the right skills to do the work needed to comply with what is being proposed. The second is that without checks of what is being shared, what *will* get shared will likely be useless, as I have found with several existing NIH-"supported" data sharing sites. The existing NIH data sharing sites already do a poor job of making sure the data submitted make sense. They focus on the existence of data, not the existence of good-quality, QC-ed, data. Some of them force the data to conform to particular formats, and can develop at least some tools to make sure the data submitted are at least nominally what might be expected. But if data start getting submitted in all sorts of different formats, it will be 1000 times harder to make sure that the right kind of data are being submitted (by right kind, I mean that integers are submitted where counting numbers would go, floating point numbers where measurements might go, etc.). Some such sites, like NDAR, force data into formats that are very non-standard and to work with both for submission and for analysis, adding a great deal of data management overhead to the local data management costs. In addition, since the people with whom the data managers consulted with are often not the people who know anything about data management or downstream data analysis, the product for analysis, in the end is so poor that it puts people off for future data use.

Other Considerations Relevant to this DRAFT Policy Proposal:

It reads like a grand plan that sounds great on paper, but will create a bureaucratic nightmare and become a colossally expensive nightmare to implement in a fashion that is not simply a drain on getting good science done.

Submission ID: 1261

Date: 12/05/19

Name: Richard A Kahn

Name of Organization: Emory University

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Basic sciences (e.g., biochemistry, cell biology, genetics)

Other Considerations Relevant to this DRAFT Policy Proposal:

Although likely viewed as perhaps not central to this proposed policy, I have long advocated for HHS spearheading the generation and distribution of software to encourage the use of affordable, common, searchable laboratory notebooks. Although there are commercial products available these are costly and not readily adapted for use in all labs. The use of such software would promote appropriate and ready sharing and storage of data, which IS the goal of your initiative. I would love to see this become a part of your efforts.

Submission ID: 1262

Date: 12/09/19

Name: Brian C Trainor

Name of Organization: UC Davis

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Behavioral neuroscience

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

There needs to be some flexibility in how and where data are made available. If there is not a central depository, where will funding come from to maintain databases?

Section VI: Data Management and Sharing Plans:

I definitely think researchers need to be able publish data first before sharing data. It would be unfair if a big lab were to come along and analyze a slower moving lab's data first.

Submission ID: 1263

Date: 12/09/19

Name: David Cormode

Name of Organization: University of Pennsylvania

Type of Data of Primary Interest: Imaging

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Contrast agents

DRAFT NIH Policy for Data Management and Sharing

Section VI: Data Management and Sharing Plans:

Broadly speaking the draft seems fine. However, the NIH should do its utmost to minimize the burden on researchers. Example data management plans should be provided for various different fields, in order to streamline this additional administrative burden for faculty and staff.

Submission ID: 1264

Date: 12/09/19

Name: Evan Mayo-Wilson

Type of Data of Primary Interest: Clinical

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

I am commenting as an individual rather than a representative of my organization.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

This policy is an important and valuable contribution. Data and code sharing have the potential to advance scientific progress and to increase return on NIH investments.

Section II: Definitions:

No comments

Section III: Scope:

The policy could go further and include standards for sharing research materials; for example, manuals needed to replicate interventions in future studies or in practice should be available publicly. The TOP guidelines provide a useful framework.

Section IV: Effective Date(s):

No comments

Section V: Requirements:

This is a helpful step towards new norms in science. The policy would be even better if it were to indicate that NIH expects data to be shared on permanent repositories, and if the policy required that investigators provide a strong rationale for refusing to share data from NIH sponsored research. Recommendations about code sharing could be stronger. Code can usually be shared at low cost and with few ethical concerns, so most code that is developed using NIH funding and used to produce results in NIH funded research could be freely available in permanent repositories.

Section VI: Data Management and Sharing Plans:

No comments

Section VII: Compliance and Enforcement:

Policies that aim to increase research transparency, including guidelines for registering and reporting clinical trials, have not been enforced. Some investigators see these as toothless policies that can be ignored. NIH should take enforcement actions where needed.

Submission ID: 1265

Date: 12/09/19

Name: Mara Mather

Name of Organization: USC

Type of Data of Primary Interest: Imaging

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

affective neuroscience

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

"Researchers with NIH-funded or conducted research projects resulting in the generation of scientific data are required to submit a Plan to the funding NIH ICO as part of Just-in-Time for extramural awards."

Why not make it a requirement to submit a plan as part of the initial research proposal? I believe that peer review would be helpful to make these plans as effective as possible.

Submission ID: 1266

Date: 12/10/19

Name: Hannah V. Carey, PhD, FASEB President

Name of Organization: Federation of American Societies for Experimental Biology (FASEB)

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: Nonprofit Research Organization

Role: Other

Role - Other: Coalition of 29 Scientific Societies

Domain of Research Most Important to You or Your Organization:

Basic Biology and Biomedical Research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

This section describes the philosophy underlying the policy and is a helpful reminder that investigators are not conducting their work within a vacuum. Highlighting the need to consider data preservation and sharing as part of the research process is critical to foster culture change. We do suggest, however, that the policy more clearly define acceptable timeframes for data sharing, as “timely manner” could be widely interpreted. These could even be conveyed as ranges to preserve flexibility.

Section II: Definitions:

FASEB thanks NIH for expanding the definition of “scientific data” to include negative results. Defining scientific data as all findings contributing to a line of research inquiry ensures transparency and improves the rigor and reproducibility of research findings.

Section V: Requirements:

While FASEB supports the requirement of a data management and sharing plan for NIH-funded or conducted research, we are concerned about varied supplementary information requirements requested by individual NIH Institutes, Centers, and Offices (ICOs). To minimize confusion and administrative burdens, we strongly encourage trans-NIH coordination of these supplemental requests and listing ICO-specific requirements as part of centralized resources associated with the final data management and sharing policy.

Section VI: Data Management and Sharing Plans:

FASEB applauds the proposal to collect data management and sharing plans as part of Just-in-Time documentation for extramural awards. Requiring submission of the plan as part of the term of award rather than the initial proposal minimizes administrative burden at the proposal stage for both the applicant and peer reviewers. Shifting the review of plans to NIH staff members rather than volunteer reviewers will also make the process more uniform and streamlined. This also allows more flexibility for grantees to make real-time updates to their plans.

One area that needs to be further clarified in the final policy is whether NIH will make data management and sharing plans publicly available. To truly fulfill the FAIR data principles, plans should be made publicly available; however, we urge further engagement with the stakeholder community to determine possible unintended consequences of this strategy. Another approach may be to share limited details about the plan to increase awareness of the work, particularly if the work leads to outputs other than publications.

Section VII: Compliance and Enforcement:

The strategy of making the data management and sharing plan a term and condition of the grant award demonstrates NIH's commitment to fostering a culture of data sharing among investigators and institutions supported by NIH funding and support.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

FASEB appreciates NIH's recognition of the costs associated with data management and sharing and applauds the inclusion of the supplemental guidance defining possible allowable costs. A concern is that the guidance only addresses those costs incurred during the term of the award but does not address costs associated with long-term data retention and accessibility.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

FASEB commends the inclusion of supplemental guidance to help investigators understand the desired elements of a data management and sharing plan. The proposed guidance offers investigators flexibility to adapt plans to their specific research needs. This, in concert with an enhanced role for NIH staff in reviewing draft plans, should help alleviate confusion regarding expectations for data management and sharing plans.

Other Considerations Relevant to this DRAFT Policy Proposal:

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to provide comments in response to NOT-OD-20-013, Request for Public

Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance. FASEB is comprised of 29 scientific societies, collectively representing over 130,000 biological and biomedical researchers who produce and use a wide variety of data, core data resources, and analytic tools.

In reviewing the draft policy and supplemental guidance documents, we were pleased to see FASEB's feedback in response to NOT-OD-19-014, Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research clearly incorporated. While we are still concerned about variability in terms of individual investigators' expectations, experience, and resource needs to ensure key data from NIH funded/supported projects are consistent with the FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles, the draft policy provides flexibility to develop a culture of data management and sharing within the NIH funded community.

We commend NIH for its careful consideration of the comments received in response to the NOT-OD-19-014. The result is a draft policy that is adaptable to the broad range of science supported by NIH and furthers the NIH goal of building the culture of data management and sharing across the biological and biomedical research community. Once the policy is finalized, we strongly encourage extensive engagement with the scientific community to clarify agency process and expectations prior to enforcing compliance as rushed implementation can result in unforeseen challenges.

Attachment:

FINAL FASEB Response_NIH Draft Data Sharing Plan_20191210_LETTERHEAD.pdf

Description:

Compiled organizational comments on letterhead



FASEB comments in response to [NOT-OD-20-013](#), “Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance”

Comments submitted electronically via [online Comment Form](#) on December 10, 2019

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to provide comments in response to NOT-OD-20-013, Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance. FASEB is comprised of 29 scientific societies, collectively representing over 130,000 biological and biomedical researchers who produce and use a wide variety of data, core data resources, and analytic tools.

In reviewing the draft policy and supplemental guidance documents, we were pleased to see FASEB’s [feedback](#) in response to [NOT-OD-19-014](#), Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research clearly incorporated. While we are still concerned about variability in terms of individual investigators’ expectations, experience, and resource needs to ensure key data from NIH funded/supported projects are consistent with the FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles, the draft policy provides flexibility to develop a culture of data management and sharing within the NIH funded community.

Comments on specific aspects of the draft policy are noted below.

Purpose: This section describes the philosophy underlying the policy and is a helpful reminder that investigators are not conducting their work within a vacuum. Highlighting the need to consider data preservation and sharing as part of the research process is critical to foster culture change. We do suggest, however, that the policy more clearly define acceptable timeframes for data sharing, as “timely manner” could be widely interpreted. These could even be conveyed as ranges to preserve flexibility.

Definitions: FASEB thanks NIH for expanding the definition of “scientific data” to include negative results. Defining scientific data as all findings contributing to a line of research inquiry ensures transparency and improves the rigor and reproducibility of research findings.

Requirements: While FASEB supports the requirement of a data management and sharing plan for NIH-funded or conducted research, we are concerned about varied supplementary information requirements requested by individual NIH Institutes, Centers, and Offices (ICOs). To minimize confusion and administrative burdens, we strongly encourage trans-NIH coordination of these supplemental requests and listing ICO-specific requirements as part of centralized resources associated with the final data management and sharing policy.

Data Management and Sharing Plans: FASEB applauds the proposal to collect data management and sharing plans as part of Just-in-Time documentation for extramural awards. Requiring submission of the plan as part of the term of award rather than the initial proposal minimizes administrative burden at the proposal stage for both the applicant and peer reviewers. Shifting the review of plans to NIH staff members rather than volunteer reviewers will also make the process more uniform and streamlined. This also allows more flexibility for grantees to make real-time updates to their plans.

One area that needs to be further clarified in the final policy is whether NIH will make data management and sharing plans publicly available. To truly fulfill the FAIR data principles, plans should be made publicly available; however, we urge further engagement with the stakeholder community to determine possible unintended consequences of this strategy. Another approach may be to share limited details about the plan to increase awareness of the work, particularly if the work leads to outputs other than publications.

Compliance and Enforcement: The strategy of making the data management and sharing plan a term and condition of the grant award demonstrates NIH's commitment to fostering a culture of data sharing among investigators and institutions supported by NIH funding and support.

Supplemental Draft Guidance – Plan Elements: FASEB commends the inclusion of supplemental guidance to help investigators understand the desired elements of a data management and sharing plan. The proposed guidance offers investigators flexibility to adapt plans to their specific research needs. This, in concert with an enhanced role for NIH staff in reviewing draft plans, should help alleviate confusion regarding expectations for data management and sharing plans.

Supplemental Draft Guidance - Allowable Costs: FASEB appreciates NIH's recognition of the costs associated with data management and sharing and applauds the inclusion of the supplemental guidance defining possible allowable costs. A concern is that the guidance only addresses those costs incurred during the term of the award but does not address costs associated with long-term data retention and accessibility.

We commend NIH for its careful consideration of the comments received in response to the NOT-OD-19-014. The result is a draft policy that is adaptable to the broad range of science supported by NIH and furthers the NIH goal of building the culture of data management and sharing across the biological and biomedical research community. Once the policy is finalized, we strongly encourage extensive engagement with the scientific community to clarify agency process and expectations prior to enforcing compliance as rushed implementation can result in unforeseen challenges.

Submission ID: 1267

Date: 12/11/2019

Name: Betsy L Humphreys

Name of Organization:

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All categories of data generated with NIH funding

Type of Organization: Not Applicable

Role: Member of the Public

Domain of Research Most Important to You or Your Organization:

All categories of research funded by NIH

DRAFT NIH Policy for Data Management and Sharing

Section VI: Data Management and Sharing Plans:

The final policy should require that (1) grant applicants describe, at least in general, all the elements of their data management and sharing plan in their initial grant applications and that (2) external peer grant reviewers review the information provided and reflect their assessment of it in scoring proposals. This will send the strongest signal that NIH is committed to advancing data management and sharing. The information described under 1. Data Type and 3. Standards in the "Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan" should be included in the initial proposal at the same level of detail described in that guidance. This information will provide important insight into the strengths and weaknesses of the science of the proposal and the team proposed to carry it out, including their knowledge and understanding of the broader scientific utility of the data they will generate and the existing standards that are applicable to those data. In addition, having external reviewers review and score data management and sharing plans will avoid problems that could arise if "just in time" review of these elements by NIH staff reveals serious flaws in a proposal that was highly ranked during external review.

Submission ID: 1268

Date: 12/11/2019

Name: Rochel Gelman

Name of Organization: Rutgers, Center for Cognitive Science and Psychology

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Videotape, Questions, Simple choice

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Cognitive and Language Development ; School-based learning- usually with normal individuals

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

To develop (a) nature of early cognitive development and the task variables that influence a child's performance; to develop learning trajectories for Science with Math in school - at all levels.

Section II: Definitions:

Habituation in infants - detection of post habituation change; Surprise reactions to critical and non-critical changes in number and cause; Preferential attention - to initial and new stimuli; Prediction re changes and Checking after the Change.

Section III: Scope:

I do not know what this question is about

Section IV: Effective Date(s):

I am retired. Have some data or printouts going back 40 years.

Section V: Requirements:

Permission from Schools, teachers, and guardians

Section VI: Data Management and Sharing Plans:

I am retired and do not plan to submit a proposal.

Section VII: Compliance and Enforcement:

I have always met all requirements, including police checks for members of lab working with young children in a school setting.

Other Considerations Relevant to this DRAFT Policy Proposal:

The requirements for working with my target samples are unreasonable. My lab is dedicated to uncovering cognitive, communication and language abilities as well as the variables that affect performance. It is typically the case that teachers are pleased to learn more about what their charges can do and often end up embedding a version of my research into their offerings. Bear in mind: I am, in one way or another asking young children "How many?", Please count as high (or equivalent) as you can. When we interact with elementary, High School and College students, the goal is to determine hidden reasons for errors, which are usually default . Then we develop experiences for teaching what can lead the students to stop defaulting and moving forward. The theoretical outcomes have informed several successful teaching programs. One way to put the matter: I ask participants to respond to items that can lead from what they know to what they still have to learn. Finally, a part of this effort involves learning how to ask questions that are understood.

Description:

Draft reply NIH Policy

Submission ID: 1269

Date: 12/11/2019

Name: Christine Morrison

Name of Organization: CDC

Type of Data of Primary Interest: Clinical

Type of Organization: Government Agency

Role: Government Official

Domain of Research Most Important to You or Your Organization:

Infectious diseases

DRAFT NIH Policy for Data Management and Sharing

Section II: Definitions:

There needs to be further clarification regarding the definition of scientific data. Raw data do not need to be shared and there are justifiable exceptions to sharing data to protect research subjects and award recipients.

Whereas 45 CFR 75.322 notes that the Federal Government has the right to:

- a. Obtain, reproduce, publish, or otherwise use the data produced under a Federal award; and
- b. Authorize others to receive, reproduce, publish, or otherwise use such data for Federal purposes,

45 CFR 75.322 goes on to define research data as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following:

Preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This “recorded” material excludes physical objects (e.g., laboratory samples).

Importantly, research data also do not include:

- (i) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and
- (ii) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.

Submission ID: 1270

Date: 12/12/2019

Name: Rhoda Au

Name of Organization: Boston University, School of Medicine/Public Health

Type of Data of Primary Interest: Clinical

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:
cognitive aging and dementia, large scale epi cohorts

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

To leverage NIH investment, lower the barriers and access to publicly funded data resources, particularly those that involve larger scale longitudinal studies. An important consideration to achieving this goal is to revisit longstanding governance policies and procedures that have been adopted by Executive Committees, whose personal professional interests do not align with proactively finding ways to facilitate much broader data sharing, rather than follow prescriptive guidelines and/or adhering to the most minimal level of data sharing. Want to stop the practice of the "illusion" of broad data sharing in face of stark evidence that there is far less data sharing than is feasible.

Section II: Definitions:

Given the technological advances, there is now the possibility of "data lending"/"data borrowing" that alleviates many of the excuses (some that are legitimate) that many investigators give that limit data sharing. Need to be much more explicit as to what is defined as true data sharing. Most data sharing plans give lip service to what is considered acceptable/sufficient but don't have a truly defined plan that would result in a much more data sharing plan. The concept of data ownership needs to be better defined so that it's clear that researchers don't "own data" - only the people who volunteer for research own their own data. Researchers responsibilities include making sure that participants' data is used to its maximum scientific potential.

Section III: Scope:

Policies have to be set that maximize data release and on a much more accelerated timeline. There should not be any pre-defined scope of what scientific questions can and cannot be addressed. Right now, the review process of use of data puts too many restrictions based on a priori beliefs/practices. The definition of innovation is doing something that hasn't been done before. Thus to drive opportunities for true discovery and innovation will necessitate getting rid of many of the reviewers presumptions of what is the "right" science and what is not. This obviously does not preclude anything that would risk participant privacy and confidentiality. But for example, the "black box" criticism of deep machine learning and quantum mathematics approaches should not be restricted because traditional researchers don't understand it.

Section IV: Effective Date(s):

It is important to explicitly define dates in which data must be released for use beyond the core research team.

Section V: Requirements:

Policies need to be much more explicit on how "embargo" periods are defined. Left to their own devices, some research teams will define the start of the embargo period after the last data point has been collected. In large scale epi studies that take upward of 5+ years to collect data from all potential participants, the traditional 1-2 year embargo period can stretch to closer to 5-6 years.

Section VI: Data Management and Sharing Plans:

Important to bring in private industry expertise - the financial sector likely has some significant expertise because they have to process millions of transactions a day, all while maintaining utmost data privacy/security. They also have to make their data available to analysts all across the world. Really need to tap enterprise level knowledge/expertise and not leave the development of a robust data management and sharing plans in the hands of researchers, clinicians and biostatisticians/epidemiologists.

Section VII: Compliance and Enforcement:

Need to provide funding to support research groups getting into compliance and then maintaining compliance. It would be really good to take a few complex high value NIH funded datasets and use them as models of how to get legacy datasets into compliance and then use them as examples of how a robust data sharing plan is enforced. These examples will also create tools and solutions for other research groups to follow.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Costs are going to depend on what datasets are involved. Large scale legacy cohorts are going to require a substantial investment to bring to compliance. A separate budget section may be needed to support data management and sharing efforts since these activities are generally

above and beyond what is needed to do the data management work to meet the aims of the proposal.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Must be future-proof. Whatever is developed must be done in a way that allows upgrading and flexibility for different methods that will emerge in the decades to come. It is critical to come up with elements that won't become quickly obsolete and no means of pivoting.

Other Considerations Relevant to this DRAFT Policy Proposal:

Need to build in career/funding incentives to compel researchers to want to share data. There are all sorts of ways to hide data and only do "surface level" sharing. The objective is to make it worth it to the researcher to fully disclose all data available for sharing and to actually share it. Too often a researcher can make claims for why some data can't be shared. This is applying a priori assumptions and biases and justifying withholding of data that might in fact be quite valuable if used in a different manner/context.

Submission ID: 1271

Date: 12/12/2019

Name: Keri Hornbuckle

Name of Organization: University of Iowa

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

human exposure, toxicology, environmental remediation, analytical chemistry

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We (Keri Hornbuckle and Brian Westra) are writing to comment on the draft NIH Policy for Data Management and Sharing . We are part of the Iowa Superfund Research Program, a multidisciplinary research center funded by the Superfund Research Program office of the National Institute for Environmental Health Sciences (SRP/NIEHS). Hornbuckle is the Director of the P42 center and Westra is a data services librarian. We have worked together to prepare data management and sharing plans as part of our application for a competitive renewal of the center.

Section II: Definitions:

no comments

Section III: Scope:

no comment

Section IV: Effective Date(s):

Concerning Section IV, we are concerned about when and under what funding circumstances the policy would be effective.

The policy should only apply to requests for applications (RFAs) released after the final policy is announced so that grantees and their institutions can develop and provide training/instruction,

resources, consultations, and other services/infrastructure. The policy should not apply to ongoing grants that have already been budgeted for the work described in the original application. Preparation of data management plans takes time and must be developed in coordination with the study being proposed.

Should the NIH decide to require data management plans from ongoing funded grants, the expectation for completion and scope cannot be the same as that required of new applications. The expectations must be proportional to the funding provided for the activity, and the additional effort required to develop a plan after the funded study has been designed and is being carried out.

Section V: Requirements:

Concerning Section V, we encourage NIH to require data management plans to be prepared and submitted with the original application for funding.

Section VI: Data Management and Sharing Plans:

Concerning Section VI, we discourage NIH from requiring data management plans only as part of the 'Just in Time' materials provided after the original application was submitted. The reason why we discourage this approach is because excellent data management plans that meet FAIR principles require a thoughtful approach that builds on, and promotes, research excellence throughout the study. Requiring the plan as 'Just in Time', may result in a rushed plan, and could be a lost opportunity for excellence.

Section VII: Compliance and Enforcement:

no comment

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Concerning the allowable costs for a data management and sharing plan, we support the emphasis on incorporating data management and sharing costs into the budget. The policy does a good job highlighting some of the potential costs, including curation.

If the data management plan will be submitted as 'Just in Time', then NIH should allow the proposed budget to be edited to account for the time and expense in completing it. The cost of preparing the data management plan is non-trivial and therefore we recommend the preparation of the data management plan, and the preparation of the budget for maintaining the plan, be required as part of the original submission, even if it is incorporated in a "Just in Time" process.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Concerning the elements of a NIH data management and sharing plan, the Supplemental Draft Guidance: Elements of a NIH Data Management Plan could be improved by:

p.1, under Data Type: add emphasis on using open data formats or converting data from proprietary formats to open formats when they exist, which will facilitate preservation and access.

p.2, "Standards": This section would be improved by stating that: "if no appropriate data standards exist, then alternatives or community best practices should be described,"

p.2: "... Timelines", bullet point 2: NIH should encourage depositing data in a repository that can provide persistent, unique identifiers (e.g., DOIs).

Other Considerations Relevant to this DRAFT Policy Proposal:

Concerning other considerations, we encourage NIH (and/or ICOs) to work with stakeholders and others within the community of interest, including the data management and curation community, to identify and share the qualities and factors of preferred repositories. There are some 'preferred' repositories for some research domains and data types (e.g., NIH provides a list of NCBI repositories, and there are many others, such as ICPSR/OpenICPSR for social science data, QDR for qualitative data). That said, institutional data repositories should remain eligible, because institutional repositories, and data curators and repository managers, can facilitate data sharing that is aligned with FAIR principles.

Lastly, we encourage NIH to include a statement in the policy indicating an expectation that data must be shared within a year, or two, at most, of the completion of the project, or upon publication of results, whichever is sooner.

Attachment:

KHornbuckle_Westra_comments_NIHdraftdatapolicy_12-12-2019.pdf

Description:

signed letter of comment



December 12, 2019

National Institutes of Health

To whom it may concern,

We are writing to comment on the draft NIH Policy for Data Management and Sharing¹. We are part of the Iowa Superfund Research Program, a multidisciplinary research center funded by the Superfund Research Program office of the National Institute for Environmental Health Sciences (SRP/NIEHS). Hornbuckle is the Director of the P42 center and Westra is a data services librarian). We have worked together to prepare data management and sharing plans as part of our application for a competitive renewal of the center.

Concerning Section I, we strongly support NIH's plan to require data management plans: "Under this Policy, individuals and entities would be required to provide a Data Management and Sharing Plan (Plan) describing how scientific data will be managed, including when and where the scientific data will be preserved and shared, prior to initiating the research study. Shared data should be made accessible in a timely manner for use by the research community and the broader public."¹

Concerning Section IV, we are concerned about when and under what funding circumstances the policy would be effective.

The policy should only apply to requests for applications (RFAs) released after the final policy is announced so that grantees and their institutions can develop and provide training/instruction, resources, consultations, and other services/infrastructure. The policy should not apply to ongoing grants that have already been budgeted for the work described in the original application. Preparation of data management plans takes time and must be developed in coordination with the study being proposed.

Should the NIH decide to require data management plans from ongoing funded grants, the expectation for completion and scope cannot be the same as that required of new applications. The expectations must be proportional to the funding provided for the activity, and the additional effort required to develop a plan after the funded study has been designed and is being carried out.

Concerning Section V, we encourage NIH to require data management plans to be prepared and submitted with the original application for funding.

Concerning Section VI, we discourage NIH from requiring data management plans only as part of the 'Just in Time' materials provided after the original application was submitted. The reason why we discourage this approach is because excellent data management plans that meet FAIR principles require a thoughtful

¹ <https://osp.od.nih.gov/draft-data-sharing-and-management/>

approach that builds on, and promotes, research excellence throughout the study. Requiring the plan as 'Just in Time', may result in a rushed plan, and could be a lost opportunity for excellence.

Concerning the allowable costs for a data management and sharing plan, we support the emphasis on incorporating data management and sharing costs into the budget. The policy does a good job highlighting some of the potential costs, including curation.

If the data management plan will be submitted as 'Just in Time', then NIH should allow the proposed budget to be edited to account for the time and expense in completing it. The cost of preparing the data management plan is non-trivial and therefore we recommend the preparation of the data management plan, and the preparation of the budget for maintaining the plan, be required as part of the original submission, even if it is incorporated in a "Just in Time" process.

Concerning the elements of a NIH data management and sharing plan, the Supplemental Draft Guidance: Elements of a NIH Data Management Plan could be improved by:

p.1, under Data Type: add emphasis on using open data formats or converting data from proprietary formats to open formats when they exist, which will facilitate preservation and access.

p.2, "Standards": This section would be improved by stating that: "if no appropriate data standards exist, then alternatives or community best practices should be described,"

p.2: "... Timelines", bullet point 2: NIH should encourage depositing data in a repository that can provide persistent, unique identifiers (e.g., DOIs).

Concerning other considerations, we encourage NIH (and/or ICOs) to work with stakeholders and others within the community of interest, including the data management and curation community, to identify and share the qualities and factors of preferred repositories. There are some 'preferred' repositories for some research domains and data types (e.g., NIH provides a list of NCBI repositories, and there are many others, such as ICPSR/OpenICPSR for social science data, QDR for qualitative data). That said, institutional data repositories should remain eligible, because institutional repositories, and data curators and repository managers, can facilitate data sharing that is aligned with FAIR principles.

Lastly, we encourage NIH to include a statement in the policy indicating an expectation that data must be shared within a year, or two, at most, of the completion of the project, or upon publication of results, whichever is sooner.

Sincerely,



Keri C. Hornbuckle, Ph.D.
Donald E. Bently Professor of Engineering
Director, [Iowa Superfund Research Program](#)



Brian Westra, M.S., M.S.I.
Data Services Librarian

Submission ID: 1272

Date: 12/13/2019

Name: Hunter Moseley

Name of Organization: University of Kentucky

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

systems biology and bioinformatics

DRAFT NIH Policy for Data Management and Sharing

Section II: Definitions:

Would recommend that validation as part of data management be better explained. Should describe validation as well as quality assessment and quality control as essential elements of data management.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

While the definition of data management in Section II includes "validation", there is no mention of validation or any quality assessment description in the Elements of an NIH Data Management and Sharing Plan.

Submission ID: 1273

Date: 12/14/19

Name: Mary Janevic

Name of Organization: University of Michigan

Type of Data of Primary Interest: Clinical

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

behavioral medicine

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

I would suggest that exemptions be allowed for some audiorecordings and transcripts of qualitative data (e.g., interviews, focus groups), as this is often hard to completely de-identify.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Researchers will need easily accessible, affordable repositories for their data. Right now repository options can be difficult to identify and it is not always clear how to use them or exactly what form the data and supporting information should be in. Along with the new Data Sharing policy, there should be a parallel effort to make storing one's data as easy and efficient as possible. NIH should consider providing repositories where its funded researchers can easily store their data, along with guidance for what kind of supporting documentation will make the data maximally usable to others.

Submission ID: 1274

Date: 12/16/19

Name: Julie Lima, Vincent Mor, Faye Dvorchak, Roe Gutman

Name of Organization: Brown University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: clinical and claims-based administrative data

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

health services/policy

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Thank you for the opportunity to respond. We reviewed the draft policy specifically as it pertains to health services as well as social and behavioral research. We support NIH's longstanding commitment, as stated in the draft policy, to making the results and outputs of the research that it funds and conducts available to the public. What can and should be shared, as it concerns individual level data, however, must be more carefully considered in concert with individual privacy concerns, technology advancements, pragmatic clinical trial designs, and regulatory requirements within the informed consent process. This is particularly critical as funding quality non-pharmacological research initiatives focusing on persons living with dementia (PLWD) has been a recent priority for NIH. Though not formally labeled a vulnerable population within federal regulations, studies involving PLWD represent unique challenges that must be considered when developing an appropriate study design and data sharing plan. These challenges are enhanced within pragmatic study designs such as cluster-randomized trials. We offer the following comments on the current draft policy with these challenges in mind, but they are relevant to a much broader array of studies as well.

Section VI: Data Management and Sharing Plans:

While it is understood that the draft policy purposely allows for flexibility across various scientific domains and is intended only to establish minimum expectations for NIH-wide data management and sharing plans, we focus on two statements in Section VI. Data Management and Sharing Plans that are overly vague and consequently could introduce potential problems.

NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public.

Who is the arbiter of what is useful to the research community or the public for a particular study? Is it the PI of the study, the federal project officer, the study subjects themselves, or perhaps an individual or groups more removed from the immediate research? Will the designation of this arbiter be flexible across studies or only across scientific disciplines? If deemed useful to the research community or public, will it ever be to the benefit of one but at the expense of the other? How will this be weighed and by whom?

NIH recognizes that certain factors (e.g., legal, ethical, technical) may limit the ability to preserve and share data.

The importance of considering what these factors are not only within disciplines and broad data types (PHI received from covered entities, e.g.) but also as a function of study design and population type should not be understated. What are the ethical considerations, for example, of sharing primary or secondary data elements collected about PLWD under a waiver of informed consent or through a legally authorized representative? Or, how do regulatory requirements affect the willingness of an individual to participate in a study if they must also consent to the potential disclosure of their data for future use? Per the Revised Common Rule, studies that involve the collection of identifiable data through the use of an informed consent document must include one of the two following statements (45 CFR 46.116 (b)) in the informed consent document -

- o (i) A statement that identifiers might be removed from the identifiable private information or identifiable biospecimens and that, after such removal, the information or biospecimens could be used for future research studies or distributed to another investigator for future research studies without additional informed consent from the subject or the legally authorized representative, if this might be a possibility; or
- o (ii) A statement that the subject's information or biospecimens collected as part of the research, even if identifiers are removed, will not be used or distributed for future research studies.

The Revised Common Rule also allows researchers the opportunity to obtain broad consent from study subjects that would allow subjects to provide or deny consent for the storage and future use of their identifiable data collected under a given study. Data for subjects who refuse consent for this future storage and use would then have to be removed and any future use of the data would not be subject to a waiver of informed consent.

Currently, the ability to make the data available in a de-identified or identifiable form (through broad consent) are options, not requirements, within the regulations. It is unclear whether this draft policy regarding data management and sharing will in effect make them requirements for NIH-funded research, at least within some divisions of NIH. The draft policy states that "NIH encourages the broadest use of scientific data resulting from NIH-funded or conducted research, consistent with privacy, security, informed consent, and proprietary issues" (p. 60402; FR 84). It does nothing to protect an investigator's judgment that the integrity of a study may be compromised if future data sharing must be guaranteed in advance. The effect that these added statements and consent procedures might have on study response rates and the resulting representativeness of study samples remains unclear. For studies involving PLWD, this is particularly worrisome as the process of consenting persons with cognitive impairments directly or through legally authorized representatives is already challenging.

Regardless of the above concerns, whether a datafile can be made sufficiently de-identified for general release is also of increasing importance as our technical and statistical capabilities increase alongside an ever-increasing volume of data collected and available for merger with potentially identifiable data. A recent National Academies of Sciences, Engineering and Medicine Committee posited that the identifiability of data is dynamic, and what might be considered de-identified today may soon be identifiable through new techniques (1). In fact, any data release has the potential to reveal information about individuals. The only way to truly protect individual level data is to reveal no information at all. Any individual privacy breach involves being able to learn about an individual in a dataset(1). When releasing individual level data, even if the released data does not contain any patient identifying information, its linkage with other sources may result in a privacy breach(1, 2). This is because linked information have more data on individuals than each file by itself(3). For example, a dataset may include an indicator that patients received care within the same facility without releasing any information about the facility. This information is valuable in identifying individuals in a different dataset because a linkage algorithm can now attempt to identify a set of individuals within a specific facility, instead of each individual across all facilities. Similar examples led the aforementioned National Academies of Science, Engineering and Medicine committee to recommend that federal statistical agencies develop and implement strategies to safeguard privacy while increasing accessibility to linked datasets for statistical purposes(1). In our opinion, neither the Safe Harbor nor the Expert Determination method of de-identification provided by the Health Insurance Portability and Accountability Act (HIPAA) Privacy rule, for example, are sufficiently

specific or rigorous to expect reasonable de-identification of protected health information (see <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>, accessed Dec 3, 2019).

References

1. In: Harris-Kojetin BA, Groves RM, editors. Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps. Washington (DC)2017.
2. Dwork C, Naor M. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*. 2010;2(1).
3. Fellegi IP, editor. Record linkage and public policy—a dynamic evolution. Proceedings of the International Workshop and Exposition, Federal Committee on Statistical Methodology, Office of Management and Budget; 1997: Citeseer.

Submission ID: 1275

Date: 12/16/19

Name: Toni Harbaugh

Name of Organization: NCI/Frederick National Laboratory for Cancer Research

Type of Data of Primary Interest: Genomic

Type of Organization: Government Agency

Role: Other

Role - Other: Server and Storage Architect

Domain of Research Most Important to You or Your Organization:

cancer research

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

In the first bullet point under section 1 'Data Type', the statement of size should include an estimated 'digital footprint' in commonly-used storage units ('Gigabytes', 'Terabytes', etc.). Repository cost estimates will require this value, and for NIH-internal storage resources the estimate provides necessary budget guidance.

Other Considerations Relevant to this DRAFT Policy Proposal:

If NIH-internal storage resources will be encumbered, the data management plan should be provided in advance to the IT personnel responsible for those resources as part of the initial request. The document should also be stored online, searchable by NIH IT support.

Submission ID: 1276

Date: 12/16/19

Name: Jennifer DeBerg

Name of Organization: University of Iowa

Type of Data of Primary Interest: Clinical

Type of Organization: University

Role: Other

Role - Other: health sciences librarian

Domain of Research Most Important to You or Your Organization:

orthopedics, nursing, audiology and speech pathology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Fully support the purpose. I find this section clear and compelling. No suggestions for improvement.

Section II: Definitions:

Definitions are well written and clear-- no changes suggested

Section III: Scope:

No changes suggested

Section IV: Effective Date(s):

No changes suggested to actual policy--

I have been struggling to find a target date completion of the policy and think that should be noted

Section V: Requirements:

No changes suggested

Section VI: Data Management and Sharing Plans:

No changes suggested

Section VII: Compliance and Enforcement:

No changes suggested

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

No changes suggested.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

No changes suggested.

Other Considerations Relevant to this DRAFT Policy Proposal:

To support the policy and to indicate that sharing is a priority, a centralized data repository should be further considered. Though I am not a data management specialist, I can report that there are many disciplines for which there is not an appropriate discipline specific repository and I believe there is a need

Submission ID: 1277

Date: 12/16/19

Name: Everett Carpenter

Name of Organization: HHS-NIH-NCI-DCCPS

Type of Data of Primary Interest: Other

Type of Organization: Government Agency

Role: Government Official

Domain of Research Most Important to You or Your Organization:

HHS-NIH-NCI-DCCPS

DRAFT NIH Policy for Data Management and Sharing

Section VI: Data Management and Sharing Plans:

Suggestion for the contract bullet

- Contracts: Statement of Work (SOWs) will require Plans. Plans will be submitted with Proposals and will be evaluated by NIH staff or NIH ICO as part of the over Technical Evaluation Panel as part of the Source Selection process.

Submission ID: 1278

Date: 12/16/19

Name: Ho Jung Yoo

Name of Organization: University of California San Diego

Type of Data of Primary Interest: Other

Type of Data of Primary InterestO-ther: General, all disciplines

Type of Organization: University

Role: Other

RoleO-ther: Data curation

Domain of Research Most Important to You or Your Organization:

General

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

As a curator of research data at an institution-based repository, I very much welcome this NIH data management and sharing policy. When the policy is in effect, how will curated repositories be able to keep up with the increased submission rates? Aside from allowable costs for project-associated data management, permanent curation staffing is limited at most universities.

Other Considerations Relevant to this DRAFT Policy Proposal:

During the webinar, you mentioned that NIH is thinking through ways to track data reuse, as a way of evaluating the effectiveness of the policy. Two of the challenges for good data reuse tracking are 1) the need for establishing community standards in the way related resources or publications are cited, and 2) for repositories and publishers to adopt those standard practices as well as to provide guidelines for authors to follow them as well. Currently, the closest thing to a standard that I'm familiar with is in the DataCite Metadata Schema (i.e., the relatedIdentifier and relationType properties). Most repositories and publishers don't accommodate relationType for citing identifiers related to the dataset being deposited (Zenodo and Dryad are two of the exceptions). So, it's likely that we're not capturing a lot of information about how data was derived from other sources, at the time of deposit. (Is the identifier a source dataset? Is it a previously published article that this work is referencing? Is it a publication that reports on the analysis of this dataset?) As a funding agency, NIH may be in a good position to recommend that proper data citation practices are followed by both repositories and investigators. Good data citation is essential to good data reuse tracking in the future. This will help with NIH's evaluation of both policy effectiveness and plan compliance.

Submission ID: 1279

Date: 12/17/19

Name: Lucia Peixoto

Name of Organization: Washington State University

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Neuroscience, genomics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

There should be a presumption that all research data underlying a publication is shared at time of publication. The current language is weak and has statements such as "shared data should be made accessible" or "not all data generated in the course of research may be necessary to validate and replicate research findings." Instead the policy should say that shared data **MUST** be made accessible, except when justified by a small number of reasons, such as participant privacy concerns that cannot be overcome by protective measures, or studies on vulnerable populations.

The draft lists an expectation of "timely" data sharing. This is a vague and unacceptable term. The release of data should follow the recommendation of the Office of Research Integrity of the HHS as follows: After a project's research has been published or patented, any information related to the project should be considered open data unless it violates the HIPAA privacy rule

Submission ID: 1280

Date: 12/17/19

Name: Fred Oswald

Name of Organization: Rice University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: psychological measurement, organizational

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

employment testing, college admissions, workforce readiness, psychological testing

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

As a general comment: Reproducibility of published results requires much more than data-sharing; it requires knowing the decisions and procedures for transforming the raw data into analyzable data (e.g., merging data, dealing with miscoded or redundant data, dealing with different types of missing data, deleting or down-weighting outliers, potentially recoding data based on the research question) and knowing how the analyses themselves were conducted (e.g., choices made between appropriate data analysis methods, the defaults and estimation methods of software packages, the decisions for conducting follow-up exploratory analyses). Data transformation and analysis choices are consequential if the central goal is reproducibility - but these only receive light/indirect attention in section 2 of the plan. A central question here is whether data sharing is sufficient if researchers provide their transformed/cleaned data to which the supplied program code is applied (because if that is done, then one is assuming the transformed/clean data are correct - it cannot be verified/reproduced from the raw/original data).

Bottom of page 1 - change 'guidance' to 'professional guidance'

Submission ID: 1281

Date: 12/18/2019

Name: Christian Murray

Name of Organization: Murtek Systems

Type of Data of Primary Interest: Clinical

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Chronic diseases

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

To enable the most flexible data consumption with the least effort by data providers, all data shall be provided via GraphQL APIs which allows semantic and automated querying of the data using the embedded relational model.

Comply with NIH data standards guidelines: <https://www.ncbi.nlm.nih.gov/books/NBK216088/>

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Data providers shall spend at least 5% of the project budget.

NIH will provide a list of vetted, knowledgeable consultants.

NIH will provide infrastructure and indefinite hosting on a private AWS account.

Submission ID: 1282

Date: 12/19/2019

Name: Pam Dixon

Name of Organization: World Privacy Forum

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: human subject research data

Type of Organization: Nonprofit Research Organization

Role: Other

Role - Other: Executive Director

Domain of Research Most Important to You or Your Organization:

data privacy and governance policies regarding human subject research, protections for subjects

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

See attached PDF

Section II: Definitions:

See attached PDF

Section III: Scope:

See attached PDF

Section IV: Effective Date(s):

Section V: Requirements:

See attached PDF

Section VI: Data Management and Sharing Plans:

See attached PDF

Section VII: Compliance and Enforcement:

See attached PDF

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing

Plan: See attached PDF

Other Considerations Relevant to this DRAFT Policy Proposal:

Please see attached PDF

Attachment:

WPF_comments_NIH_DraftGuidance_Research_Privacy_Dec2019_fs.pdf

Description:

Comments of World Privacy Forum to NIH re DRAFT Policy Proposal



WORLD **PRIVACY** FORUM

Comments of the World Privacy Forum

To

National Institutes of Health

Regarding Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Comments submitted via: <https://osp.od.nih.gov/draft-data-sharing-and-management>

Andrea Jackson-Dipina, Dr.PH
Director of the Division of Scientific Data Sharing Policy
Office of Science Policy, NIH, 6705 Rockledge Drive
Suite 750
Bethesda, MD 20892

December 18, 2019

The World Privacy Forum is pleased to have the opportunity respond to the request of the National Institutes of Health for public comments on a draft NIH policy for data management and sharing and supplemental draft guidance. The request appeared in the Federal Register on November 8, 2019, 84 Fed. Reg. 60398, <https://www.federalregister.gov/documents/2019/11/08/2019-24529/request-for-public-comments-on-a-draft-nih-policy-for-data-management-and-sharing-and-supplemental>.

The World Privacy Forum is a nonprofit, non-partisan 501(c)(3) public interest research group. The WPF focuses on privacy, with health privacy among our central focuses. We publish and maintain a large body of health privacy work, including a patient's guide to HIPAA, reports on and FAQs for victims of medical identity theft; and reports on genetic privacy, precision medicine, electronic health records, and other topics. We regularly testify before Congress and federal agencies, and we submit comments on HIPAA and other regulations with relevance to privacy and security. More about our work and our reports, data visualizations, testimony, consumer guides, and public comments can be found at <http://www.worldprivacyforum.org>.

The World Privacy Forum supports the broad goals of the draft policy. Responsible research and the appropriate sharing of research data are worthy objectives. Our comments address the policy as it relates to the privacy and security of data about human subjects.

We preface our suggestions with a few observations about researchers. While there are many responsible researchers, we find that too many researchers want everyone's data but are unwilling to accept or implement the level of responsibility required to provide meaningful privacy and security for this personal data. At present, we must accept the Institutional Review Board process as it is today. However, we also note there are meaningful gaps in protections even when institutional review boards oversee research activities.

Privacy and security are only sometimes adequately addressed in the IRB process because only some IRBs have the knowledge, motivation, and interest to require that research projects maintain proper privacy and security protections. A factor in this difficulty is that IRB members only occasionally have the needed expertise in privacy or security. We recognize that this is not the place to address generally the shortcomings of the IRB process. We also recognize that the IRB process is an area that would benefit from increased policy attention, particularly as it relates to human subject research, including research that is incorporating AI and machine learning aspects.

Some cities are undertaking innovative work on IRBs, for example, Columbus, OH has a community IRB process. See: <https://orpp.osu.edu/irb/research-participants/community-engaged-research/>. The city of Cambridge, MA has an open data review board, <https://data.cambridgema.gov/General-Government/Cambridge-Open-Data-Ordinance-092115/tf4d-q3qs>. We hope that the smaller efforts seeking to update IRB processes will continue, and will spark larger scale projects updating IRBs.

In the meantime, the point is that no one can assume that existing mechanisms (like IRBs) or that the researchers themselves can guarantee suitable protections for privacy and security. Thus, casual references to privacy and security in a summary list of requirements for data management and sharing in the guidance is not likely to make a meaningful difference. Much health research data in the hands of researchers is not subject to the privacy or security rules in HIPAA. **Indeed, most research data about individuals is not subject to any existing privacy law in the United States.** This contrasts with the situation in the European Union and much of the rest of the world, where researchers are generally be subject to the same data protection rules as others who process personal data.

NIH is one of the few institutions that has the clout to impose more specific privacy and security obligations on researchers. We do not suggest, however, that NIH use the proposed guidance to promulgate privacy and security regulations on those who receive NIH funding. Still, NIH can do better than a few casual references to privacy and security.

For example, we suggest that guidance include specific references to current NIST security guidance and to HIPAA security standards. Telling researchers that they must address security is one thing. Telling researchers that their security measures must be as rigorous as those from specifically identified and generally authoritative sources is more likely to be noticed and to

result in a reasonable level of security. We observe that the supplementary information for the draft guidance includes a specific reference to several NIH genome policy documents and to other NIH policy materials. We suggest something similar here. More references to appropriate security documents – especially ones that the authors of those documents keep up to date – would make the guidelines more useful to data users and more beneficial to data subjects. Referencing standards would also help during project evaluation.

We make the same suggestions for privacy. NIH documents and standards like the Common Rule are filled with vague and general references to privacy and confidentiality in research. All lack meaningful standards to tell researchers what they should do. Recent revisions to the Common Rule failed to include the more specific privacy and security obligations for researchers that the draft rule proposed. NIH should point to specific privacy policies used in existing research as models for everyone to follow. Telling researchers that research projects will be evaluated by NIH in part on the basis of specific privacy and security standards has the potential to make a difference.

We make the same suggestion yet again for data de-identification obligations. Numerous organizations maintain best practices for data de-identification. NIH should select several as examples, choosing those where the authors keep the documents up to date. De-identification is a much more prominent legislative issue now at the state level in the US and globally. We expect that researchers need to be much more aware going forward about what the proper standards are for de-identification in various research contexts. We note that privacy-related laws drafted or passed the last few years introduced more precise language and requirements around de-identification. See, for example, various state level laws in the US, including the CCPA, and see for example, the GDPR in Europe, and most recently, India's Data Protection Bill 2019.

Where the NIH draft guidance addresses data sharing agreements, we think that providing references to sample agreements would be valuable. Providing a wide range of models would be much more effective than a vague admonition to do *something* appropriate with respect to privacy and security.

Further, we suggest that NIH require – or at the very least, suggest – that all data sharing agreements expressly include language stating that data subjects are third party beneficiaries of the agreement. Unless data subjects have the ability to enforce the privacy and security requirements of a data sharing agreement when and if the need arises, violations of an agreement will never be pursued because the parties to the agreement will likely have no interest in doing so. We do not suggest a third-party beneficiary clause as a cure-all, but it will offer a possible enforcement tool that would otherwise be absent.

In closing, we appreciate that NIH's draft guidance focuses on the importance of privacy and security in data sharing. We observe, however, that the draft's reference to protective measures “that are consistent with applicable federal, tribal, state, and local laws, regulations, statues [sic], guidance, and institutional policies” has little meaning as the research world largely falls outside of any privacy and security rules in the US.

Further, even this broad suggestion to comply with “applicable” rules is inadequate. Much research is international in scope/ The conduct of research and the international transfer and location of research data about individuals implicates data protection rules in other countries. Nearly every other country in the world has generally applicable data protection rules, and the United States is the only major outlier. NIH should use its guidance to tell U.S. researchers that they need to be aware of the consequences of international activities.

NIH still has a narrow band of time in which it can be proactive regarding privacy, noting that privacy legislation is under active consideration in the Congress and in other countries as well. To proactively address privacy issues, NIH needs to do more than it proposes in its guidance to properly advise the research community and to protect data subjects.

Incorporating more specific and relevant guidance is an important starting point. We are only a front-page scandal away from the imposition of new state and federal laws that would provide the type of privacy and security protections now lacking in the research world in the US. NIH needs to step up and do its part to provide more meaningful and more useful guidance for the research community before someone else does. The health sector, including the kinds of research the NIH draft guidance seeks to address, is enormously complex. Legislation can sometimes be a blunt tool that does not acknowledge those complexities. We urge NIH to be as proactive as possible in its guidance.

Thank you for the opportunity to comment on the draft guidance.

Respectfully submitted,

s/

Pam Dixon
Executive Director, World Privacy Forum
www.worldprivacyforum.org
3 Monroe Parkway, Suite P #148
Lake Oswego, OR 97035

Submission ID: 1283

Date: 12/19/2019

Name: Michael Hoffman

Name of Organization: Princess Margaret Cancer Centre

Type of Data of Primary Interest: Genomic

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

raw sequencing reads from genomic assays, and processed versions thereof

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

There should be a presumption that all research data underlying a publication is shared at time of publication. The current language is weak and has statements such as "shared data should be made accessible" or "not all data generated in the course of research may be necessary to validate and replicate research findings." Instead the policy should say that shared data **MUST** be made accessible, except when justified by a small number of reasons, such as participant privacy concerns that cannot be overcome by protective measures, or studies on vulnerable populations.

The draft lists an expectation of "timely" data sharing. This should be defined as generally at the time of publication. Funding opportunities specifically designated to create a shared resource should specify a date by which data must be available even in the absence of a publication. This aspect is a step backwards from previous NIH policy which clearly defines "timely" as "no later than the acceptance for publication of the main findings from the final data set." The relaxation of this existing requirement is not justified.

Section II: Definitions:

Should include definitions of FAIR data and the 15 FAIR principles.

Section III: Scope:

Scope should make clear that the policy continues to apply for scientific data produced by funding in whole or in part from NIH after the NIH funding period is over.

Section IV: Effective Date(s):

The current absence of an effective data management and sharing policy and lack of enforcement causes a serious negative impact on health research and enables an ongoing waste of public funds. The noncommittal implementation date of the draft is unacceptable. The final policy should have a "no later than" date for implementation, ideally 12 months after issuance of the final policy.

Section V: Requirements:

To ensure good data management, any data described as collected in a progress report must be deposited independently and an accession code or digital object identifier (DOI) supplied. Except when specified by the funding opportunity announcement, researchers may embargo this data until publication. Grant opportunities specifically designated to create a shared resource should specify a date by which data must be available even in the absence of a publication.

It should be clear that these requirements apply not just to research project grants and contracts, but most other forms of requests for support that will lead to the creation of scientific data. This includes cooperative agreements, career grants, fellowships, scholarships, and training grants.

Absent a compelling reason otherwise, contract solicitations should specify that collected data is the property of NIH. They should also include specific requirements that data should be made publicly available in a third-party repository as a periodic deliverable, upon which further funding can be conditioned.

There are a large number of digital repositories with different policies. You should require that acceptable digital repositories must not allow recipients to unilaterally change or delete deposited data. The repositories, may, however, allow adding new versions of data advertised in metadata for the original dataset.

It is important to protect human participant privacy but it is also important that concerns about human participant privacy not be abused to eliminate appropriate data sharing. It is especially worth considering that many human participants expect that data from their participation will

be shared with other qualified researchers. Ineffective sharing of the resulting data (assuming appropriate protective measures such as de-identification are in place) is unethical as it wastes human participants' contributions to research and may result in more patients being exposed to harm. Therefore it should be an explicit goal of this policy and any submitted Data Management and Sharing Plans to maximize access subject to necessary restrictions.

Section VI: Data Management and Sharing Plans:

The draft states that NIH encourages scientific data to be made available. Instead, it should REQUIRE that scientific data are shared.

An effective Data Management and Sharing Plan should increase the overall impact of a grant and an ineffective one will decrease it. It is important that Data Management and Sharing Plans be provided to NIH peer reviewers and ICO advisory council review so they can consider the plan's effect on the application's overall impact, significance, and approach. Guidance to reviewers on how to score review criteria such as significance and approach should include review of the Data Management and Sharing Plan.

Therefore, NIH should require Data Management and Sharing Plans at the regular submission due date for an application, and not as a Just-in-Time submission. Overcoming deficiencies in the Data Management and Sharing Plan identified in summary statements could be provided as a Just-in-Time submission.

NIH should require that data management plans must describe how the researchers address each of the 15 FAIR Principles.

NIH should publish data management plans for funded grants and contracts alongside abstracts in public databases such as RePORTER. This will increase transparency and let other researchers and the public know what the grantees promised to NIH. This is the only thing that will make enforcement of individual plan items possible, given that NIH does not have the resources for exhaustive, systematic checks on compliance. Grantees knowing that their data management and sharing promises are readily available to the public will provide some measure of self-enforcement. Currently data sharing plans are available through Freedom of Information Act requests, and putting them on RePORTER will reduce the burden on data requesters.

The draft says that only data "deemed useful to the research community or the public" need be shared. It should be clear that applicants do not get to unilaterally decide what data is deemed useful. Any exceptions to the general principle that scientific data must be shared must be justified and funding conditioned on prior approval by an NIH advisory committee of data management experts that includes data scientists and librarians.

For intramural research, you should not give a single NIH official (such as Scientific Director or Clinical Director) the ability to assess Data Management and Sharing Plans without oversight. Data Management and Sharing Plans must be reviewed and approved by Boards of Scientific Counselors and ICO advisory councils during the existing periodic peer review and site visit process.

Section VII: Compliance and Enforcement:

It is currently unclear where to turn when NIH data sharing expectations and policies are not followed. To solve this, RePORTER should list, for each grant, contact information to request corrective action for violations of the Data Management and Sharing policy or published Data Management and Sharing plans. This should include contact email addresses for the principal investigators/project directors of the grant, contact email addresses for officials representing the grantee institution, and a contact email address at NIH. That will allow for solving issues at the most local level, when possible, and escalation when the previous proves ineffective. Similar information should be available for contracts and for intramural research projects.

In addition to reviewing progress reports and addressing complaints, NIH ICOs should also perform more thorough random audits to ensure grantees are performing data management as expected.

Current sanctions listed in the draft policy are incredibly weak and will have no deterrent effect. The policy should mention that failure to follow the Data Management and Sharing policy can be considered research misconduct by NIH. The policy should specify that violating the policy in place at the time of competing award at any time thereafter (including after the end of the award period) can result in sanctions. These sanctions can include publication of a notice describing the violation in the NIH Guide to Grants and Contracts, debarment and suspension from contracting, subcontracting, or financial assistance from the federal government, and prohibition of service to the Public Health Service on advisory committees, boards, or peer review committees, or as a consultant. Because it touches on potential research misconduct, this policy must be reviewed by the HHS Office of Research Integrity.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The guidance should specify that fees that preserve data beyond the funding period are allowed, as are personnel expenses related to data sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

An entry of "to be determined" in a Plan is not acceptable. This language will encourage useless Plans and should be removed.

Statements like "NIH does not expect researchers to share all scientific data generated in a study" defeat the purpose of this policy. Instead NIH should make clear that they do expect and require sharing of scientific data except in limited exceptions, justified by the applicant, and prior approval by peer reviewers, program staff, and an NIH advisory committee of data management experts that includes data scientists and librarians.

Section 1 describes "consistency with community practices" as a potential rationale for deciding which data are preserved and shared. In many scientific disciplines, community practices lag far behind general best practices and what the public expects for data management and sharing. This language allows certain communities to settle for mediocrity in data management and sharing, defeats the aim of this policy to improve data management and sharing. It should be removed. This also illustrates why decisions to withhold scientific data from sharing should not only be reviewed by study section members trained in the same discipline but also an NIH advisory committee of data management experts that includes data scientists and librarians.

Section 4 says that "if an existing data repository(ies) will not be used, consider indicating why not". This policy should require the use of established repositories, except when exceptions are justified and approved. It should not be up to applicants to unilaterally decide not to use standard established repositories and to not even justify the same.

Section 5 anticipates that applicants may have restrictions on sharing imposed by existing or future agreements. This provides a major loophole in the policy in that applicants may choose to enter into more restrictive agreements than necessary so that they can avoid data sharing. This can be overcome by (1) providing data sharing plans as part of initial peer-review so that peer reviewers can appropriately score any decrease in impact that may come about from restrictions on sharing, and (2) review by an NIH advisory committee that includes data scientists and librarians.

Other Considerations Relevant to this DRAFT Policy Proposal:

I applaud your efforts to establish an excellent research data management and sharing policy. As written, I do not think this policy will provide a substantive change in data sharing. To maximize the benefit to the public of providing research funds, it is essential that the policy and enforcement be strengthened as described in this response.

In general, the draft policy is overly cautious and fails to consider the burden an ineffective policy will place on researchers who seek to use shared NIH-funded scientific data. The current system is incredibly burdensome on those seeking to obtain shared data because when data are not available as per existing NIH expectations, investigators can stonewall requests. There is no enforcement and the way to request enforcement is unclear. My most serious concern about this policy is that it is too vague on requirements in some places and lacks sufficient detail on enforcement.

A policy with ineffective, vague requirements and no real enforcement will have a serious negative impact on researchers who seek to use scientific data produced with public funds. There is a huge waste of researcher time and money attempting to obtain data that is lost, improperly described, or withheld. Failure to follow good data management practices leads to great inefficiency and slows the work of many researchers. There is also a large impact on our research communities, which lose opportunities to aggregate data and create a whole that is greater than the sum of its parts.

It is good to have both requirements and incentives to encourage high-quality data management. I suggest that an "Incentives for High-Quality Data Management and Sharing" section be added to the policy, including the following incentives:

1. Add to the NIH biosketch a section for key personnel to describe their most significant contributions to data management and resource sharing (including data, code, reagents, samples, and other materials). This should be separate from other contributions to avoid it getting short shrift due to lack of space. The past record of the principal investigator and other key personnel should be explicitly added to the scored review criteria.
2. NIH should create awards to recognize and cultivate excellence in data management and resource sharing, both at the individual researcher and institutional levels.

Submission ID: 1284

Date: 12/19/2019

Name: Susanna-Assunta Sansone

Name of Organization: FAIRsharing

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: all

Type of Organization: Other

Type of Organization - Other: Community-driven initiative

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

on behalf of the FAIRsharing Community: <https://fairsharing.org>

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The FAIRsharing Community (<https://fairsharing.org/communities>) welcome the opportunity to submit comments on the Draft NIH Policy for Data Management and Sharing. The stress on both data sharing and management is a very welcome addition.

We in FAIRsharing have long advocated for funders and journal publishers' data policies to ensure that datasets and other digital products associated with their articles are deposited and made accessible via the appropriate repositories, in line with the FAIR Principles and to support data stewardship and reproducibility.

Based at the University of Oxford in the UK, the FAIRsharing operational team works with an international Advisory Board (incl. Mike Huerta, NIH-NLM), stakeholders and adopters that formally endorse and recommend FAIRsharing (incl. ELIXIR, the Research Data Alliance, European Commission H2020, and major journal publishers) to provide a curated, informative and educational resource on data and metadata standards, inter-related to repositories and data policies. FAIRsharing guides consumers to discover, select and use these resources with confidence, and producers to make their resources more findable, more widely adopted and cited.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We suggest to add a reference to FAIRsharing (<https://fairsharing.org>) at page 2, under Standards. There are only few standards mentioned as an example. Pointing to FAIRsharing would help researchers to: find the right standards for the data and metadata type; understand which standard is implemented by which repositories, in order to format and annotated data and metadata for the deposition process; and be aware of which standards and repositories are recommended by the journals they wish to target to publish their work, and associated data and metadata.

FAIRsharing has also grouped the NIH-supported repositories (https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html) in this Collection <https://fairsharing.org/collection/NIHsupporteddatarepositories>; in FAIRsharing, the records describing the repositories show which standards, if any, these repositories implements, as well as their relationships (<https://fairsharing.org/graph/#/collection/bsg-c000002>), therefore providing information that a flat list cannot. Currently this FAIRsharing Collection has been created by the FAIRsharing team; however, as we do with other organizations (e.g. <https://fairsharing.org/recommendation/WellcomeOpenResearch>), the Collection should be maintained by a NIH officer, group, or list, and we welcome your advice on who the contact should be.

Lastly, we suggest to add the formal citation to the FAIR Principles:
<https://doi.org/10.1038/sdata.2016.18>

Submission ID: 1285

Date: 12/19/2019

Name: Bruce Stillman

Name of Organization: Cold Spring Harbor Laboratory

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: Nonprofit Research Organization

Role: Institutional Official

Attachment:

NIH-DATA MANAGEMENT AND SHARING request response_Dec2019.pdf

Description:

NIH-DATA MANAGEMENT AND SHARING request response



Cold Spring Harbor Laboratory

Bruce Stillman, PhD, FRS, FAA
President and Chief Executive Officer

William J. Matheson Professor
of Cancer Biology

December 19, 2019

Thank you for the opportunity to comment in response to the National Institutes of Health (NIH) Notice (NOT-OD-20-013) regarding the DRAFT NIH Policy for Data Management and Sharing and supplemental DRAFT guidance. The draft policy was shared widely at Cold Spring Harbor Laboratory (CSHL) and feedback was gathered from our scientific and administrative communities. On behalf of CSHL, I express support for the commitment of the NIH to making the results and accomplishments of the research it funds and conducts available to the public. I offer the following comments for consideration as you finalize the policy.

1. Having the policy call for the sharing of data in a “timely manner for use by the research community” is appropriate and should be fostered and encouraged. However, having the NIH policy and practice “in general” be that scientific data should be made available “independent of award period and publication schedule” is of concern. Consideration must be given to the timely review, analysis and curation of the specific data involved by those generating it, and the need to avoid creating a competitive disadvantage resulting from the early release of data prior to the completion of projects and publications. We agree that data sharing must be timely, however, the appropriate timing must be agreed upon between the investigators conducting the work and the NIH Institute, Center or Office (ICO), considering the time needed for proper data curation and publishing the results based on the science involved rather than administratively driven mandates. There should be an option that allows investigators to appeal ICO mandated data sharing requirements to the NIH Policy Office should the requirements be considered unreasonable or inappropriate by the Principal Investigator(s) involved, without fear of reprisal. This suggestion does not apply for large-scale genome sequencing projects where immediate data release has been the standard.
2. ICOs should be encouraged to use common formats and data standards whenever possible for collecting the necessary data and information that can be applied across ICOs for given types of data, specifying acceptable elements.
3. There should be a centralized location on the NIH Policy Office website where all additional ICO specific requirements for Data Management and Sharing are located so that this information is readily accessible in a single location to investigators rather than having them search each individual ICO websites.
4. We recommend funding mechanisms that utilize modular budgets be allowed to include additional modules, beyond the outdated direct cost limit of \$250k, to accommodate data management and sharing costs. This would incentivize early stage investigators who tend to utilize modular budgeting to include such costs without having to compromise supply and other cost needs. The existing NIH modular budget parameters established in 1998 have remained unchanged for 21 years and must be adjusted accordingly. We also recommend NIH require all ICOs to provide the flexibility of allowing additional data management costs to be added at JIT based on the final Program negotiated data management plan that may require additional support. This option should also be available at the time of progress reporting should additional costs for data management be required.

5. The supervisory and responsibility functions and expectations for this role should be further defined and appropriately supported. Will there be a requirement to certify expertise in data collection, analysis and submission, and what is the expectation for such oversight expertise?
6. The implementation of a Data Management and Sharing Plan must allow for adequate lead time. The “effective date” must provide sufficient time for all constituent parties to familiarize themselves and their teams with the requirements to effectively implement the Plan with the intended maximum benefit to the research community. We recommend that implementation be effective with new awards issued in NIH fiscal year 2021.

Sincerely,

A handwritten signature in blue ink that reads "Bruce Stillman". The signature is written in a cursive style with a horizontal line underneath the name.

Bruce Stillman

Submission ID: 1286

Date: 12/20/2019

Name: Data Services Team

Name of Organization: NYU Health Sciences Library

Type of Organization: University

Role: Other

Role - Other: Librarians

Domain of Research Most Important to You or Your Organization:

NYU Langone Health

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Although it is understandable why the policy doesn't provide a specific timeline for what constitutes a "timely manner" for shared data to be made accessible, more detailed information about what that phrase means would be helpful. In particular, if there is an expectation that data will be shared upon publication in cases where a publication results from the data, that should be noted as one marker.

Reference is made to data being preserved, but "preservation" needs to be defined. Many researchers will conflate storage with preservation, so the distinction should be clarified.

Section II: Definitions:

Definition of metadata: Rather than saying that metadata makes data more understandable, it would be clearer and more complete to say that metadata ensures that data can be discovered, analyzed, and interpreted.

Definition of scientific data: The exclusion of lab notebooks is confusing, as they often contain data and/or metadata.

Section VI: Data Management and Sharing Plans:

While it is clearly necessary to allow for researchers to follow laws, regulations, and policies in regards to sharing human subjects research data, there should be additional space that allows researchers to follow their own ethical compass. Recent studies on the re-identification of research data (including but not limited to MRI data) demonstrate that technology is advancing in a way that outpaces how regulators are dealing with the identifiability of research data. Researchers may become aware of risks to research subjects' privacy sooner than others. If there is not space within this data sharing policy for researchers to decide not to share data because of a personal concern about re-identification when that personal concern is not supported by a law or regulatory body, this policy may force researchers to share data that they think may lead to a violation of their research subjects' privacy.

It is unclear how, if the Plan is required as part of the Just-in-Time, the researcher will be able to properly budget for data management and sharing in their submission.

The statement that data should be made available "as long as it is deemed useful to the research community" is very vague, perhaps intentionally so, but it could be strengthened by at least some examples of how one might ascertain this.

It states that "NIH encourages the use of established repositories" without including any explanation of what that means. Again, some examples would be helpful. In addition, would NIH discourage a researcher from sharing their data through an institutional repository, even if recently established?

Submission ID: 1287

Date: 12/20/2019

Name: Michael McDonell

Name of Organization: Washington State University

Type of Data of Primary Interest: Clinical

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Sensitive information related to drug and alcohol use and mental health, data from American Indian and Alaska Native communities

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

I would strongly recommend that you specifically call out an exemption or recognition that studies conducted in partnership with American Indian and Alaska Native communities, particularly those conducted with tribal organizations or on Reservations are allowed to specify their data sharing as restrictive. Or if appropriate, no data sharing will be conducted if that is the request of our tribal partners. As you know many AI/AN communities have been harmed by the misuse of their data, including data gathered through NIH grants. As a person who partners with many Native communities on NIH funded alcohol and drug treatment research, I want to make sure that Native communities can dictate data sharing/access on their terms. I know that Native communities will not participate in NIH research if we do not allow them to determine how their data is use.

Submission ID: 1288

Date: 12/21/2019

Name: Casey Greene

Name of Organization: University of Pennsylvania

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

computational biology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The policy states that "Shared data should be made accessible in a timely manner...". Timely should be defined.

Section VI: Data Management and Sharing Plans:

This draft states: "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." NIH should require, not encourage, data to be shared. It is unclear who would be responsible for deeming data as useful.

In this policy, the Data Management Plan will be submitted as part of the Just-in-Time. This signals that the plan is not a valued part of the application and is in fact an afterthought. It should be required as part of the application so that appropriate sharing costs can be budgeted for at the time of application, and the plan can be included as part of the review process.

This section states that, "NIH may make Plans publicly available". NIH should make these plans publicly available to ensure transparency with the public who has funded the work, as well as to help enforce compliance.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The draft guidance does not specify whether fees that preserve data beyond the duration of the funded grant are allowed. The draft guidance does not specify whether personnel costs are allowable expenses related to data sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Page 1 states that, "Providing a rationale for decisions about which scientific data are to be preserved and made available for sharing, taking into consideration...consistency with community practices". This particular wording, allowing for researchers to remain consistent with community practices of sharing (or not sharing), is weak and does not move the needle on improving sharing practices across all scientific disciplines.

Page 2 states that "If an existing data repository(ies) will not be used, consider indicating why not...". The word "consider" should be removed. This policy should recommend the use of established repositories, and if there is a specific reason as to why this isn't feasible, then it should be justified.

Other Considerations Relevant to this DRAFT Policy Proposal:

This draft policy proposal suggests a desire to move in the right direction, but it ignores reality in ways that suggest that the most likely outcome is a new piece of paperwork during grant submission that produces no meaningful change in data sharing. The primary challenge that this effort aims to address is that retaining data to the maximum extent possible can advantage investigators who can then trade those data for authorship on manuscripts, positions on grant applications, or other scientific currencies of meaningful value. The draft solution, proposed here, is a mandatory document that states how data produced under the funding would be shared.

There are major weaknesses to the proposed solution:

* There does not appear to be a statement requiring data to be shared which is necessary to ensure that researchers will share.

- * There are no parameters around when data are to be shared. This provides flexibility, but does not help to ensure data are actually accessible in a reasonable time frame.
- * The Data Management Plan is only required as part of Just-in-Time, signaling that this is not a valued part of the application.
- * It is possible to technically share data while withholding key information that is necessary to make those data valuable for reuse. The key information can then be exchanged for authorship, position on proposals, or other scientific currencies.
- * It is likely to be inordinately time consuming for program officers, who appear to be the primary means of enforcement for extramurally funded projects, to verify each shared data output meets the commitments described in the sharing plan.

Because the NIH deals with many different fields, the only sustainable solution would appear to be one in which investigators are not just held to some minimum difficult-to-enforce bar but instead where they must compete to share data that become reused. Adjusting this policy to support the following would promote a culture where investigators are incentivized to produce datasets that are valuable, reusable, and available:

1. Include the uptake and impact of previously shared data (if any).
2. Include the sharing plans for the proposed work as a required part of the complete application package to be evaluated by reviewers (included in the current structure).
3. Be evaluated as a separate scoring criterion alongside resource sharing ("Resource and Data Sharing") on Extramural Awards with comparable consideration for spending through Contract, Intramural Research Projects, and other funding agreement mechanisms.
4. Publicly releasing the plans to ensure transparency and compliance.

Submission ID: 1289

Date: 12/22/2019

Name: Joshua Batson

Name of Organization: Chan Zuckerberg Biohub

Type of Data of Primary Interest: Genomic

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Infectious Diseases

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

There should be a presumption that all research data underlying a publication is shared at time of publication. The current language is weak and has statements such as "shared data should be made accessible" or "not all data generated in the course of research may be necessary to validate and replicate research findings." Instead the policy should say that shared data **MUST** be made accessible, except when justified by a small number of reasons, such as participant privacy concerns that cannot be overcome by protective measures, or studies on vulnerable populations.

The draft lists an expectation of "timely" data sharing. This should be defined as generally at the time of publication. Funding opportunities specifically designated to create a shared resource should specify a date by which data must be available even in the absence of a publication. This aspect is a step backwards from previous NIH policy which clearly defines "timely" as "no later than the acceptance for publication of the main findings from the final data set." The relaxation of this existing requirement is not justified.

Section II: Definitions:

Should include definitions of FAIR data and the 15 FAIR principles.

Section III: Scope:

Scope should make clear that the policy continues to apply for scientific data produced by funding in whole or in part from NIH after the NIH funding period is over.

Section IV: Effective Date(s):

The current absence of an effective data management and sharing policy and lack of enforcement causes a serious negative impact on health research and enables an ongoing waste of public funds. The noncommittal implementation date of the draft is unacceptable. The final policy should have a "no later than" date for implementation, ideally 12 months after issuance of the final policy.

Section V: Requirements:

To ensure good data management, any data described as collected in a progress report must be deposited independently and an accession code or digital object identifier (DOI) supplied. Except when specified by the funding opportunity announcement, researchers may embargo this data until publication. Grant opportunities specifically designated to create a shared resource should specify a date by which data must be available even in the absence of a publication.

It should be clear that these requirements apply not just to research project grants and contracts, but most other forms of requests for support that will lead to the creation of scientific data. This includes cooperative agreements, career grants, fellowships, scholarships, and training grants.

Absent a compelling reason otherwise, contract solicitations should specify that collected data is the property of NIH. They should also include specific requirements that data should be made publicly available in a third-party repository as a periodic deliverable, upon which further funding can be conditioned.

There are a large number of digital repositories with different policies. You should require that acceptable digital repositories must not allow recipients to unilaterally change or delete deposited data. The repositories, may, however, allow adding new versions of data advertised in metadata for the original dataset.

It is important to protect human participant privacy but it is also important that concerns about human participant privacy not be abused to eliminate appropriate data sharing. It is especially worth considering that many human participants expect that data from their participation will

be shared with other qualified researchers. Ineffective sharing of the resulting data (assuming appropriate protective measures such as de-identification are in place) is unethical as it wastes human participants' contributions to research and may result in more patients being exposed to harm. Therefore it should be an explicit goal of this policy and any submitted Data Management and Sharing Plans to maximize access subject to necessary restrictions.

Section VI: Data Management and Sharing Plans:

The draft states that NIH encourages scientific data to be made available. Instead, it should REQUIRE that scientific data are shared.

An effective Data Management and Sharing Plan should increase the overall impact of a grant and an ineffective one will decrease it. It is important that Data Management and Sharing Plans be provided to NIH peer reviewers and ICO advisory council review so they can consider the plan's effect on the application's overall impact, significance, and approach. Guidance to reviewers on how to score review criteria such as significance and approach should include review of the Data Management and Sharing Plan.

Therefore, NIH should require Data Management and Sharing Plans at the regular submission due date for an application, and not as a Just-in-Time submission. Overcoming deficiencies in the Data Management and Sharing Plan identified in summary statements could be provided as a Just-in-Time submission.

NIH should require that data management plans must describe how the researchers address each of the 15 FAIR Principles.

NIH should publish data management plans for funded grants and contracts alongside abstracts in public databases such as RePORTER. This will increase transparency and let other researchers and the public know what the grantees promised to NIH. This is the only thing that will make enforcement of individual plan items possible, given that NIH does not have the resources for exhaustive, systematic checks on compliance. Grantees knowing that their data management and sharing promises are readily available to the public will provide some measure of self-enforcement. Currently data sharing plans are available through Freedom of Information Act requests, and putting them on RePORTER will reduce the burden on data requesters.

The draft says that only data "deemed useful to the research community or the public" need be shared. It should be clear that applicants do not get to unilaterally decide what data is deemed useful. Any exceptions to the general principle that scientific data must be shared must be justified and funding conditioned on prior approval by an NIH advisory committee of data management experts that includes data scientists and librarians.

For intramural research, you should not give a single NIH official (such as Scientific Director or Clinical Director) the ability to assess Data Management and Sharing Plans without oversight. Data Management and Sharing Plans must be reviewed and approved by Boards of Scientific Counselors and ICO advisory councils during the existing periodic peer review and site visit process.

Section VII: Compliance and Enforcement:

It is currently unclear where to turn when NIH data sharing expectations and policies are not followed. To solve this, RePORTER should list, for each grant, contact information to request corrective action for violations of the Data Management and Sharing policy or published Data Management and Sharing plans. This should include contact email addresses for the principal investigators/project directors of the grant, contact email addresses for officials representing the grantee institution, and a contact email address at NIH. That will allow for solving issues at the most local level, when possible, and escalation when the previous proves ineffective. Similar information should be available for contracts and for intramural research projects.

In addition to reviewing progress reports and addressing complaints, NIH ICOs should also perform more thorough random audits to ensure grantees are performing data management as expected.

Current sanctions listed in the draft policy are incredibly weak and will have no deterrent effect. The policy should mention that failure to follow the Data Management and Sharing policy can be considered research misconduct by NIH. The policy should specify that violating the policy in place at the time of competing award at any time thereafter (including after the end of the award period) can result in sanctions. These sanctions can include publication of a notice describing the violation in the NIH Guide to Grants and Contracts, debarment and suspension from contracting, subcontracting, or financial assistance from the federal government, and prohibition of service to the Public Health Service on advisory committees, boards, or peer review committees, or as a consultant. Because it touches on potential research misconduct, this policy must be reviewed by the HHS Office of Research Integrity.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The guidance should specify that fees that preserve data beyond the funding period are allowed, as are personnel expenses related to data sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

An entry of "to be determined" in a Plan is not acceptable. This language will encourage useless Plans and should be removed.

Statements like "NIH does not expect researchers to share all scientific data generated in a study" defeat the purpose of this policy. Instead NIH should make clear that they do expect and require sharing of scientific data except in limited exceptions, justified by the applicant, and prior approval by peer reviewers, program staff, and an NIH advisory committee of data management experts that includes data scientists and librarians.

Section 1 describes "consistency with community practices" as a potential rationale for deciding which data are preserved and shared. In many scientific disciplines, community practices lag far behind general best practices and what the public expects for data management and sharing. This language allows certain communities to settle for mediocrity in data management and sharing, defeats the aim of this policy to improve data management and sharing. It should be removed. This also illustrates why decisions to withhold scientific data from sharing should not only be reviewed by study section members trained in the same discipline but also an NIH advisory committee of data management experts that includes data scientists and librarians.

Section 4 says that "if an existing data repository(ies) will not be used, consider indicating why not". This policy should require the use of established repositories, except when exceptions are justified and approved. It should not be up to applicants to unilaterally decide not to use standard established repositories and to not even justify the same.

Section 5 anticipates that applicants may have restrictions on sharing imposed by existing or future agreements. This provides a major loophole in the policy in that applicants may choose to enter into more restrictive agreements than necessary so that they can avoid data sharing. This can be overcome by (1) providing data sharing plans as part of initial peer-review so that peer reviewers can appropriately score any decrease in impact that may come about from restrictions on sharing, and (2) review by an NIH advisory committee that includes data scientists and librarians.

Other Considerations Relevant to this DRAFT Policy Proposal:

[These comments are to reinforce those of Michael Hoffman, who clearly articulated the issues and opportunities in this new proposal. For my work in infectious disease, it is essential that all data from a study may be available, because parts of the data gathered (eg, in microbiome studies) which are seemingly not of interest to the researcher may hold the key to understanding the disease. For example, a divergent virus in what was assumed to be a bacterial infection. Similarly, the submission of sequences for controls is essential for understanding contamination.]

I applaud your efforts to establish an excellent research data management and sharing policy. As written, I do not think this policy will provide a substantive change in data sharing. To maximize the benefit to the public of providing research funds, it is essential that the policy and enforcement be strengthened as described in this response.

In general, the draft policy is overly cautious and fails to consider the burden an ineffective policy will place on researchers who seek to use shared NIH-funded scientific data. The current system is incredibly burdensome on those seeking to obtain shared data because when data are not available as per existing NIH expectations, investigators can stonewall requests. There is no enforcement and the way to request enforcement is unclear. My most serious concern about this policy is that it is too vague on requirements in some places and lacks sufficient detail on enforcement.

A policy with ineffective, vague requirements and no real enforcement will have a serious negative impact on researchers who seek to use scientific data produced with public funds. There is a huge waste of researcher time and money attempting to obtain data that is lost, improperly described, or withheld. Failure to follow good data management practices leads to great inefficiency and slows the work of many researchers. There is also a large impact on our research communities, which lose opportunities to aggregate data and create a whole that is greater than the sum of its parts.

It is good to have both requirements and incentives to encourage high-quality data management. I suggest that an "Incentives for High-Quality Data Management and Sharing" section be added to the policy, including the following incentives:

Add to the NIH biosketch a section for key personnel to describe their most significant contributions to data management and resource sharing (including data, code, reagents, samples, and other materials). This should be separate from other contributions to avoid it getting short shrift due to lack of space. The past record of the principal investigator and other key personnel should be explicitly added to the scored review criteria.

NIH should create awards to recognize and cultivate excellence in data management and resource sharing, both at the individual researcher and institutional levels.

Submission ID: 1290

Date: 12/22/2019

Name: Charles Warden

Name of Organization: City of Hope National Medical Center

Type of Data of Primary Interest: Genomic

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

General Genomics

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

I believe "explicit consent" is currently required for genomics studies (for either public or controlled-access deposit), which is good.

However, I think "explicit consent" for public data deposit should be required for cell lines, since enforcement of controlled access will be difficult after the cell line is shared with others

Submission ID: 1294

Date: 12/24/2019

Name: Anna Greene

Name of Organization: Alex's Lemonade Stand Foundation

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: As a funder, we want all data types to be shared.

Type of Organization: Other

Type of Organization - Other: Nonprofit funder

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

Pediatric cancer research.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

There should be a requirement for data to be shared, and it should be stated up front. Furthermore, the use of repositories, should be expected in order to make data accessible, rather than generally inaccessible under a "data are available upon request" approach. The policy states that "Shared data should be made accessible in a timely manner...". Timely should generally be defined as by the time of publication or earlier if defined by the needs of specific FOAs, such as those creating a shared resource.

Section IV: Effective Date(s):

An effective date should be stated for the policy to go into full effect, ideally no later than 12 months after issuance of the final policy.

Section VI: Data Management and Sharing Plans:

This draft states: "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." NIH should require, not encourage,

data to be shared. It is unclear who would be responsible for deeming data as useful, as applicants are certainly not in a position to make an unbiased call.

In this policy, the Data Management Plan will be submitted as part of the Just-in-Time. This signals that the plan is not a valued part of the application and is in fact an afterthought. It should be required as part of the application so that appropriate sharing costs can be budgeted for at the time of application, and the plan can be included as part of the review process.

This section states that, "NIH may make Plans publicly available". NIH should make these plans publicly available to ensure transparency with the public who has funded the work, as well as to help enforce compliance. The plans could be provided through RePORTER.

Section VII: Compliance and Enforcement:

Contact information for those responsible for sharing and enforcing sharing of data should be provided in RePORTER, so that issues may be addressed when data sharing plans are not followed.

There should be sanctions for those who fail to uphold their sharing plan.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The draft guidance does not specify whether fees that preserve data beyond the duration of the funded grant are allowed. The draft guidance does not specify whether personnel costs are allowable expenses related to data sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Page 1 states that, "Providing a rationale for decisions about which scientific data are to be preserved and made available for sharing, taking into consideration...consistency with community practices". This particular wording, allowing for researchers to remain consistent with community practices of sharing (or not sharing), is weak and does not help to move the needle on improving sharing practices across all scientific disciplines which should be the goal of this policy.

Page 2 states that "If an existing data repository(ies) will not be used, consider indicating why not...". The word "consider" should be removed. This policy should recommend the use of established repositories, and if there is a specific reason as to why this isn't feasible, then it should be justified.

Other Considerations Relevant to this DRAFT Policy Proposal:

This draft policy proposal suggests a desire to move in the right direction, but it ignores reality in ways that suggest that the most likely outcome is a new piece of paperwork during grant submission that produces no meaningful change in data sharing. The primary challenge that this effort aims to address is that retaining data to the maximum extent possible can advantage investigators who can then trade those data for authorship on manuscripts, positions on grant applications, or other scientific currencies of meaningful value. The draft solution, proposed here, is a mandatory document that states how data produced under the funding would be shared. There are major weaknesses to the proposed solution:

- * There does not appear to be a statement requiring data to be shared which is necessary to ensure that researchers will share.
- * There are no parameters around when data are to be shared. This provides flexibility, but does not help to ensure data are actually accessible in a reasonable time frame.
- * The Data Management Plan is only required as part of Just-in-Time, signaling that this is not a valued part of the application.
- * It is possible to technically share data while withholding key information that is necessary to make those data valuable for reuse. The key information can then be exchanged for authorship, position on proposals, or other scientific currencies.
- * It is likely to be inordinately time consuming for program officers, who appear to be the primary means of enforcement for extramurally funded projects, to verify each shared data output meets the commitments described in the sharing plan.

Because the NIH deals with many different fields, the only sustainable solution would appear to be one in which investigators are not just held to some minimum difficult-to-enforce bar but instead where they must compete to share data that become reused. Adjusting this policy to support the following would promote a culture where investigators are incentivized to produce datasets that are valuable, reusable, and available:

1. Include the uptake and impact of previously shared data (if any).
2. Include the sharing plans for the proposed work as a required part of the complete application package to be evaluated by reviewers.
3. Be evaluated as a separate scoring criterion alongside resource sharing ("Resource and Data Sharing") on Extramural Awards with comparable consideration for spending through Contract, Intramural Research Projects, and other funding agreement mechanisms.
4. Publicly releasing the plans to ensure transparency and compliance.

Submission ID: 1295

Date: 12/24/2019

Name: Emma Grace

Name of Organization: The Chicago School of Professional Psychology - Washington, DC

Type of Data of Primary Interest: Clinical

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

clinical psychology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The purpose is well-defined and explained.

Section II: Definitions:

Data management starts with data collection, therefore, it should be included in the respective definition as follows, "Data Management : The process of collecting, validating, organizing, securing, maintaining, and processing scientific data, and of determining which scientific data to preserve."

Section III: Scope:

The scope is clear and well-defined.

Section IV: Effective Date(s):

This section is good.

Section V: Requirements:

it would be helpful if the NIH could create a standard template for the Data Management and Sharing Plan that applicants/researchers could fill in. It could be an online template or an offline fillable PDF form. Having such a standard template would save time and efforts for both the NIH and the applicants and the NIH would not need to request any "...additional or specific

information to be included within the Plan in order to meet expectations..." because it would have already been included in the standard template.

Section VI: Data Management and Sharing Plans:

The standard online template that I proposed in Section V above would also help with the following requirement in the draft NIH Policy: "Plans may be updated by researchers (with appropriate NIH ICO approval) during regular reporting intervals if changes are necessary or at the request of the NIH ICO to reflect changes in the previously documented approach to data management and data sharing throughout the research project, as appropriate." It would make the updating process much faster and easier for both the researchers and the NIH since both would have instant access to the updates.

Section VII: Compliance and Enforcement:

I would recommend changing "may be" to "must be" or "should be" in the following phrase: "...non-compliance with the NIH ICO-approved Plan may be..."

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

I think this is a technical error with spacing but it makes the following sentence look incomplete: "To assist individuals and entities who may be subject to a future NIH Policy for..."

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

For the "4. Data Preservation, Access, and Associated Timelines," I'm wondering why the NIH does not have its own repository? I think if the NIH could create its own repository, it would be much easier for the researchers who want to reuse the data find it and it would also allow the NIH see how efficient the data sharing is, e.g., how many researchers are looking for shared data and what types of data are on demand, and etc. It would also give more credibility to data and more confidence to researchers that they are using a trusted repository. As a researcher, I would feel more confident using data from the NIH repository than from any other sources.

Submission ID: 1296

Date: 12/25/2019

Name: Boris Barbour

Name of Organization: The PubPeer Foundation

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: The PubPeer web site is open to discussion of all science.

Type of Organization: Other

Type of Organization - Other: Non-profit Foundation promoting post-publication peer review

Role: Other

Role - Other: Organiser of PubPeer web site (also a researcher in neuroscience)

Domain of Research Most Important to You or Your Organization:

The PubPeer web site is open to discussion of all science.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Mandatory data sharing can have an important preventative action against low-quality and fraudulent research.

As an organiser of the PubPeer web site, I have accumulated a long experience of low-quality, erroneous and fraudulent research, as whistleblowers often identify such problems in publications via our site. My experience informs these comments addressing the benefits of data sharing for research quality and integrity, which I believe have been neglected in the draft policy. Thus, in addition to the benefits of data sharing correctly listed in the draft policy, I wish to stress that data sharing can also have a broad preventative action against low quality research and fraud. Research quality would be improved because authors will be more careful about their work if they know it will be exposed to public scrutiny. Analogously, authors who might falsify or fabricate research would, under data sharing, have the very difficult task of falsifying/fabricating an entire, coherent data set; the risk of detection would be greatly increased. Without data sharing, authors need only falsify/fabricate the occasional illustration (something we see highlighted very frequently on the PubPeer web site). Finally, if all the data underlying a publication is accessible, this greatly facilitates and accelerates in-depth investigation of any potential problems. Most of the apparently serious problems highlighted

on the PubPeer web site could be confirmed or resolved by access to the underlying data, but this is rarely available or provided.

Section III: Scope:

Public sharing of all data should be mandated, with very few exceptions.

Section IV: Effective Date(s):

All new funding should include the condition of mandatory public data sharing.

Section V: Requirements:

Public data sharing should be mandatory except under exceptional circumstances.

The principal benefits of data sharing for research quality and integrity will be lost if data sharing is not made mandatory or, at the very least, encouraged strongly by awarding credit during review for plans to share data publicly and then enforcing compliance with these plans. This is because the worst actors - those producing low-quality research or engaging in misconduct - are precisely those who will choose not to share their data. They will continue to compete unfairly with honest researchers trying to perform high-quality research. The perverse incentives rewarding researchers who cut corners and worse will be perpetuated.

Section VI: Data Management and Sharing Plans:

Public data sharing should be mandatory except under exceptional circumstances. At the very least, the data management plans should be scored during review with significant credit being awarded for plans to share data publicly. The draft proposal for submission of a data management plan as "just-in-time" and the absence of any mandatory sharing result in a policy that in reality provides absolutely no encouragement beyond pretty words to share data.

Section VII: Compliance and Enforcement:

If, as suggested, public data sharing is made mandatory, compliance and enforcement should obviously ensure that the mandated sharing is indeed implemented. It will be important to identify clearly institutional and funder-level contact persons (not the authors) to whom members of the public (including other researchers) may complain if the data sharing is inadequate in some way. These contact persons must have powers to ensure that data-sharing problems are rectified. Authors (and institutions) who do not implement the mandated data sharing adequately should be subject to effective penalties - for instance, reimbursement of funding and restriction of future funding opportunities.

Submission ID: 1297

Date: 12/31/2019

Name: Chris Brown

Name of Organization: Atlas Research

Type of Data of Primary Interest: Clinical

Type of Organization: Other

Type of Organization - Other: Government Contractor

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Bioinformatics across all disciplines.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Part of the data that should be managed as part of the plan is not only the raw data, but any training and test data sets generated through the use of machine learning techniques. A large amount of research is done using machine learning, and without access to the training and test data sets, bioinformatics analysts cannot reliably recreate results from the algorithms used in the research

Section II: Definitions:

Atlas Research (Atlas) recommends that NIH include machine learning training and test data in the definition of scientific data.

Section III: Scope:

No comments.

Section IV: Effective Date(s):

The effective date for the policy should occur after the comment period when the policy is approved. All new research activities funded by NIH—whether grants, intermural, or other activities—should comply with the policy and policy compliance should be written into all new contracts.

Section V: Requirements:

Given the investment NIH has made in a number of "data commons," Atlas recommends including a requirement stating that the data must be loaded to the appropriate data commons at the end of the contract. This would ensure that NIH always has a copy of the data to share after the contract ends. Additionally, it might be appropriate to state that the owner of the data can or cannot charge for access to the data. NIH may want to ensure the data is freely available to prevent an organization from monetizing the data.

NIH should ensure organizations that do charge for data usage meet with their government Grants Management Specialist to understand how they are allowed to use the funds. Program income for the grant:

https://grants.nih.gov/grants/policy/nihgps/html5/section_8/8.3.2_program_income.htm

Section VI: Data Management and Sharing Plans:

Atlas recommends removing "...as long as it is deemed useful to the research community or the public" from the section. The phrase presents a way for someone to not share the data because they deem it not useful for the research community or public. Consider specifying some of the areas to be addressed in this section to give an overview and then detailing the guidance in the "Elements of An NIH Data Management and Sharing Plan."

Section VII: Compliance and Enforcement:

NIH should require that data be loaded to an NIH data commons for future access after the funding period. Organizations will not want to continue to pay for allowing access to the data after funding has expired. Given the data was created with NIH funds, NIH should take control of the data and ensure it continues to be available.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Section 3 Local data management considerations. While storing and using the data is part of the overhead of an organization during the course of the project, costs associated with sharing data is not part of the overhead as organizations normally do not share their data externally; this would include costs for bandwidth, possible cloud costs, managing users for security purposes, and cybersecurity activities. NIH should consider paying for these costs or moving the data into an NIH-controlled environment.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

In section 1 (Data Type), the phrase "...estimated amount of scientific data that will result from NIH-funded or conducted research, which scientific data will be preserved and shared, and the rationale for these decisions" should be removed. All data that is created from NIH-funded or conducted research should be considered scientific data. This section should focus on the description, valid values, type of data (integer, string, float, etc.), among other metadata.

In section 2 (Related Tools, Software, and/or Code), NIH should recommend that data be stored in common formats, such as comma separated values or other formats that are widely usable without the need for special tools. This should be written into grants and contracts as it will allow the possibility of the widest use of the data across the scientific community.

In section 4 (Data Preservation, Access, and Associated Timelines), NIH should specify how long data should be made available. Leaving this timeframe open allows data providers to discontinue data after the funding or period of performance, which will result in the loss of data for researchers to utilize.

Other Considerations Relevant to this DRAFT Policy Proposal:

Overall, NIH must ensure that data management policies are incorporated into grants and contracts. Too often in the research community, people do not share their data. With the advent of machine learning techniques, it is more important than ever for research to gain access to data—both high and low quality—to train new algorithms to make new discoveries. The medical community is held back by the research community keeping data to themselves. This is not unique to NIH as the intelligence community has suffered from the same issues over the years—even after 9/11. NIH should ensure that data is available to the research community so researchers can make new discoveries.

Submission ID: 1298

Date: 1/2/2020

Name: Stephan Bour

Name of Organization: Digital Infuzion, Inc.

Type of Data of Primary Interest: Clinical

Type of Organization: Other

Type of Organization - Other: Technology company

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

Infectious diseases, patient engagement, precision medicine

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The policy and associated guidance are well thought-out and clearly explained. We have only a few comments that we believe could strengthen the policy without undue burden on compliance.

Section III: Scope:

While the physical specimen themselves may not be "scientific data", the metadata about them is. There is tremendous potential value in disseminating information about biospecimen and the context in which they were acquired, and the assays that were performed on them.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Section 4: "Where scientific data will be archived to ensure long-term preservation (i.e., which repository(ies))." This requirement may be too open-ended. Several government-funded public repositories are available to archive common data such as sequencing (GenBank, SRA), pathogens (BRCs), epitopes (IEDB), and so on. The NIH should consider mandating specific repositories for the most common data types. If an existing repository is not used, a strong justification should be required.

"How the scientific data will be findable." The ability of others to find and use the data is the ultimate benefit of the entire effort. We therefore believe that the guidelines to make the data Findable should be more detailed. For example, while each sequence submitted to GenBank would receive a "persistent unique identifier" in the form of an accession number, a Bioproject would be needed to group all the data related to the study the data sharing plan is developed for.

"When the scientific data will be made available to other users." This section could be strengthened by defining a few boundaries. For example, that genomic data needs to be shared no later than 90 days after sequencing/assembly. Similar boundaries should be created for the most common data types.

Submission ID: 1299

Date: 1/3/2020

Name: Greg Raschke

Name of Organization: NC State University Libraries

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

North Carolina State University's research enterprise is broad and interdisciplinary, encompassing, among other areas, a wide range of genomics, health, and life sciences disciplines such as bioinformatics, environmental health science, genetics and genomics, molecular biology, translational regenerative medicine, and all aspects of veterinary medicine. Scholars and researchers from diverse backgrounds collaborate with each other and with public and private sector partners to address a wide range of research topics.

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

Page 2 discusses allowable costs and refers the reader to the Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing, which states, "Budget estimates should not include infrastructure costs typically included in institutional overhead (e.g., Facilities and Administrative costs), nor costs associated with the routine conduct of research."

An area of frustration is the cost of storing data on institutionally managed systems. Most universities charge for storage over a certain limit, but the NIH (and other funders) do not allow the cost of that storage. These costs should be allowable for the proposed project.

Section VI: Data Management and Sharing Plans:

"The NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public."

This language is too vague because it is not clear how and by whom the data would be "deemed useful to the research community or the public." More context is needed here.

"Plans should also identify strategies or approaches to ensure data security and compliance with privacy protections are in place throughout the life of the scientific data."

It would be helpful to point to documentation, guidance, and/or examples about what is expected here. Researchers are not always clear on what is expected of them in terms of data security and compliance with privacy protections, and they may not understand what type of information they need to provide and/or steps they need to take in order to meet this requirement.

"NIH may make Plans publicly available."

As more of a comment than a suggestion, there is some concern about the implications for stakeholders who may be written into a plan but not consulted, such as entities at an institution that provide resources (e.g., Information Technology office) but that the researcher did not consult and misinterpreted when creating their plan. Who is the responsible party in this instance-- the researcher or the institution? Additionally, what about potential security concerns that may be shared in plans — will the NIH redact sensitive information, or would that responsibility fall to the author of the plan?

"NIH encourages the use of established repositories for preserving and sharing scientific data."

It is recommended that the NIH provide a list of repositories that are deemed "established" or, at the very least, a list of attributes of a repository that would make it "established." There is work being done and published in the library science and information field about repository selection and criteria that could be useful. Furthermore, the NIH should be explicit about what it expects in terms of preservation from a data repository.

"Extramural Awards: Plans will undergo a programmatic assessment by NIH staff within the proposed funding NIH ICO. NIH encourages potential awardees to work with NIH staff to address any potential concerns regarding the Plan prior to submission."

It would be helpful to describe what type of concerns an NIH staff member could address.

It would be helpful to describe what kind of training in data management the NIH staff members will undergo that would be relevant to assessment of data plans.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Data Type

"Providing a rationale for decisions about which scientific data are to be preserved and made available for sharing, taking into consideration scientific utility, validation of results, availability of suitable data repositories, privacy and confidentiality, cost, consistency with community practices, and data security."

It would be helpful for the NIH to prioritize this list.

What is meant by "suitable data repository"? There should be guidance about what constitutes "suitable." The NIH should provide information about how much data should be shared and preserved when a project produces or collects an amount of data larger than what any repository will allow.

Related Tools, Software, and/or Code

The NIH should require that a researcher make a case if they are not using an open source tool.

Overall, we are pleased that the NIH has included a section for this.

Standards

The content in this section is not very easy to understand for a lot of researchers. We recommend providing more information about what standards and common data elements are, where to find them, and why they are useful. While some examples are listed, we suggest a more exhaustive list. Overall, more guidance is needed.

"While many scientific fields have developed and adopted common data standards, others have not. In such cases, the Plan may indicate that no appropriate data standards exist for the data to be collected, preserved, and shared."

We recommend that rather than allowing a plan to indicate that no standards exist, the NIH require a researcher to describe some kind of local standard it will use (and how it will be described for reuse).

The language in this section makes it sound like there will only be one metadata standard needed for the whole project when, in reality, there are different metadata standards for

different types of data and different things that will be classified as scientific data. Information is needed to indicate that there is the chance a researcher will need to use more than one metadata standard depending on the data.

Data Preservation, Access, and Associated Timelines

The first bullet mentions long-term storage plans and where the data will be archived. The researcher should be asked to include language about what happens if the repository is decommissioned or loses funding.

The NIH should also require (rather than suggest) the inclusion of the name and URL of the chosen repository.

The second bullet point mentions backups. The NIH should make it clear whether the researcher is to explain about the backup procedures of a chosen repository or backups that the researchers / research group need to be doing.

The second bullet point also mentions persistent unique identifiers. The NIH should consider the mention of favoring a DOI for this.

Due to the growing size of data, as well as size limits from most current data repositories, the NIH should include language advising researchers how to determine how much data to share when their dataset exceeds the standard size accepted at most repositories (e. g., limits range from 5GB- 50GB per dataset for Zenodo, Open Science Framework, and Figshare).

Data Sharing Agreements, Licenses, and Other Use Limitations

All bullet points are about proprietary issues. Information should be added about the other areas mentioned in the introduction.

"Any other considerations that may result in limitations on the ability to broadly share scientific data."

This is very broad. What is in the range of valid other considerations? Ethical, cultural, license that has patent grant language, strategic reasons, etc.?

Oversight of Data Management

"An indication of the individual(s) who will be responsible for executing various components (e. g., data collection, data analysis, data submission) of the Plan over the course of the research project and the roles of the individual(s) in data management, and a description of the appropriate expertise for oversight."

This section would benefit from either use examples of what the NIH means for appropriate expertise or examples of the types of roles and the expected expertise for an individual in such a role.

Other Considerations Relevant to this DRAFT Policy Proposal:

There should be a statement required somewhere in the plan that speaks to who benefits from the data and perhaps who could reuse the data, specifically, and how would that benefit and help advance the science. We think it would be beneficial for the NIH to consider rewarding projects that can articulate who else would benefit from the data being generated and shared. It would be valuable for the application and rubric to be explicit about the value of open and reusable data and to ask researchers to tell a story about the value of their open data.

While it is important for individuals to be responsible, it is unclear what the responsibility of the institution is in the oversight of data management. The NIH should be clear about what responsibility it believes falls on the researcher and what falls on the institution.

Submission ID: 1300

Date: 1/3/2020

Name: Stephanie Fox-Rawlings

Name of Organization: National Center for Health Research

Type of Data of Primary Interest: Other

Type of Organization: Other

Type of Organization - Other: Think Tank

Role: Other

Domain of Research Most Important to You or Your

Organization: Public Health

Attachment:

NCHR Comments on NIH Policy for Data Sharing.pdf

Description:

full comments



**NATIONAL CENTER FOR
HEALTH RESEARCH**
The Voice For Prevention, Treatment And Policy

**National Center for Health Research Public Comments on
NIH's Request for Public Comments on a DRAFT NIH Policy for Data Management
and Sharing and Supplemental DRAFT Guidance**

Thank you for the opportunity to comment on the draft Policy for Data Management and Sharing and associated draft guidances.

The National Center for Health Research (NCHR) is a nonprofit think tank that conducts, analyzes, and scrutinizes research, policies, and programs on a range of issues related to health and safety. We do not accept funding from companies that make products that are the subject of our work.

We commend NIH for efforts to encourage data management and sharing. Our Center has supported data sharing, and particularly for data funded by federal agencies or submitted to federal agencies as part of application materials to the FDA and other federal agencies. Data sharing between scientists is an invaluable tool for confirming the accuracy of reported research findings and enabling other scientists to replicate results and understand any conflicting findings. Taxpayers deserve to have NIH maximize the usefulness of the funds they've invested in research, through data sharing and other means. We are confident that requiring responsible data management for NIH grant recipients will benefit the scientific community and the public.

Data sharing is an issue that has been debated and considered for several decades. We understand that it takes time to finalize and implement these policies in an efficient and easy-to-use manner, in the meantime, NIH should require researchers to share their data with other researchers upon request.

National Center for Health Research can be reached at info@center4research.org or at (202) 223-4000.

Submission ID: 1301

Date: 1/3/2020

Name: Scott Kahn

Name of Organization: Helmsley Charitable Trust

Type of Data of Primary Interest: Clinical

Type of Organization: Other

Type of Organization - Other: Funder

Role: Other

Role - Other: Data Science and Data Use Policy Consultant

Domain of Research Most Important to You or Your Organization:

Type 1 Diabetes and Crohn's Disease

DRAFT NIH Policy for Data Management and Sharing

Section II: Definitions:

- The term "data standards" should be defined as a set of community accepted definitions of data formats and data semantics that are publicly accessible
- Data Management and Sharing Plan should assert the use of data standards (e.g., CDISC for clinical trial data) that would be documented in the plan. (It is important to stress the use of standards to make the data interoperable and reusable) Lacking standards, most data are very challenging to use in practical terms.
- Metadata should assert the use of data standards for semantics so that descriptions are transferrable from data set to data set. Each research community would be well served to create "metadata standards" that lever common vocabularies!

Section V: Requirements:

- Submission of a Data Management and Sharing Plan should also mention that there should be limits on embargo restrictions on shared data. It is unreasonable that valuable data is withheld indefinitely (in some cases) for publicly funded research and the NIH should develop limits on data embargos. Waiting until all possible publications are submitted and accepted limits the value to research in general that might be achievable with shared data (in a timely manner).

Section VI: Data Management and Sharing Plans:

- "Plans should explain how scientific data generated by a research study will be managed and which of these scientific data will be shared" via existing data standards.
- "NIH encourages shared scientific data to be made available" via accepted standard data models ...
- ... "i.e., through de-identification or other protective measures" and through summarization of intended (acceptable) use(s).
- Extramural Awards – add a statement that "Scoring of new applications for funding will consider compliance with past Data Management and Sharing Plans and/or the investigator's demonstrated history of sharing data previously published and its reuse."

Section VII: Compliance and Enforcement:

- Extramural Awards – change "may affect future funding decisions" to "will affect future funding decisions". This is where the "stick" needs to be communicated to grantees/applicants. There is a second instance of changing "may" to "will" in the section on Post Funding or Support Period".

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Supplemental DRAFT Guidance: Elements of a Data Management and Sharing Plan section 3

- Local data management should be mandated to include periodic backups to prevent data loss. The NIH might consider pushing for use of Cloud resources for the security and longevity of project data.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)

Section 1 – Data Type

- "Identifying metadata" ... and the relationship with existing data standards.
- "For scientific data derived from human participants" ... "regulations, statutes, guidance, and institutional policies" and that anticipate advances in re-identification methods.

Section 2 – Related Tools, Software and/or Code

- ... "only from the research team or some other source)" when standards are not available.

Section 3 – Standards

- Remove the "if any" from the opening sentence. As stated, this makes standards usage optional. The added text in the previous section covers those cases where no standards exist.

Section 4 – Data Preservation

- Much more guidance is required for grantees of where data can be deposited/archived for subsequent reuse. Our experience is that there are no good tools or services for researchers to leverage in the search for appropriate repositories.
- A discussion of embargo periods should be made here. This should include the delaying of submission to a repository – a de facto embargo. Our experience is that a one-year delay from the collection of the data to its disposition is a reasonable compromise. We do not feel that association of the delay release of data to the submission or acceptance of a publication is justifiable.

Attachment:

Helmsley feedback on NIH Draft Data Sharing Policy - Jan 2020.docx

Description:

Summary of feedback on data sharing policy - easier to read!

Feedback on NIH Draft Policy for Data Management and Sharing and Supplemental Draft Guidance

Submitted on Behalf of the Helmsley Charitable Trust

II. Definitions

- The term “data standards” should be defined as a set of community accepted definitions of data formats and data semantics that are publicly accessible
- Data Management and Sharing Plan should assert the use of data standards (e.g., CDISC for clinical trial data) that would be documented in the plan. (It is important to stress the use of standards to make the data interoperable and reusable) Lacking standards, most data are very challenging to use in practical terms.
- Metadata should assert the use of data standards for semantics so that descriptions are transferrable from data set to data set. Each research community would be well served to create “metadata standards” that lever common vocabularies!

V. Requirements

- Submission of a Data Management and Sharing Plan should also mention that there should be limits on embargo restrictions on shared data. It is unreasonable that valuable data is withheld indefinitely (in some cases) for publicly funded research and the NIH should develop limits on data embargos. Waiting until all possible publications are submitted and accepted limits the value to research in general that might be achievable with shared data (in a timely manner).

VI. Data Management and Sharing Plans

- “Plans should explain how scientific data generated by a research study will be managed and which of these scientific data will be shared” via existing data standards.
- “NIH encourages shared scientific data to be made available” via accepted standard data models
...
- ...”i.e., through de-identification or other protective measures” and through summarization of intended (acceptable) use(s).
- Extramural Awards – add a statement that “Scoring of new applications for funding will consider compliance with past Data Management and Sharing Plans and/or the investigator’s demonstrated history of sharing data previously published and its reuse.”

VII. Compliance and Enforcement

- Extramural Awards – change “may affect future funding decisions” to “will affect future funding decisions”. This is where the “stick” needs to be communicated to grantees/applicants. There is a second instance of changing “may” to “will” in the section on Post Funding or Support Period”.

Supplemental DRAFT Guidance: Elements of a Data Management and Sharing Plan section 3

- Local data management should be mandated to include periodic backups to prevent data loss. The NIH might consider pushing for use of Cloud resources for the security and longevity of project data.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)

Section 1 – Data Type

- “Identifying metadata”... and the relationship with existing data standards.
- “For scientific data derived from human participants” ... “regulations, statutes, guidance, and institutional policies” and that anticipate advances in re-identification methods.

Section 2 – Related Tools, Software and/or Code

- ...” only from the research team or some other source)” when standards are not available.

Section 3 – Standards

- Remove the “if any” from the opening sentence. As stated, this makes standards usage optional. The added text in the previous section covers those cases where no standards exist.

Section 4 – Data Preservation

- Much more guidance is required for grantees of where data can be deposited/archived for subsequent reuse. Our experience is that there are no good tools or services for researchers to leverage in the search for appropriate repositories.
- A discussion of embargo periods should be made here. This should include the delaying of submission to a repository – a *de facto* embargo. Our experience is that a one-year delay from the collection of the data to its disposition is a reasonable compromise. We do not feel that association of the delay release of data to the submission or acceptance of a publication is justifiable.

Submission ID: 1302

Date: 1/3/2020

Name: Douglas P. Kiel, MD, MPH

Name of Organization: Marcus Institute for Aging Research, Hebrew SeniorLife

Type of Data of Primary Interest: Clinical

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Aging

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

Because each organization and their IRBs have separate standards for what can be shared and with whom, the informed consent form becomes a key document in facilitating the requirements of the policy. Our group recommended that the NIH should provide some direct guidance on this, and language that could be recommended for consent forms.

Section VI: Data Management and Sharing Plans:

Deceased participants: The NIH and our IRB have very different viewpoints about whether deceased individuals' data is protected from sharing without consent. The NIH does not consider deceased individuals to be human subjects, so all protections to them/their data are not considered. However, some IRBs do not agree with this, and still consider deceased subjects to be human subjects, and that their data should be handled as they agreed to in the ICF. This should be clarified in the revised policy.

The issue of de-identification of data is important. If data CAN be completely de-identified and de-linked with no way to re-identify subjects, then IRB/HIPAA waivers are possible. However, for some types of data this is not possible (e.g., those with dates—dates are considered HIPAA identifiers). The revised policy should be very clear about this. This point, and the preceding paragraph, highlights the frequent gap between NIH policy and IRB guidance. Historically the NIH has remained at arm's length from IRB related matters; however, this leaves investigators caught between required policy and IRB rules. We recommend that the revised policy address these gaps.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

In terms of the costs of data sharing, our Institute investigators underscored the importance of funding the costs of producing shared data. If these costs are to be included in the usual direct costs of a project, this will reduce funding to complete the research. We recommend that the support for data sharing not be diverted from other direct costs of research, especially since the cap on direct costs not requiring pre-approval has not changed for many years. We recommend that extra funds be allocated outside of the annual cap on grants.

Other Considerations Relevant to this DRAFT Policy Proposal:

Comments provide represent a compilation of suggestions from senior faculty in our Institution

Attachment:

Final comments on data sharing.docx

Description:

Complete document of above points

These comments come from investigators from the Hinda and Arthur Marcus Institute for Aging Research in Boston, MA. The Institute is an affiliate of Harvard Medical School.

In commenting on the DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance, our investigators wanted to highlight several key areas that require specific attention in revising this draft:

1. Because each organization and their IRBs have separate standards for what can be shared and with whom, the informed consent form becomes a key document in facilitating the requirements of the policy. Our group recommended that the NIH should provide some direct guidance on this, and language that could be recommended for consent forms.
2. Deceased participants: The NIH and our IRB have very different viewpoints about whether deceased individuals' data is protected from sharing without consent. The NIH does not consider deceased individuals to be human subjects, so all protections to them/their data are not considered. However, some IRBs do not agree with this, and still consider deceased subjects to be human subjects, and that their data should be handled as they agreed to in the ICF. This should be clarified in the revised policy.
3. The issue of de-identification of data is important. If data CAN be completely de-identified and de-linked with no way to re-identify subjects, then IRB/HIPAA waivers are possible. However, for some types of data this is not possible (e.g., those with dates—dates are considered HIPAA identifiers). The revised policy should be very clear about this. This point, and point number 2 above, highlights the frequent gap between NIH policy and IRB guidance. Historically the NIH has remained at arm's length from IRB related matters; however, this leaves investigators caught between required policy and IRB rules. We recommend that the revised policy address these gaps.
4. In terms of the costs of data sharing, our Institute investigators underscored the importance of funding the costs of producing shared data. If these costs are to be included in the usual direct costs of a project, this will reduce funding to complete the research. We recommend that the support for data sharing not be diverted from other direct costs of research, especially since the cap on direct costs not requiring pre-approval has not changed for many years. We recommend that extra funds be allocated outside of the annual cap on grants.

Submission ID: 1303

Date: 1/6/2020

Name: REBECCA LI

Name of Organization: Vivli

Type of Data of Primary Interest: Clinical

Type of Organization: Nonprofit Research Organization

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

clinical research

Attachment:

Vivli NIH comments .docx

Vivli Center for Global Clinical Research Data

Submitted electronically via <https://osp.od.nih.gov/draft-data-sharing-and-management>

Comment re: Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

As the world's largest funder of research, NIH has a major leadership opportunity to significantly impact data sharing by updating and aligning its data sharing policy with contemporary best practices. Sadly, this recently issued draft policy does not reflect the major step forward we had hoped for, but is simply an incremental change over the last policy.

Vivli is a non-profit organization founded in 2018 that manages the world's largest clinical trial data sharing platform. We provide a single point of search and request to participant-level data from over 4700 trials representing 2.2 million participants from 109 countries. Our comments are restricted to clinical trial data sharing, which we believe has the broadest and most immediate impact on advancing human health by accelerating new findings through data reuse. Moreover, clinical trial data sharing also respects trial participants' assumption of personal risk to contribute to science by maximizing the value of their contributions.

The plan as currently drafted is significantly weakened by choosing "deliberate flexibility" over a robust and clear mandate for clinical trial data sharing. Typically, flexibility in the conduct of science is a benefit; however, in this instance this approach significantly weakens our accountability to participants. We recommend at a minimum that the following be mandated elements within the data sharing plan with respect to clinical trials data rather than "flexible" elements managed at the discretion of the investigator:

- The current proposal leaves open the timeframe for when data would be made available to users at the discretion of researchers. We recommend that NIH funded clinical trials require reporting of individual participant-level data (IPD) to an approved repository within a reasonable time period. The National Academy of Medicine Report <http://www.nationalacademies.org/hmd/Reports/2015/Sharing-Clinical-Trial-Data.aspx> has suggested a practical timeframe of 18 months post-trial completion.
- The current proposal does not bind clinical trial proposals to declare a particular trial repository in the data sharing plan. We strongly recommend that NIH establish clear standards, criteria and best practices for clinical trial data sharing repositories, maintain a list of these approved repositories, promote awareness among researchers of this list, and require investigators to declare which approved repository they will be using.
- For clinical trial proposals, NIH should institute a requirement that demonstrates a rigorous search of prior relevant summary and IPD results in the research plan section. This would ensure that duplicative trials are not initiated, and we are respecting our participants contributions by leveraging them to the fullest.



In conclusion, Vivli strongly supports mandatory data sharing for clinical trial proposals. Investigators who do not meet minimum thresholds should have direct measurable rewards and consequences based on their data sharing performance. Perhaps the single most impactful change to the current draft policy would be to score the data sharing plan during the grant review process and ensure that this score impacts the funding decision. We have waited for 15 years for this important update to the NIH's data sharing policy. As this new policy lacks any effective mandate for sharing of clinical trial data, it in effect relinquishes NIH's responsibility to the research community, researchers and patients. This incremental proposal if enacted would signal to researchers that clinical trial data sharing is a voluntary endeavor, which breaks trust with trial participants' strong desire to share. We can do better.

NIH is in the enviable position of being able to alter incentives and investigator behavior in ways to produce lasting changes for future generations of patients if this proposal is crafted carefully. We appreciate the opportunity to share our perspective, and we urge the NIH to consider our suggestions carefully in the next draft of its data sharing policy and to lead in clinical trial data sharing.

Submission ID: 1304

Date: 1/6/2020

Name: Steve Pieper

Name of Organization: isomics, Inc.

Type of Data of Primary Interest: Imaging

Type of Organization: Biotech/Pharmaceutical Company

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization: Cancer imaging and image analysis software

DRAFT NIH Policy for Data Management and Sharing

Section VI: Data Management and Sharing Plans:

I believe that the Data Management plan should be reviewed by the study section as a scored criteria. The importance of the data should be addressed in the Significance section and the methods of sharing should be described in the Approach. Making clear that funding depends on data sharing will be the most effective way to convince investigators to take the process seriously when planning their research.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

It would be great if NIH could provide ways to make data citations valuable to researchers in a way similar to the value of publications. The data sharing plan guidelines could provide guidance about how to request that users of the data should cite the original data collection effort in the most meaningful and valuable way.

Other Considerations Relevant to this DRAFT Policy Proposal:

Sharing data is one of the most valuable things scientists can do, so as I stated above it should be a major review criterion and not something tacked on the end.

Anything that can be done to reward investigators for sharing valuable data should be recognized and encouraged. This could be in the form of special citation options, special funding programs for data sharing, and more educational materials demonstrating best practices.

Submission ID: 1305

Date: 1/7/2020

Name: Sean McGurn

Name of Organization: Triple Point Security, NIH Extramural Data Security Team (EDST)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Sensitive and Controlled Data

Type of Organization: Government Agency

Role: Other

Role - Other: Federal Contractor - NIH Office of the Director (OD) / Office of the Chief Information Officer (OCIO)

Domain of Research Most Important to You or Your Organization:

Overall Cybersecurity and Risk Management

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Recommendation: Expand explanation of policy objectives to include security and privacy (as underlying principles) to secure data from unauthorized use and disclosure. Also include language promoting confidentiality, integrity, and availability (CIA) as foundational principles for data management and sharing to be based on.

Recommendation: With regards to "describing how scientific data will be managed" the plan should be expanded to describe how the scientific data will be protected from the cybersecurity perspective. This recommendation complements comment 1.

Recommendation: The current expectations of the Data Management and Sharing Plans describe a document that should be tailored at the discretion of the ICO. The EDST recommends adding a supplemental document or linking to a scientific data security and privacy policy principles and framework (to be developed) that the plan would be based on, along with detailed baseline requirements to establish performance measures in development and review.

Question: Is the purpose of the policy to put the sole responsibility of implementation on the ICOs and the organizations handling scientific data? Relating to this notion, what is the role of the NIH OCIO in managing compliance?

Section II: Definitions:

Recommendation: The definition of data sharing should be enhanced to include data sensitivity as a key role in determining the methods in how data is transacted among authorized partners. Additional data types, such as Controlled Unclassified Information (CUI) (NIST SP 800-171) should be listed and defined to correlate data type with sensitivity and data handling requirements.

Recommendation: The definition of metadata and scientific data should be expanded to further define how factors such as data sensitivity and Federally funded intellectual property play a role in data identification.

Question: Is it permissible that metadata and scientific data be disclosed to the public? Are metadata and scientific data controlled, and only permitted use by authorized personnel?

Section III: Scope:

Question: To clarify the scope, does this policy apply to data in the Precision Medicine Initiative (PMI)? Also, does this policy apply to and/or supersede the Genomics Data Sharing Policy?

Section IV: Effective Date(s):

Question: In the event that an extramural grant is modified after the effective date, will the policy requirement for the "plan" be enforced? Examples regarding the question: if the modification includes additional funds, a conversion from a grant to a cooperative agreement, or an expansion of scope.

Section V: Requirements:

Question: Who will be responsible for approving and signing off on the risks associated with the Data Management and Sharing Plan?

Question: With regard to the statement noting that the funding ICO may require additional information (supplemental requirements) in the plan, what are the baseline requirements that are being imposed by the policy?

Section VI: Data Management and Sharing Plans:

Question: What constitutes a significant change to data management and sharing processes (e.g. Change Management Process)? Also, what is an example of the timeframe for reporting intervals? Are these processes documented?

Question: At what level of detail are strategies and approaches to data security and privacy required in the plan, and how will security and privacy requirements be documented for implementation?

For example, will the granularity of the plan dictate the need to show the implementation of NIST security and privacy controls (NIST SP 800-53), and best practices aligning with the NIST Risk Management Framework (NIST SP 800-37)?

Question: Does this statement indicate that an organization's plan will be available for review within an NIH controlled environment, or published on the public Internet? Will the plans be stored in a centralized repository, and if so, what is the centralized repository?

Recommendation: De-identified data should be considered sensitive and controlled data (in cases where the re-identification of data in question is also sensitive and controlled) due to the emerging risks of re-identifying data through data set correlation and cloud computing machine learning capabilities being widespread.

Recommendation: Training should be developed to establish a standard operating procedure(SOP) for the "programmatic assessment" process to enable NIH to conduct consistent plan reviews and assessments across the agency.

Question: What Security Compliance mechanisms will NIH ICOs or funding agencies have to adhere to when Data is Shared with out of country researchers and academicians and vice versa? What aspects need to be covered by the DSP?

Section VII: Compliance and Enforcement:

Question: What are the baseline requirements for an ICO to determine compliance with the Plan and policy? Also, will the ICOs develop their own baseline requirements?

Question: How will the policy compliance requirement of a plan be enforced with Other Transactions (OT), which currently are bound to an alternate process outside of the NIH GPS? The referenced "applicable NIH policies" should be listed as best case/scenario examples.

Question: Who will be responsible for making the determination that a "Plan" is non-compliant and what are the benchmarks and performance measures?

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

No Comments

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Recommendation: Develop a new section titled, "Security and Privacy Guidance for Plan," which should include NIH OCIO best practices for security, privacy, and risk management implementation. NIH OCIO would be responsible for providing this guidance and documentation content for OSP to include in the supplemental guidance for Plan Elements, in addition to referencing this content in the policy.

Recommendation: Remove the page count restriction as two pages may not be sufficient to discuss the data management/sharing/security/privacy aspects of an organization.

Recommendation: The plan should state the current and future state of data management and sharing, as the language suggests that the document will only include "proposed" content.

Recommendation: NIH OCIO/OSP develops and provides a template for the NIH Data Management and Sharing Plan during the Funding Opportunity Announcement (FOA) stage.

Question: How will the data types be identified? Will data type risk level be associated with FIPS 199 Security Categorization, and NIST SP 800-60 Vol 1/2? Alternatively, will data that is considered Controlled Unclassified Information (CUI) be identified as noted in the NARA CUI registry?

Question: How will the types of data be associated with sensitivity - and then associated with a corresponding baseline of requirements?

Question: What de-identification standards are being recommended and/or required in the event that they are utilized?

Question: How will the authors of the plan know what tools are being specified , and how access should be limited (e.g. best practices)? In addition, will additional guidance be provided to the authors of the plan to help identify what types of data and tools could generate risk to the research organization and/or NIH?

For example, what requirements in Access Control, Audit and Accountability, Configuration Management, among other security control families, are being required in the Plan?

Question: What are the minimum security and privacy requirements (pre-established standard) for the Plan?

Question: Are there any aspects of data retention of participant data (ex. Personally Identifiable Information (PII) originating from a System of Record) that need to be addressed separately in the Plan? For example, if a research organization is ingesting various different types of data sourced from different data sources (spanning Federal and Non-Federal systems), how are unique terms and conditions regarding data use and retention analyzed, collated, and integrated into a process?

In addition, how are these processes publicized for privacy concerns and transparency in the event that the data was collected from the public by the Federal Government?

Question: What are the requirements for system backups, contingency planning, and encryption standards?

Question: With regard to "a description of the appropriate expertise for oversight" is there supplemental guidance available that defines what the appropriate level of expertise is required for an oversight role?

Other Considerations Relevant to this DRAFT Policy Proposal:

Please see the attached spreadsheet, which includes comments and inquiries relating to the Draft NIH Policy for Data Management and Sharing. If your team has any questions relating to our submission, please contact the NIH Extramural Data Security Team via E-mail at Sean.McGurn@nih.gov and Annie.Chitre@nih.gov.

Attachment:

Comment_Matrix_NIH_PDMS_01072020.xlsx

Description:

Comment Matrix for the review of the draft NIH Policy for Data Management and Sharing

Comment #	Document Name	Section	Page #	Content in Question	Comment(s)	Reviewer
1	Draft NIH Policy for Data Management and Sharing	I. Purpose	1	"NIH encourages data management and data sharing practices consistent with the NIH Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research and the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles." Proposed Text: NIH also encourages the implementation of security, privacy, and risk management safeguards as appropriate and necessary for the preservation of scientific data in operations relating to research and downstream data transactions with other research entities.	Recommendation: Expand explanation of policy objectives to include security and privacy (as underlying principles) to secure data from unauthorized use and disclosure. Also include language promoting confidentiality, integrity, and availability (CIA) as foundational principles for data management and sharing to be based on.	EDST - TriplePoint Security
2	Draft NIH Policy for Data Management and Sharing	I. Purpose	1	"Under this Policy, individuals and entities would be required to provide a Data Management and Sharing Plan (Plan) describing how scientific data will be managed, including when, where, [Proposed Text: (and how) the scientific data will be (safeguarded) and shared, prior to initiating the research study.]"	Recommendation: With regards to "describing how scientific data will be managed" the plan should be expanded to describe how the scientific data will be protected from the cybersecurity perspective. This recommendation complements comment 1.	EDST - TriplePoint Security
3	Draft NIH Policy for Data Management and Sharing	I. Purpose	1	"This Policy is intended to establish expectations for Data Management and Sharing Plans upon which other NIH Institutes, Centers and Offices (ICO) may supplement as appropriate." Proposed Text: To provide additional background on the recommended security, privacy, and risk management components, which may be leveraged to bolster the plan, see the NIH Scientific Data Security Policy Principles and Framework."	Recommendation: The current expectations of the Data Management and Sharing Plans describe a document that should be tailored at the discretion of the ICO. The EDST recommends adding a supplemental document or linking to a scientific data security and privacy policy principles and framework (to be developed) that the plan would be based on, along with detailed baseline requirements to establish performance measures in development and review.	EDST - TriplePoint Security
4	Draft NIH Policy for Data Management and Sharing	I. Purpose	1	"This Policy is intended to establish expectations for Data Management and Sharing Plans upon which other NIH Institutes, Centers and Offices (ICO) may supplement as appropriate."	Question: Is the purpose of the policy to put the sole responsibility of implementation on the ICOs and the organizations handling scientific data? Relating to this notion, what is the role of the NIH OClO in managing compliance?	EDST - TriplePoint Security
5	Draft NIH Policy for Data Management and Sharing	II. Definitions	1	"Data Sharing: The act of making scientific data available for use by others (e.g., researchers, institutions, the broader public)."	Recommendation: The definition of data sharing should be enhanced to include data sensitivity as a key role in determining the methods in how data is transacted among authorized partners. Additional data types, such as <i>Controlled Unclassified Information</i> (CUI) (NIST SP 800-171) should be listed and defined to correlate data type with sensitivity and data handling requirements.	EDST - TriplePoint Security
6	Draft NIH Policy for Data Management and Sharing	II. Definitions	1	" Metadata: Data describing scientific data that provide additional information to make such scientific data more understandable (e.g., date, independent sample and variable description, outcome measures, and any intermediate, descriptive, or phenotypic observational variables)."	Recommendation: The definition of metadata and scientific data should be expanded to further define how factors such as data sensitivity and Federally funded intellectual property play a role in data identification. Question: Is it permissible that metadata and scientific data be disclosed to the public? Are metadata and scientific data controlled, and only permitted use by authorized personnel?	EDST - TriplePoint Security
7	Draft NIH Policy for Data Management and Sharing	III. Scope	2	"This Policy applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data."	Question: To clarify the scope, does this policy apply to data in the Precision Medicine Initiative (PMI)? Also, does this policy apply to and/or supersede the Genomics Data Sharing Policy?	EDST - TriplePoint Security
8	Draft NIH Policy for Data Management and Sharing	IV. Effective Date(s)	2	First Bullet	Question: In the event that an extramural grant is modified after the effective date, will the policy requirement for the "plan" be enforced? Examples regarding the question: if the modification includes additional funds, a conversion from a grant to a cooperative agreement, or an expansion of scope.	EDST - TriplePoint Security

9	Draft NIH Policy for Data Management and Sharing	V. Requirements	2	" Submission of a Data Management and Sharing Plan (Plan) outlining how scientific data will be managed and shared, taking into account any potential restrictions or limitations."	Question: Who will be responsible for approving and signing off on the risks associated with the Data Management and Sharing Plan?	EDST - TriplePoint Security
10	Draft NIH Policy for Data Management and Sharing	V. Requirements	2	"The funding NIH ICO may request additional or specific information to be included within the Plan in order to meet expectations for data management and data sharing in support of programmatic priorities or to expand the utility of the scientific data generated from the research."	Question: With regard to the statement noting that the funding ICO may require additional information (supplemental requirements) in the plan, what are the baseline requirements that are being imposed by the policy?	EDST - TriplePoint Security
11	Draft NIH Policy for Data Management and Sharing	VI. Data Management and Sharing Plans	3	"Plans may be updated by researchers (with appropriate NIH ICO approval) during regular reporting intervals if changes are necessary or at the request of the NIH ICO to reflect changes in the previously documented approach to data management and data sharing throughout the research project, as appropriate."	Question: What constitutes a significant change to data management and sharing processes (e.g. Change Management Process)? Also, what is an example of the timeframe for reporting intervals? Are these processes documented?	EDST - TriplePoint Security
12	Draft NIH Policy for Data Management and Sharing	VI. Data Management and Sharing Plans	3	"Plans should also identify strategies or approaches to ensure data security and compliance with privacy protections are in place throughout the life of the scientific data."	Question: At what level of detail are strategies and approaches to data security and privacy required in the plan, and how will security and privacy requirements be documented for implementation? For example, will the granularity of the plan dictate the need to show the implementation of NIST security and privacy controls (NIST SP 800-53), and best practices aligning with the NIST Risk Management Framework (NIST SP 800-37)?	EDST - TriplePoint Security
13	Draft NIH Policy for Data Management and Sharing	VI. Data Management and Sharing Plans	3	"NIH may make Plans publicly available."	Question: Does this statement indicate that an organization's plan will be available for review within an NIH controlled environment, or published on the public Internet? Will the plans be stored in a centralized repository, and if so, what is the centralized repository?	EDST - TriplePoint Security
14	Draft NIH Policy for Data Management and Sharing	VI. Data Management and Sharing Plans	3	"...human participants' privacy, rights, and confidentiality will be protected, i.e., through [data] de-identification..."	Recommendation: De-identified data should be considered sensitive and controlled data (in cases where the re-identification of data in question is also sensitive and controlled) due to the emerging risks of re-identifying data through data set correlation and cloud computing machine learning capabilities being widespread.	EDST - TriplePoint Security
15	Draft NIH Policy for Data Management and Sharing	VI. Data Management and Sharing Plans	3	"Extramural Awards: Plans will undergo a programmatic assessment by NIH staff within the proposed funding NIH ICO."	Recommendation: Training should be developed to establish a standard operating procedure(SOP) for the "programmatic assessment" process to enable NIH to conduct consistent plan reviews and assessments across the agency.	EDST - TriplePoint Security
16	Draft NIH Policy for Data Management and Sharing	VII. Compliance and Enforcement	4	"During the funding period, compliance with the Plan will be determined by the funding NIH ICO."	Question: What are the baseline requirements for an ICO to determine compliance with the Plan and policy? Also, will the ICOs develop their own baseline requirements?	EDST - TriplePoint Security
17	Draft NIH Policy for Data Management and Sharing	VII. Compliance and Enforcement	4	"Other funding agreements: Compliance with and enforcement of the Plan will be consistent with applicable NIH policies"	Question: How will the policy compliance requirement of a plan be enforced with Other Transactions (OT), which currently are bound to an alternate process outside of the NIH GPS? The referenced "applicable NIH policies" should be listed as best case/scenario examples.	EDST - TriplePoint Security
18	Draft NIH Policy for Data Management and Sharing	VII. Compliance and Enforcement	4	"...non-compliance with the NIH ICO-approved Plan may be taken into account by the funding NIH ICO for future funding decisions for the recipient institution..."	Question: Who will be responsible for making the determination that a "Plan" is non-compliant and what are the benchmarks and performance measures?	EDST - TriplePoint Security
19	Draft NIH Policy for Data Management and Sharing	General	N/A	N/A	Question: What Security Compliance mechanisms will NIH ICOs or funding agencies have to adhere to when Data is Shared with out of country researchers and academicians and vice versa? What aspects need to be covered by the DSP?	EDST - TriplePoint Security
20	Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)	General	N/A	N/A	Recommendation: Develop a new section titled, "Security and Privacy Guidance for Plan," which should include NIH OCIO best practices for security, privacy, and risk management implementation. NIH OCIO would be responsible for providing this guidance and documentation content for OSP to include in the supplemental guidance for Plan Elements, in addition to referencing this content in the policy.	EDST - TriplePoint Security

21	Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)	Introduction Paragraph	1	"A Plan should describe in two pages or less the proposed approach to data management and sharing that the specific research will employ."	<p>Recommendation: Remove the page count restriction as two pages may not be sufficient to discuss the data management/sharing/security/privacy aspects of an organization.</p> <p>Recommendation: The plan should state the current and future state of data management and sharing, as the language suggests that the document will only include "proposed" content.</p> <p>Recommendation: NIH OCIO/OSP develops and provides a template for the NIH Data Management and Sharing Plan during the Funding Opportunity Announcement (FOA) stage.</p>	EDST - TriplePoint Security
22	Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)	1. Data Type	1	"A description of the types and estimated amount of scientific data that will result from NIH-funded or conducted research, which scientific data will be preserved and shared, and the rationale for these decisions."	<p>Question: How will the data types be identified? Will data type risk level be associated with FIPS 199 Security Categorization, and NIST SP 800-60 Vol 1/2? Alternatively, will data that is considered Controlled Unclassified Information (CUI) be identified as noted in the NARA CUI registry?</p> <p>Question: How will the types of data be associated with sensitivity - and then associated with a corresponding baseline of requirements?</p>	EDST - TriplePoint Security
23	Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)	1. Data Type	1	"For scientific data derived from human participants or specimens, outlining plans for providing appropriate protections of privacy and confidentiality (i.e., through de-identification or other protective measures)"	<p>Question: What de-identification standards are being recommended and/or required in the event that they are utilized?</p>	EDST - TriplePoint Security
24	Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)	2. Related Tools, Software, and/or Code	2	"Consider specifying how needed tools can be accessed, (i.e., open source and freely available, generally available for a fee in the marketplace, or available only from the research team or some other source)."	<p>Question: How will the authors of the plan know what tools are being specified, and how access should be limited (e.g. best practices)? In addition, will additional guidance be provided to the authors of the plan to help identify what types of data and tools could generate risk to the research organization and/or NIH?</p> <p>For example, what requirements in Access Control, Audit and Accountability, Configuration Management, among other security control families, are being required in the Plan?</p>	EDST - TriplePoint Security
25	Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)	3. Standards	2	An indication of what standards, if any, will be applied to the scientific data and associated metadata to be collected, including data formats, data identifiers, definitions, unique identifiers, and other data documentation.	<p>Question: What are the minimum security and privacy requirements (pre-established standard) for the Plan?</p>	EDST - TriplePoint Security

Submission ID: 1306

Date: 1/7/2020

Name: Chris Bourg

Name of Organization: Massachusetts Institute of Technology - MIT Libraries

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Our institution produces and uses a broad array of data

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

Our institution produces and uses a broad array of data in a number of research domains

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

No comments.

Section II: Definitions:

Metadata: The phrase "more understandable" is too subjective. To better align with the FAIR guiding principles and reinforce that the metadata provided should allow others to replicate and/or reproduce the study, consider clarifying this phrase in terms of "useable" and "shareable."

Data Sharing: Accessibility of data seems to be overall missing in the definitions and could be a useful enhancement in the context of data sharing, bolstering the value of and need for reproducibility.

Scientific Data: We would suggest expanding this definition to include examples of expected data types, similar to that presented in the DOE Policy for Digital Research Data Management Glossary (<https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management-glossary>), "The term digital data encompasses a wide variety of information

stored in digital form including: experimental, observational, and simulation data; codes, software and algorithms; text; numeric information; images; video; audio; and associated metadata. It also encompasses information in a variety of different forms including raw, processed, and analyzed data, published and archived data." This expansion would prompt an expanded view of the types of data necessary for validation and replication and disambiguate the data included in "recorded factual material" as some may overlook analysis environments, workflows, and scripts in this phrase.

In addition to the included definitions, we strongly recommend defining the term "preservation" as its meaning can be ranging and difficult to pin down for researchers. Within the framing of data management the CASRAI glossary provides definitions for both digital preservation (<https://casrai.org/term/digital-preservation/>) and preservation (<https://casrai-test.evision.ca/glossary-term/preservation/>).

Section III: Scope:

No comments.

Section IV: Effective Date(s):

No comments.

Section V: Requirements:

Compliance as written appears restricted to the submission of the plan to the NIH ICO for approval. It's unclear whether ICO's have the facility and authority to develop baseline and publicly available standards for the data produced under their purview. This, as opposed to ad hoc and supplemental standards of assessment, would aid both the researchers developing plans and the ICO's assessing them.

Section VI: Data Management and Sharing Plans:

Regarding specific phrasing of this section:

"NIH encourages shared scientific data to be made available as long as it is deemed useful...": "As long as" could refer to both conditional or temporal situations and "useful" can be subjectively scoped to be a subset that does not lend to replication or reproducibility. It's also uncertain who deems the data to be useful and by what standards. We suggest clarifying the language of this statement to provide more definitive conditions for when data should not be shared.

"NIH may make Plans publicly available.": Under what caveats would a successful application's plan not be made available? If there are privacy or other concerns that would restrict access,

these should be articulated to provide direction to researchers that sets the appropriate and consistent expectation for the public sharing of plans.

The Plan Assessment subsection should include a statement about evaluating plans as an "additional review consideration" (as stated in a previous drafting) or details on what a "programmatic assessment" would entail. We further recommend including data management plans in the overall grant proposal impact score. This provides NIH an opportunity to make a strong data management statement that reinforces the importance of both the Plan and establishing groundwork for data sharing from the beginning rather than as an afterthought at a project's completion. This would also embed compliance at the onset. This does, however, place a burden of responsibility on reviewers to evaluate content that may exceed their familiarity with best data management practices and infrastructure. Including the data management plan in the impact score as we suggest may require substantial training and guidance for reviewers, which ultimately may align with the workforce development goal of the NIH Strategic Plan for Data Science. Strong consideration of the development of this reviewer support is needed.

Section VII: Compliance and Enforcement:

During the Funding or Support Period: We support compliance enforcement that carries forward onto future funding decisions as specified for extramural awards. It is unclear, however, how these carry-forward consequences will be managed and/or shared with future peer reviewers. Attention should be given to this workflow and the plan for implementation included in this section.

Post Funding or Support Period: "...non-compliance with the NIH ICO-approved Plan may be taken into account by the funding NIH ICO for future funding decisions for the recipient institution..." The use of "may" here and in the previous section is too ambiguous. We strongly recommend replacing with "will" and providing information for how this will be carried out.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Item 1: We recommend expanding this item to include work necessary for supporting interoperability, particularly machine-interoperability as part of FAIR data, as this may support improving current community standards. e.g., "...community standards for sharing and interoperability," or "...standards and supporting FAIR principles."

Item 2. We encourage the NIH to clarify the sentence, "When proposing to use a repository that charges recurring fees, budgets may include costs that would be incurred for preserving and sharing data." A clearer statement is required if the intent is to allow the budget to support recurring costs for long-term preservation beyond the funded period of the project.

Item 2: We suggest strengthening the statement "If the Plan proposes use of multiple repositories, consider including costs associated with use of each proposed repository," by rewording it to "If the Plan proposes use of multiple repositories, costs associated with the use of each proposed repository may be included".

Item 3: In the final sentence, the sole example of data access fees may obscure its intent. We recommend expanding this example as follows: (e.g., data access, licensing, or subscription fees).

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We support the existence of guidance outlining the expected elements of a Data Management and Sharing Plan (Plan). We encourage the NIH to consider that two pages may be insufficient to describe an effective Plan dealing with multiple types or sources of data with diverse security or licensing requirements and multiple analysis methods and tools. As an alternative, the NIH may consider requiring a brief, two-page limited Plan be submitted for all applications, with a more in-depth Plan submitted as part of Just-in-Time or similarly timed submissions.

The wording of the final phrase of the introductory paragraph is unclear. We recommend rephrasing to clarify the intent and audience, e.g., "Elements of Plan should include:".

Element 1:

For scientific data derived from human participants or specimens, consider including community representation and community participation in addition to privacy protections and confidentiality measures.

Element 2:

In addition to noting potential analysis tools, there should be some consideration for indicating how analysis scripts, codes, workflows, and details of the computing environment may be shared. These are often necessary for understanding and reproducing results. If these details are not prospectively known, providing a plan of how and when the Plan will be updated with these details and how documentation of tools, software, etc. will occur is desirable.

Element 3:

The current phrasing of "no appropriate data standards exist for the data to be collected, preserved, and shared" implies the need for a single standard that is applicable to collection, preservation, and sharing. Consider rewording to "for the data to be collected, preserved, or shared" to allow for a variety of standards that may differentially serve a selection rather than the totality of these stages.

Element 4:

We suggest that the NIH clarify wording around storage, preservation, and archiving activities to avoid confusion regarding expectations. Consider re-wording the first bullet point to replace "archived" with "stored," as most digital repositories do not engage in activities understood as digital archiving.

The content of this element addresses issues of storage and access rather than the complexities normally associated with preservation processes or specific timelines. In the absence of a definition of "preservation" as recommended in our comments for "Section II: Definitions" of the policy, the NIH should consider retitling this element to accurately reflect its content, e.g., "Data Storage, Access, and Associated Timelines: An indication of the activities and timelines for data storage and access, considering:" Additionally, we would encourage the concept of data citation be used to frame the motivation and emphasize the importance of these activities, e.g., "...to ensure long-term preservation and enable citation...", "How the scientific data will be findable, whether a persistent unique identifier will be used to enable citation, what standard indexing tools will be used, and any provisions made for maintaining..."

Bullet 5: The use of the term "preserving" is misleading as preservation activities should be happening throughout the research process, from data collection and analysis through sharing and storage, and beyond. It is unclear if "preserving" is referring to data deposit, sharing, or another activity.

Element 5:

This section focuses on documenting usage limitations. Of equal or greater use is the documentation of usage permissions, as the lack of this documentation often leaves the usability of data in question. We recommend an additional bullet indicating the description of permissions in addition to restrictions. Such descriptions of permissions or allowable uses might be most easily affected by the application of the most open license that is appropriate.

Element 6:

In addition to the active project data management roles outlined, the NIH should consider including post-project completion roles and responsibilities regarding ongoing access to shared data, data products, or associated resources. This would require another Plan sub-element that provides for what happens to data products and resources developed after the grant is done and a budgeting allowance for continuing costs associated with these activities that may not be linked to a repository. Data management responsibilities do not end at project completion.

Other Considerations Relevant to this DRAFT Policy Proposal:

Reflecting on the history of and experience with the NIH Public Access Policy, we encourage the NIH to take these learnings into account when it comes to data management and sharing. Strengthening statements on expectations for data management and sharing plans and on associated compliance measures will shorten the timeline for realizing their benefits. This effect was observed with the Guide Notice NOT-OD-13-042

(<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-13-042.html>), where statement of the results of non-compliance with the NIH Public Access Policy brought forward the realization of the promise of the Public Access Policy.

We appreciate the need for flexibility based on community needs and expectations, which is why we encourage peer review of data management and sharing plans, in addition to NIH staff review. In order to effectively use the time of these reviewers, we encourage a more standardized approach to the necessary elements of the Plan. Standard minimal requirements would speed plan writing and review and, upon a structured implementation, make it possible to use machine-facilitated methods for review, updating, and compliance. There are sufficient extant similar policies from which to derive a suite of minimally viable required Plan elements to achieve the most basic of these goals, resulting in more efficient Plan writing and review.

We also appreciate the motivation to lighten the load of applicants and reviewers by restricting the submission of Plans to later in the grant review process with reviews to be carried out by NIH staff, but we believe this minimizes the importance of data management and sharing. Centering these activities rather than relegating them to an afterthought outside of normal peer review grant processes emphasizes their importance, raises community awareness of how others are managing and sharing data, and enhances community consensus building on expectations for data management and sharing. Data management and sharing plans should be viewed as outlining one of the major contributions of any research project to the research community and should avoid being positioned as an administrative burden.

Submission ID: 1307

Date: 1/7/2020

Name: American Society of Bone and Mineral Research

Name of Organization: American Society of Bone and Mineral Research

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Clinical/Basic and Biomedical/Other

Type of Organization: Professional Org/Association

Role: Other

Role - Other: Scientific Researcher, Medical Provider

Domain of Research Most Important to You or Your Organization:

Bone, mineral, and musculoskeletal research

DRAFT NIH Policy for Data Management and Sharing

Section II: Definitions:

It would be helpful to add examples of what scientific data include, in addition to what it does not include.

Section V: Requirements:

The description of what is required by the NIH policy could be interpreted many ways, some of which may not fulfill the NIH mission (the data will be incomplete or buried in an obscure website). As it is written now, it is fairly open to interpretation.

It would be very helpful for NIH to provide examples for multiple types of projects – e.g., projects involving human subjects vs model organism, example of data sharing plans for big data (including metagenomic and metabolic data).

For the biological data, in vitro and in vivo data, tests and analyses, the guidelines make sense. The area regarding IP contents of the data does not seem well specified (e.g., technical validation, soft codes, imaging processing to obtain high quality data and tests, and design data). It would be better to provide guidelines for investigators on how to have a plan to share these kinds of data.

Some journals, for example those managed by the Nature Publishing Group, already require the submission of all raw data associated with accepted manuscripts in an Excel form and have these data available on the journal web site as a "Source data." We recommend a shared platform with journals so investigators do not have to prepare and upload twice the same dataset in different format and on different databases.

Section VI: Data Management and Sharing Plans:

Plan assessment sub-section: for Extramural Awards the "programmatic assessment by NIH staff" needs to be better defined. It is unclear if a submitted application will be voided if the plan is not appropriate, or the application returned with some comments and allowed to be resubmitted in the same review cycle.

Section VII: Compliance and Enforcement:

The new NIH policy is rather opaque on the implementation and costs of creating data that is usable, accessible to outsiders, and at the same time does not compromise confidentiality of the participants. This may not be feasible in small studies that have barely enough person time on the statistical side to perform the main statistical analysis. This policy would have unknown impact on new investigators. Finally, it is not feasible for studies with data use agreements that preclude sharing of such data. The Center for Medicaid Services is very particular how data can be accessed and used.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

In addition to saving time if raw data are associated with publication and their submission is coordinated between NIH and journals, it will save costs.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

This document should say something about availability of specific new coding methods for data analysis.

Section 1: Data Type:

Some raw data will be more explicit than what is shown in papers (for example, the number of electric shock in behavioral tests, but there are many others). Since these databases are meant

to be searchable by the general public, such experimental data involving live animals may be used to target labs or generate anti-animal research propaganda by animal right groups. It is unclear how these risks will be addressed in the Sharing plan.

For the raw data, the description should specify for images whether all images should be shown or if only a representative image is required when associated with a quantitative analysis (histomorphometry, densitometry...)

Section 2: Related Tools, Software and/or Code:

"An indication of whether specialized tools are needed to access or manipulate shared data to support replication or reuse, and name(s) of the needed tool(s) and software."

Data repositories such as GEO require us to submit data in the 'raw form'. If someone wants access to these data, they can utilize established tools to retrieve it and analyze it on their end. The above statement may make this process harder (i.e., need for special software/tools) and not easier as intended.

Section 4: Data Type:

"How the scientific data will be findable and whether a persistent unique identifier or other standard indexing tools will be used." To decrease burden for investigators, if there is a common platform to all journals this identifier could be the PubMed ID.

"In general, scientific data should be made available as soon as practicable, independent of award period and publication schedule." This sentence is unrealistic. Data cannot be made available before publication because of competition and requirement of "novelty" by top-tiers journals.

Other Considerations Relevant to this DRAFT Policy Proposal:

In general, how the Sharing Plan will be considered in study sections should be defined. If it is not to be considered as a criteria for review or even only as a non-scorable item, this should be stated. Because a Data Sharing Plan is discussed at the end of each review at the Study Section, examples would be very helpful.

Submission ID: 1308

Date: 1/7/2020

Name: Harry W. Orf, PhD

Name of Organization: Massachusetts General Hospital

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All types stated above

Type of Organization: Other

Type of Organization - Other: Academic Medical Center

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

All domains

Attachment:

NIH Policy for Data Management and Sharing - Dr. Orf 1-2020.pdf



MASSACHUSETTS
GENERAL HOSPITAL

RESEARCH INSTITUTE



HARVARD
MEDICAL SCHOOL

55 Fruit Street, Bui 240E
Boston, MA 02114-2696
Tel: 617-724-9079
Fax: 617-724-3377
E-mail: horf@mgh.harvard.edu

Harry W. Orf, Ph.D.
*Senior Vice President for Research/
Membership & General Hospital
Principals Association / Harvard
Medical School*

January 7 2020

Dr. Andrea Jackson-Dipina
Director of the Division of Scientific Data Sharing Policy
National Institutes of Health
Office of Science Policy
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

subject: Comments to DRAFT NIH Policy for Data Management and Sharing Plan (DMSP)

Dear Dr. Jackson-Dipina:

I am writing on behalf of Massachusetts General Hospital (MGH), a founding member of the Partners HealthCare system and an affiliate of the Harvard Medical School, MGH is ranked first in receipt of NIH funding. In FY 2019, the MGH research budget from all funding sources reached one billion dollars. MGH researchers generate significant amounts of research data which they in turn share with internal and external investigators and other institutions. My colleagues and I greatly appreciate the opportunity to respond to the NIH draft policy. While we recognize that data sharing is vitally important to the conduct of research, of equal importance are the tools and resources necessary to share data. A realistic, cost-effective approach is essential for determining how to promote a culture of data-sharing across all scientific disciplines. This has prompted us also to include some recommendations that go beyond the actual policy.

We were pleased to see the draft policy has been modified to require submission of data sharing plans at "Just-In-Time" (JIT) rather than at the initial application thereby reducing applicant burden. Based on the timing, we assume that plan details will not be considered part of merit review. While JIT submission will allow researchers more time to focus on the science being proposed, one potential drawback is that it will be challenging to budget costs for a plan at time of application when the details will be later finalized with NIH Program staff. We recommend that NIH provide the flexibility of allowing additional data management costs to be added to the budget at JIT based on the final negotiated data management plan. Furthermore, because many grantee institution offices are involved in reviewing and approving components of the data management sharing plans (DMSP), having feedback available on the status of the NIH review of the plan for those involved in the development and review process at the institution will be extremely helpful in order to manage a plan.

January 7, 2020
Page Two

We also thank the NIH for creating a DRAFT Policy that allows for flexibility across scientific disciplines by outlining minimal specific expectations for the NIH-wide Data Management and Sharing Plan and allowing each NIH *VC* to supplement with additional requirements as appropriate. We are somewhat concerned, however, that the possibility of separate requirements for each of the twenty-seven institutes and centers will create confusion in the awardee community. We strongly urge NIH, to the extent possible, to harmonize and develop consistent data sharing plans across all *VCs* for collecting the necessary information. At MGH, investigators are often funded by different NIH Institutes. Having to comply with different formats, dates, and requirements increases the potential for confusion and non-compliance. Rather than having each researcher develop their own plans in two pages or less, we recommend that NIH provide a basic common template to which investigators may add or subtract information. This would streamline the process and reduce burden for all submitters and reviewers. We also recommend that NIH consider a centralized location to host *UC* specific requirements as opposed to individual websites. One central location hosting information pertinent to data sharing will improve transparency and monitoring practices for both public and grantee communities. Standardization of data fields across *I/Cs* may also make the data more useful for future meta analyses of previously collected data.

Allowing faculty to create the specific plans applicable to their data is important to ensure that data are not made public before any security, privacy or IP restrictions or concerns are met. We strongly recommend that NIH include in the policy, or in its implementation resource, information to help researchers and the public understand what the legal, ethical, technical, security, or privacy restrictions might be and to ensure that appropriate options exist to address the myriad ways that these restrictions may present themselves. Such coordination across sensitive data sets left to the researchers alone would significantly add unfunded administrative burden. NIH has the unique opportunity to lead the community by creating field-specific data repositories. NIH-led data repositories will allow both the agency and the awardees to leverage resources, avoid duplication and disaggregation of valuable knowledge, and to curate and provide data in ways that maximize the public benefit.

We also recognize the importance of NIH guidance to assure consistent application of the draft policy at the institution level, and we would ask NIH for additional guidance on standards for uncontrolled access, de-identification, application of the NIH Certificate of Confidentiality Policy, consequences of participant withdrawal or ability for a participant to decline data sharing, and how requirements such as the Health Insurance Portability and Accountability Act and other data protection laws (e.g. European Union General Data Protection Regulation) apply, especially as the data could be used for commercial purposes through uncontrolled access.

The DRAFT policy indicates that "non-compliance with the NIH [*I/CJ*]-approved Plan may be taken into account by the funding NIH [*VC*] for future funding decisions for the recipient institution." It would be helpful to know more about how non-compliance will be assessed by NIH, particularly since a data management and sharing plan (DMSP) is by definition a *plan*, whose implementation is dependent on the progress of the research and requires descriptions such as anticipated timeframes and anticipated agreements that could limit the ability to share

January 7, 2020

Page Three

scientific data broadly. For example, if deposited data were not yet analyzed and ready for publication, it is unlikely to meet the overall intent of "reproducibility."

A finding of non-compliance after the end of the funding period should be limited to the situation where there was failure to follow a DSMP *during* the funding period, or to other actions related to data sharing and management *during* the funding period. NIH should consider whether the policy applies to the data set that is available at the end of the funding period, or whether the data desired and requested must necessarily rely on more fully contemplated resources needed after the end of the award period. One potential solution would be to create a data sharing mechanism using modular budgeting that could be a supplement and extension to every award - a *de facto* addition of a sixth year to each standard RO1 or an appropriate equivalent for each funding mechanism.

The DRAFT policy contains the following statement, "*NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public*". If a repository with recurring fees is the only viable option, will the grantee be required to cover the costs once the project is over, and if so, for how long? Also, the determination of usefulness is necessarily a subjective one that is best made by the investigator. We ask that NIH continue to discuss the allowable costs guidance of the data sharing policy with stakeholders at future roundtable meetings or other public forums.

We also note that the DRAFT Policy applies to all scientific data generated from NIH-funded or conducted research and is written with the expectation that reasonable efforts will be made to digitize all scientific data. The February 22, 2013 [memo from OSTP](#) to departments and agencies significantly applies only to digital data. This expectation that non-digital data will be digitized creates a new, complex and costly burden for researchers, administrators and their institutions. This could serve as a disincentive to participate in research for smaller institutions, new or junior faculty, or interdisciplinary scientists. This could also create undue competition and unfunded requirements on non-profit institutions whose mission is primarily educational or patient care in the case of hospitals. There should be an option that allows investigators to appeal J/C mandated data sharing requirements to the NIH Policy Office should the requirements be considered unreasonable or inappropriate by the Principal Investigator(s) involved, without fear of reprisal.

The DRAFT policy for data management and sharing indicates that the plans should consider the life of the scientific data, and we applaud NIH for recognizing that each scientific area will have different length life cycles for the use of the data. However, all fields will be affected by evolution of technology, which over time will render current hardware and software necessary for accessing data obsolete. Migrating data to be compatible with future technology will be costly. In its policy guidance NIH should recognize that technological changes are inevitable and should not require investigators to attempt to predict such changes nor require institutions to incur such costs in the future.

The recommendation to apply this draft policy to **all** projects instead of at the current \$500K threshold will require significant additional resources, training, and time to implement. This

January 7, 2020

Page Four

should be taken into consideration when establishing an implementation date. We recommend that the policy apply to applications submitted on or after January 25, 2021 and apply only to new, competitive awards, as opposed to new and non-competing continuation awards. NIH should consider clarifying that the policy does not apply to awards (or activity codes) for which no data management plan is required as a condition of the award.

Finally, we suggest that NIH take into account the feedback received by OSTP in its [Request for Information](#), particularly with respect to research rigor and reproducibility. We also suggest that, prior to the implementation of a policy, NIH consider the creation of a Good Research Practices (similar to Good Clinical Practices) standard that addresses DMSP, including standards for research data standards for archival, and standards for data collection/design/purpose. We also recommend that NIH consider the issues and potential solutions related to data sharing raised in the publication 'Good Practices for University Open-Access Policies' published by the Harvard Open Access Project (available via wiki at http://cyber.harvard.edu/hoap/Good_practice_for_university_open-access_policy). While this work was primarily aimed at open access for scholarly articles its principles can also be applied to data sets.

Thank you for the opportunity to comment. If you have any questions please do not hesitate to contact me.

Yours sincerely



Harry W. Orf, PhD
Senior Vice President for Research

Submission ID: 1309

Date: 1/7/2020

Name: Kerry Ressler, MD, PhD

Name of Organization: McLean Hospital

Type of Data of Primary Interest: Genomic

Type of Organization: Other

Type of Organization - Other: Psychiatric Hospital

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

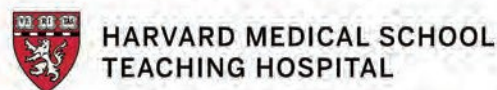
Insert neuroscience, psychiatric, clinical and basic science

Attachment:

NIH DMSP McLean Hospital 1.7.20.docx

Description:

Comment Letter



Kerry Ressler, MD, PhD

Chief Scientific Officer
James and Patricia Poitras Chair in Psychiatry
Chief, Division of Depression & Anxiety Disorders
McLean Hospital

Professor of Psychiatry, Harvard Medical School

January 7, 2020

Dr. Andrea Jackson-Dipina
Director of the Division of Scientific Data Sharing Policy
National Institutes of Health
Office of Science Policy
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Subject: Comments to DRAFT NIH Policy for Data Management and Sharing Plan (DMSP)

Dear Dr. Jackson-Dipina:

I am writing on behalf of McLean Hospital a member of the Partners HealthCare system and the leading psychiatric hospital affiliate of Harvard Medical School. McLean researchers generate significant amounts of research data which they in turn share with internal and external investigators and other institutions. My colleagues and I greatly appreciate the opportunity to respond to the NIH draft policy. While we recognize that data sharing is vitally important to the conduct of research, of equal importance are the tools and resources necessary to share data. A realistic, cost-effective approach is essential for determining how to promote a culture of data-sharing across all scientific disciplines. This has prompted us also to include some recommendations that go beyond the actual policy.

We were pleased to see the draft policy has been modified to require submission of data sharing plans at “Just-In-Time” (JIT) rather than at the initial application thereby reducing applicant burden. Based on the timing, we assume that plan details will not be considered part of merit review. While JIT submission will allow researchers more time to focus on the science being proposed, one potential drawback is that it will be challenging to budget costs for a plan at time of application when the details will be later finalized with NIH Program staff. We recommend that NIH provide the flexibility of allowing additional data management costs to be added to the budget at JIT based on the final negotiated data management plan. Furthermore, because many grantee institution offices are involved in reviewing and approving components of the data management sharing plans (DMSP), having feedback available on the status of the NIH review of the plan for those involved in

115 Mill Street, Mail Stop 212, Belmont, MA 02478-1064
T: 617.855.4210 F: 617.855.4213 E: kressler@mclean.harvard.edu

www.mcleanhospital.org



McLean Hospital is a member of Partners HealthCare.

the development and review process at the institution will be extremely helpful in order to manage a plan.

We also thank the NIH for creating a DRAFT Policy that allows for flexibility across scientific disciplines by outlining minimal specific expectations for the NIH-wide Data Management and Sharing Plan and allowing each NIH I/C to supplement with additional requirements as appropriate. We are somewhat concerned, however, that the possibility of separate requirements for each of the twenty-seven institutes and centers will create confusion in the awardee community. We strongly urge NIH, to the extent possible, to harmonize and develop consistent data sharing plan formats across all I/Cs for collecting the necessary information. At McLean investigators are often funded by different NIH Institutes. Having to comply with different formats, dates, and requirements increases the potential for confusion and non-compliance. Rather than having each researcher develop their own plans in two pages or less, we recommend that NIH provide a basic common template to which investigators may add or subtract information. This would streamline the process and reduce burden for all submitters and reviewers. We also recommend that NIH consider a centralized location to host I/C specific requirements as opposed to individual websites. One central location hosting information pertinent to data sharing will improve transparency and monitoring practices for both public and grantee communities. Standardization of data fields across I/Cs may also make the data more useful for future meta analyses of previously collected data.

Allowing faculty to create the specific plans applicable to their data is important to ensure that data are not made public before any security, privacy or IP restrictions or concerns are met. We strongly recommend that NIH include in the policy, or in its implementation resource, information to help researchers and the public understand what the legal, ethical, technical, security, or privacy restrictions might be and to ensure that appropriate options exist to address the myriad ways that these restrictions may present themselves. Such coordination across sensitive data sets left to the researchers alone would significantly add unfunded administrative burden. NIH has the unique opportunity to lead the community by creating field-specific data repositories. NIH-led data repositories will allow both the agency and the awardees to leverage resources, avoid duplication and disaggregation of valuable knowledge, and to curate and provide data in ways that maximize the public benefit.

We also recognize the importance of NIH guidance to assure consistent application of the draft policy at the institution level, and we would ask NIH for additional guidance on standards for uncontrolled access, de-identification, application of the NIH Certificate of Confidentiality Policy, consequences of participant withdrawal or ability for a participant to decline data sharing, and how requirements such as the Health Insurance Portability and Accountability Act and other data protection laws (e.g. European Union General Data Protection Regulation) apply, especially as the data could be used for commercial purposes through uncontrolled access.

The DRAFT policy indicates that “non-compliance with the NIH [I/C]-approved Plan may be taken into account by the funding NIH [I/C] for future funding decisions for the recipient institution.” It would be helpful to know more about how non-compliance will be assessed by NIH, particularly since a data management and sharing plan (DMSP) is by definition a

plan, whose implementation is dependent on the progress of the research and requires descriptions such as anticipated timeframes and anticipated agreements that could limit the ability to share scientific data broadly. For example, if deposited data were not yet analyzed and ready for publication, it is unlikely to meet the overall intent of “reproducibility.”

A finding of non-compliance after the end of the funding period should be limited to the situation where there was failure to follow a DSMP *during* the funding period, or to other actions related to data sharing and management *during* the funding period. NIH should consider whether the policy applies to the data set that is available at the end of the funding period, or whether the data desired and requested must necessarily rely on more fully contemplated resources needed after the end of the award period. One potential solution would be to create a data sharing mechanism using modular budgeting that could be a supplement and extension to every award – a *de facto* addition of a sixth year to each standard R01 or an appropriate equivalent for each funding mechanism.

The DRAFT policy contains the following statement, “*NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public*”. If a repository with recurring fees is the only viable option, will the grantee be required to cover the costs once the project is over, and if so, for how long? Also, the determination of usefulness is necessarily a subjective one that is best made by the investigator. We ask that NIH continue to discuss the allowable costs guidance of the data sharing policy with stakeholders at future roundtable meetings or other public forums.

We also note that the DRAFT Policy applies to all scientific data generated from NIH-funded or conducted research and is written with the expectation that reasonable efforts will be made to digitize all scientific data. The February 22, 2013 [memo from OSTP](#) to departments and agencies significantly applies only to digital data. This expectation that non-digital data will be digitized creates a new, complex and costly burden for researchers, administrators and their institutions. This could serve as a disincentive to participate in research for smaller institutions, new or junior faculty, or interdisciplinary scientists. This could also create undue competition and unfunded requirements on non-profit institutions whose mission is primarily educational or patient care in the case of hospitals. There should be an option that allows investigators to appeal I/C mandated data sharing requirements to the NIH Policy Office should the requirements be considered unreasonable or inappropriate by the Principal Investigator(s) involved, without fear of reprisal.

The DRAFT policy for data management and sharing indicates that the plans should consider the life of the scientific data, and we applaud NIH for recognizing that each scientific area will have different length life cycles for the use of the data. However, all fields will be affected by evolution of technology, which over time will render current hardware and software necessary for accessing data obsolete. Migrating data to be compatible with future technology will be costly. In its policy guidance NIH should recognize that technological changes are inevitable and should not require investigators to attempt to predict such changes nor require institutions to incur such costs in the future.

The recommendation to apply this draft policy to *all* projects instead of at the current \$500K threshold will require significant additional resources, training, and time to implement. This should be taken into consideration when establishing an implementation date.

We recommend that the policy apply to applications submitted on or after January 25, 2021 and apply only to new, competitive awards, as opposed to new and non-competing continuation awards. NIH should consider clarifying that the policy does not apply to awards (or activity codes) for which no data management plan is required as a condition of the award.

Finally, we suggest that NIH take into account the feedback received by OSTP in its current [Request for Information](#) , particularly with respect to research rigor and reproducibility. We also suggest that, prior to the implementation of a policy, NIH consider the creation of a Good Research Practices (similar to Good Clinical Practices) standard that addresses DMSP, including standards for research data, standards for archival, and standards for data collection/design/purpose. We also recommend that NIH consider the issues and potential solutions related to data sharing raised in the publication “Good Practices for University Open-Access Policies” published by the Harvard Open Access Project (available via wiki at https://cyber.harvard.edu/hoap/Good_practices_for_university_open-access_policies). While this work was primarily aimed at open access for scholarly articles, its principles can also be applied to data sets.

Thank you for the opportunity to comment. If you have any questions, please do not hesitate to contact me.

Yours sincerely,

A handwritten signature in black ink, appearing to read "Kerry Ressler". The signature is fluid and cursive, with a long horizontal stroke extending to the right.

Kerry Ressler, MD, PhD
Chief Scientific Officer

Submission ID: 1310

Date: 1/7/2020

Name: Paul Anderson

Name of Organization: BWH

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Clinical, genomic, basic and all types cited above

Type of Organization: Other

Type of Organization - Other: Academic Medical Center

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

All domains

Attachment:

NIH Policy Form.pdf

Paul J. Anderson, M.D., Ph.D.

Chief Academic Officer and Senior Vice President of Research
 K Frank Austen Professor of Medicine, Harvard Medical School

75 Francis Street, Boston MA 02115
 Tel: 617-732-8990, Fax: 617-732-5343
 Email: panderson@bwh.harvard.edu

January 7, 2020

Dr. Andrea Jackson-Dipina
 Director of the Division of Scientific Data Sharing Policy
 National Institutes of Health
 Office of Science Policy
 6705 Rockledge Drive, Suite 750
 Bethesda, MD 20892
 Subject: Comments to DRAFT NIH Policy for Data Management and Sharing Plan (DMSP)

Dear Dr. Jackson-Dipina:

I am writing on behalf of Brigham and Women's Hospital (BWH), a founding member of the Partners HealthCare system and an affiliate of the Harvard Medical School, BWH is ranked second in receipt of NIH funding. BWH researchers generate significant amounts of research data which they in turn share with internal and external investigators and other institutions. My colleagues and I greatly appreciate the opportunity to respond to the NIH draft policy. While we recognize that data sharing is vitally important to the conduct of research, of equal importance are the tools and resources necessary to share data. A realistic, cost-effective approach is essential for determining how to promote a culture of data-sharing across all scientific disciplines. This has prompted us also to include some recommendations that go beyond the actual policy.

We were pleased to see the draft policy has been modified to require submission of data sharing plans at "Just-In-Time" (JIT) rather than at the initial application thereby reducing applicant burden. Based on the timing, we assume that plan details will not be considered part of merit review. While JIT submission will allow researchers more time to focus on the science being proposed, one potential drawback is that it will be challenging to budget costs for a plan at time of application when the details will be later finalized with NIH Program staff. We recommend that NIH provide the flexibility of allowing additional data management costs to be added to the budget at JIT based on the final negotiated data management plan. Furthermore, because many grantee institution offices are involved in reviewing and approving components of the data management sharing plans (DMSP), having feedback available on the status of the NIH review of the plan for those involved in the development and review process at the institution will be extremely helpful in order to manage a plan.

We also thank the NIH for creating a DRAFT Policy that allows for flexibility across scientific disciplines by outlining minimal specific expectations for the NIH-wide Data Management and Sharing Plan and allowing each NIH I/C to supplement with additional requirements as

appropriate. We are somewhat concerned, however, that the possibility of separate requirements for each of the twenty-seven institutes and centers will create confusion in the awardee community. We strongly urge NIH, to the extent possible, to harmonize and develop consistent data sharing plan formats across all I/Cs for collecting the necessary information. At BWH, investigators are often funded by different NIH Institutes. Having to comply with different formats, dates, and requirements increases the potential for confusion and non-compliance. Rather than having each researcher develop their own plans in two pages or less, we recommend that NIH provide a basic common template to which investigators may add or subtract information. This would streamline the process and reduce burden for all submitters and reviewers. We also recommend that NIH consider a centralized location to host I/C specific requirements as opposed to individual websites. One central location hosting information pertinent to data sharing will improve transparency and monitoring practices for both public and grantee communities. Standardization of data fields across I/Cs may also make the data more useful for future meta analyses of previously collected data.

Allowing faculty to create the specific plans applicable to their data is important to ensure that data are not made public before any security, privacy or IP restrictions or concerns are met. We strongly recommend that NIH include in the policy, or in its implementation resource, information to help researchers and the public understand what the legal, ethical, technical, security, or privacy restrictions might be and to ensure that appropriate options exist to address the myriad ways that these restrictions may present themselves. Such coordination across sensitive data sets left to the researchers alone would significantly add unfunded administrative burden. NIH has the unique opportunity to lead the community by creating field-specific data repositories. NIH-led data repositories will allow both the agency and the awardees to leverage resources, avoid duplication and disaggregation of valuable knowledge, and to curate and provide data in ways that maximize the public benefit.

We also recognize the importance of NIH guidance to assure consistent application of the draft policy at the institution level, and we would ask NIH for additional guidance on standards for uncontrolled access, de-identification, application of the NIH Certificate of Confidentiality Policy, consequences of participant withdrawal or ability for a participant to decline data sharing, and how requirements such as the Health Insurance Portability and Accountability Act and other data protection laws (e.g. European Union General Data Protection Regulation) apply, especially as the data could be used for commercial purposes through uncontrolled access.

The DRAFT policy indicates that “non-compliance with the NIH [I/C]-approved Plan may be taken into account by the funding NIH [I/C] for future funding decisions for the recipient institution.” It would be helpful to know more about how non-compliance will be assessed by NIH, particularly since a data management and sharing plan (DMSP) is by definition a *plan*, whose implementation is dependent on the progress of the research and requires descriptions such as anticipated timeframes and anticipated agreements that could limit the ability to share scientific data broadly. For example, if deposited data were not yet analyzed and ready for publication, it is unlikely to meet the overall intent of “reproducibility.”

A finding of non-compliance after the end of the funding period should be limited to the situation where there was failure to follow a DSMP *during* the funding period, or to other actions related to data sharing and management *during* the funding period. NIH should consider whether the policy applies to the data set that is available at the end of the funding period, or whether the data

desired and requested must necessarily rely on more fully contemplated resources needed after the end of the award period. One potential solution would be to create a data sharing mechanism using modular budgeting that could be a supplement and extension to every award – a *de facto* addition of a sixth year to each standard R01 or an appropriate equivalent for each funding mechanism.

The DRAFT policy contains the following statement, “*NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public*”. If a repository with recurring fees is the only viable option, will the grantee be required to cover the costs once the project is over, and if so, for how long? Also, the determination of usefulness is necessarily a subjective one that is best made by the investigator. We ask that NIH continue to discuss the allowable costs guidance of the data sharing policy with stakeholders at future roundtable meetings or other public forums.

We also note that the DRAFT Policy applies to all scientific data generated from NIH-funded or conducted research and is written with the expectation that reasonable efforts will be made to digitize all scientific data. The February 22, 2013 [memo from OSTP](#) to departments and agencies significantly applies only to digital data. This expectation that non-digital data will be digitized creates a new, complex and costly burden for researchers, administrators and their institutions. This could serve as a disincentive to participate in research for smaller institutions, new or junior faculty, or interdisciplinary scientists. This could also create undue competition and unfunded requirements on non-profit institutions whose mission is primarily educational or patient care in the case of hospitals. There should be an option that allows investigators to appeal I/C mandated data sharing requirements to the NIH Policy Office should the requirements be considered unreasonable or inappropriate by the Principal Investigator(s) involved, without fear of reprisal.

The DRAFT policy for data management and sharing indicates that the plans should consider the life of the scientific data, and we applaud NIH for recognizing that each scientific area will have different length life cycles for the use of the data. However, all fields will be affected by evolution of technology, which over time will render current hardware and software necessary for accessing data obsolete. Migrating data to be compatible with future technology will be costly. In its policy guidance NIH should recognize that technological changes are inevitable and should not require investigators to attempt to predict such changes nor require institutions to incur such costs in the future.

The recommendation to apply this draft policy to *all* projects instead of at the current \$500K threshold will require significant additional resources, training, and time to implement. This should be taken into consideration when establishing an implementation date. We recommend that the policy apply to applications submitted on or after January 25, 2021 and apply only to new, competitive awards, as opposed to new and non-competing continuation awards. NIH should consider clarifying that the policy does not apply to awards (or activity codes) for which no data management plan is required as a condition of the award.

Finally, we suggest that NIH take into account the feedback received by OSTP in its current [Request for Information](#), particularly with respect to research rigor and reproducibility. We also suggest that, prior to the implementation of a policy, NIH consider the creation of a Good Research Practices (similar to Good Clinical Practices) standard that addresses DMSP, including standards for research data, standards for archival, and standards for data

collection/design/purpose. We also recommend that NIH consider the issues and potential solutions related to data sharing raised in the publication “Good Practices for University Open-Access Policies” published by the Harvard Open Access Project (available via wiki at https://cyber.harvard.edu/hoap/Good_practices_for_university_open-access_policies). While this work was primarily aimed at open access for scholarly articles, its principles can also be applied to data sets.

Thank you for the opportunity to comment. If you have any questions, please do not hesitate to contact me.

Yours sincerely,

A handwritten signature in black ink that reads "Paul Anderson". The signature is written in a cursive style with a large initial "P" and "A".

Paul Anderson, MD, PhD
Chief Academic Officer

Submission ID: 1311

Date: 1/7/2020

Name: Ravi Thadhani, M.D., MPH

Name of Organization: Partners HealthCare

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: clinical, genomic, and all types cited above

Type of Organization: Other

Type of Organization - Other: Academic Medical Center

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

All Domains

Attachment:

Comments-DMSP-fromRThadhani-Jan2020.pdf

Description:

comment letter

January 7, 2020

Dr. Andrea Jackson-Dipina
Director of the Division of Scientific Data Sharing Policy
National Institutes of Health
Office of Science Policy
6705 Rockledge Drive, Suite 750
Bethesda MD 20892

Subject: Comments to DRAFT NIH Policy for Data Management and Sharing Plan (DMSP)

Dear Dr. Jackson-Dipina:

I am writing on behalf of Partners HealthCare, a not-for-profit healthcare system committed to patient care, research, teaching, and service to the local community. Several Partners hospitals, most notably Brigham and Women's Hospital (BWH), Massachusetts General Hospital (MGH), McLean Hospital, and Massachusetts Eye and Ear Infirmary, are affiliated with Harvard Medical School and are recipients of substantial NIH research funding. MGH and BWH are ranked first and second respectively in NIH funding. Partners hospitals generate significant amounts of research data which in turn are shared with internal and external investigators and institutions. We greatly appreciate the opportunity to respond to the NIH draft policy. While we recognize that data sharing is vitally important to the conduct of research, of equal importance are the tools and resources necessary to share data. A realistic, cost-effective approach is essential for determining how to promote a culture of data-sharing across all scientific disciplines. This has prompted us also to include in our comments some recommendations that go beyond the actual policy.

We were pleased to see the draft policy has been modified to require submission of data sharing plans at "Just-In-Time" (JIT) rather than at the initial application thereby reducing applicant burden. Based on the timing, we assume that plan details will not be considered part of merit review. While JIT submission will allow researchers more time to focus on the science being proposed, one potential drawback is that it will be challenging to budget costs for a plan at time of application when the details will be later finalized with NIH Program staff. We recommend that NIH provide the flexibility of allowing additional data management costs to be added to the budget at JIT based on the final negotiated data management plan. Furthermore, because many grantee institution offices are involved in reviewing and approving components of the data management sharing plans (DMSP), having feedback available on the status of the NIH review of the plan for those involved in the development and review process at the institution will be extremely helpful in order to manage a plan.

Ravi T. Thadhani, M.D., MPH.

Chief Academic Officer, Partners HealthCare Professor of Medicine, Harvard Medical School
Director of Academic Programs at Partners HealthCare, Harvard Medical School

Letter to Dr. Andrea Jackson-Dipina from Dr. Ravi Thad.bani

Page2

We also thank the NIH for creating a DRAFT Policy that allows for flexibility across scientific disciplines by outlining minimal specific expectations for the NIH-wide Data Management and Sharing Plan and allowing each NIH I/C to supplement with additional requirements as appropriate. We are somewhat concerned, however, that the possibility of separate requirements for each of the twenty-seven institutes and centers will create confusion in the awardee community. We strongly urge NIH, to the extent possible, to harmonize and develop consistent data sharing plans across all I/Cs for collecting the necessary information. At Partner institutions, investigators are often funded by different NIH Institutes. Having to comply with different formats and requirements increases the potential for confusion and non-compliance. Rather than having each researcher develop their own plans in two pages or less, we recommend that NIH provide a basic common template to which investigators may add or subtract information. This would streamline the process and reduce burden for all submitters and reviewers. We also recommend that NIH consider a centralized location to host I/C specific requirements as opposed to individual websites. One central location hosting information pertinent to data sharing will improve transparency and monitoring practices for both public and grantee communities. Standardization of data fields across I/Cs may also make the data more useful for future meta analyses of previously collected data.

Allowing faculty to create the specific plans applicable to their data is important to ensure that data are not made public before any security, privacy or IP restrictions or concerns are met. We strongly recommend that NIH include in the policy or in its implementation resource, information to help researchers and the public understand what the legal ethical technical, security or privacy restrictions might be and to ensure that appropriate options exist to address the myriad ways that these restrictions may present themselves. Much coordination across sensitive data sets left to the researchers alone would significantly add unfunded administrative burden. NIH has the unique opportunity to lead the community by creating field-specific data repositories. NIH-led data repositories will allow both the agency and the awardees to leverage resources, avoid duplication and disaggregation of valuable knowledge, and to curate and provide data in ways that maximize the public benefit.

We also recognize the importance of NIH guidance to assure consistent application of the draft policy at the institution level and we would ask NIH for additional guidance on standards for uncontrolled access, de-identification, application of the NIH Certificate of Confidentiality Policy, consequences of participant withdrawal or ability for a participant to decline data sharing, and how requirements such as the Health Insurance Portability and Accountability Act and other data protection laws (e.g. European Union General Data Protection Regulation) apply, especially as the data could be used for commercial purposes through uncontrolled access.

The DRAFT policy indicates that 'non-compliance with the NIH [I/C]-approved Plan may be taken into account by the funding NIH [I/C] for future funding decisions for the recipient institution.' It would be helpful to know more about how non-compliance will be assessed by NIH, particularly since a data management and sharing plan (DMSP) is by definition a *plan* whose implementation is dependent on the progress of the research and requires descriptions such as anticipated timeframes and anticipated agreements that could limit the ability to share scientific data broadly. For example if deposited data were not yet analyzed and ready for publication, it is unlikely to meet the overall intent of 'reproducibility.'

A finding of non-compliance after the end of the funding period should be limited to the situation where there was failure to follow a DSMP *during* the funding period, or to other actions related to data sharing and management *during* the funding period. III should consider whether the policy applies to the data set that is available at the end of the funding period, or whether the data desired and requested must necessarily rely on more fully contemplated resources needed after the end of the award period. One potential solution would be to create a data sharing mechanism using modular budgeting that could be a supplement and extension to every award - a *de facto* addition of a sixth year to each standard RO1 or an appropriate equivalent for each funding mechanism.

The DRAFT policy contains the following statement, "*NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public*". If a repository with recwTing fees is the only viable option, will the grantee be required to cover the costs once the project is over and if so, for how long? Also the determination of usefulness is necessarily a subjective one that is best made by the investigator. We ask that NIH continue to discuss the allowable costs guidance of the data sharing policy with stakeholders at future roundtable meetings or other public forums.

We also note that the DRAFT Policy applies to all scientific data generated from NIH-funded or conducted research and is written with the expectation that reasonable efforts will be made to digitize all scientific data. The February 22, 2013 [memo from TP](#) to departments and agencies significantly applies only to digital data. This expectation that non-digital data will be digitized creates a new, complex and costly burden for researchers, administrators and their institutions. This could serve as a disincentive to participate in research for smaller institutions new or junior faculty, or interdisciplinary scientists. This could also create undue competition and unfunded requirements on non-profit institutions whose mission is primarily educational or patient care in the case of hospitals. There should be an option that allows investigators to appeal I/C mandated data sharing requirements to the NIH Policy Office should the requirements be considered unreasonable or inappropriate by the Principal Investigator(s) involved, without fear of reprisal.

The DRAFT policy for data management and sharing indicates that the plans should consider the life of the scientific data, and we applaud NIH for recognizing that each scientific area will have different length life cycles for the use of the data. However all fields will be affected by evolution of technology, which over time will render current hardware and software necessary for accessing data obsolete. Migrating data to be compatible with future technology will be costly. In its policy guidance NIH should recognize that technological changes are inevitable and should not require investigators to attempt to predict such changes nor require institutions to incur such costs in the future.

The recommendation to apply this draft policy to *all* projects instead of at the current \$500K threshold will require significant additional resources training, and time to implement. This should be taken into consideration when establishing an implementation date. We recommend that the policy apply to applications submitted on or after January 25, 2021 and apply only to new, competitive awards, as opposed to new and non-competing continuation awards. NIH

Letter to Dr. Andrea Jackson-Dipina from Dr. Ravi Thadhani
Page 4

should consider clarifying that the policy does not apply to awards (or activity codes for which no data management plan is required as a condition of the award.

Finally, we suggest that NIH take into account the feedback received by OSTP in its current [Request for Information](#), particularly with respect to research rigor and reproducibility. We also suggest that, prior to the implementation of a policy, NIH consider the creation of a Good Research Practices (similar to Good Clinical Practices) standard that addresses DMSR, including standards for research data, standards for archival, and standards for data collection/design/purpose. We also recommend that NIH consider the issues and potential solutions related to data sharing raised in the publication "Good Practices for University Open-Access Policies" published by the Harvard Open Access Project (available via wiki at https://cyber.harvard.edu/hoap/Good_practices_for_university_open-access_policy). While this work was primarily aimed at open access for scholarly articles, its principles can also be applied to data sets.

Thank you for the opportunity to comment. If you have any questions, please do not hesitate to contact me.

Yours sincerely,



Ravi Thadhani, MD, MPH

Submission ID: 1312

Date: 1/7/2020

Name: Lauren Gross

Name of Organization: The American Association of Immunologists

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Data of Primary Interest - Other:

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Domain of Research Most Important to You or Your Organization:

immunology and related fields

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The American Association of Immunologists (AAI) is the nation's largest professional society of research scientists and physicians who are dedicated to understanding the immune system through basic, translational, and/or clinical research. Founded in 1913, AAI members are scientists at all career stages and from all sectors of research – academic, government, corporate, and non-profit. The vast majority of AAI members receives NIH funding to conduct research on critically important and promising areas of immunology. These discoveries have laid the foundation for extraordinary advances in preventing and treating disease; some recent advances, such as immunotherapies to treat certain cancers, have achieved unimaginable success.

While AAI supports the principle of data sharing and agrees with the need for effective data sharing, preservation, and management, AAI is concerned about some of the details, including those related to implementation, that are described in the draft policy; these concerns are addressed in more detail below. As a preliminary matter, AAI believes it is essential for NIH to describe clearly the specific goals of the policy, including the currently perceived deficit in data sharing. Clarification of the goals will help the research community determine exactly which data should be shared, and at what point it should be shared. NIH should also explain how this policy will interact with preexisting data sharing policies. Furthermore, NIH should describe what metrics will be used to evaluate the policy, especially to determine the value added by this policy and to ensure that there are not unintended consequences. Finally, once the policy

is modified as a result of comments submitted, it is crucial that NIH provide stakeholders with the opportunity to comment on the revised proposed policy.

Section II: Definitions:

AAI believes that additional clarification of the terms "scientific data" and "preliminary analyses," as well as definitions of what would be considered "negative data" and "unpublished data," are needed. NIH should provide specific examples of what would and would not be covered under each. While AAI agrees that there is value in sharing some negative/unpublished results, there is minimal value in sharing uninterpretable data (for example, experiments in which critical controls failed). NIH should recognize the large variability in the quality of data, and therefore should provide clarification about exactly which data NIH seeks to be included.

Section III: Scope:

AAI agrees with the scope specified: that the policy should apply "to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data."

Section IV: Effective Date(s):

AAI appreciates having the opportunity to comment on the draft policy. However, as stated above, AAI urges NIH to issue another request for comments after revisions are made to the draft policy. Additional policy details are needed for the community to be able to provide the most thoughtful and thorough feedback. As a result, it may be that the anticipated effective date (2022) will have to be delayed so that NIH can address these additional comments.

Section V: Requirements:

In developing a final policy, AAI hopes that NIH will address the following questions: 1) how the data will be curated and determined to be useful; 2) how to ensure the quality and monitoring of data (including how it will be updated and/or corrected if needed); 3) how to ensure the data will be accessible; 4) how long data will need to be stored; and 5) how to ensure the needed infrastructure is available (given that established repositories may not be able to handle the volume of data that would be stored as a result of this policy). Further, in order to reduce administrative burden with this new policy, AAI requests that NIH coordinate with, clarify, and disseminate any new policies or expectations by individual Institutes, Centers, and Offices (ICO).

Section VI: Data Management and Sharing Plans:

AAI greatly appreciates that the NIH draft policy provides significant flexibility to investigators as they develop their individual data sharing plans; by so doing, NIH recognizes the diversity of data generated and the need for limitations in sharing due to patient privacy, intellectual property protections (including those rights impacted by the date of publication), biosecurity implications or threats to public health or safety, and other relevant issues. However, to ensure that investigators are well-positioned to submit an appropriate and useful plan, NIH should

clarify exactly what data needs to be shared, in what format (e.g., raw v. processed), where (which repositories would be acceptable), and when. Without some level of specificity, NIH could receive a wide range of plans, some of which may not satisfy NIH goals or result in more burden on investigators and program officers.

In addition, AAI is concerned that, as a result of this (appreciated) flexibility, there may be a significant lack of standardization within and across ICOs. If the discretion to approve a plan is given, per the draft policy, to ICO program officers, AAI encourages NIH to train, and set parameters for, program officers, in order to ensure fairness and basic consistency. Furthermore, NIH should make researchers aware of the criteria that program officers will use to evaluate the plans.

AAI also greatly appreciates that NIH will require the plan to be submitted during Just-in-Time for extramural awards; this will minimize the potential increase in administrative burden.

AAI also strongly encourages NIH to consider how to ensure the security of shared data, as well as the implications if plans were to be made publicly available. This is an area that requires additional stakeholder input following NIH's release of a draft revised policy.

Section VII: Compliance and Enforcement:

AAI requests that NIH clarify who will be responsible for monitoring and enforcing the policy. If compliance is expected beyond the award term, clarification is also needed as to how this is to be monitored and enforced.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

As there are significant costs associated with data sharing and management, AAI is pleased that NIH has included guidance in this area. However, AAI is concerned that this could essentially become an unfunded mandate if additional funds are not made available for this purpose. Clarification is needed as to whether these costs can be covered by a supplement to a grant, or if they must be supported by the grant itself (especially problematic for modular grants). Before a final policy is issued, AAI encourages NIH to determine the total cost associated with this policy, including the cost to researchers and institutions for data storage and retrieval. If the policy will also be enforced after the award term has ended, AAI requests further clarification as to how costs related to long-term data retention and management will be addressed.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

AAI appreciates that NIH has provided supplemental draft guidance. However, per the comments above, AAI urges NIH to strike a balance between the need for flexibility and the need for adequate guidance to ensure fairness and basic consistency.

Other Considerations Relevant to this DRAFT Policy Proposal:

AAI believes that this policy, if clarified as requested above, implemented with NIH's intended flexibility, and evaluated based on agreed-upon metrics, may enhance reliability in, and the reproducibility of, research findings, and be especially helpful to scientists seeking to re-use hard-to-generate data.

Attachment:

AAI response to NIH draft policy on data management and sharing.January 72020.pdf

Description:

AAI comments on draft NIH policy on data management and sharing

Comments of The American Association of Immunologists (AAI) on the Draft NIH Policy for Data Management and Sharing and Supplemental Draft Guidance

January 7, 2020

*Submitted on behalf of AAI by Lauren G. Gross, J.D.,
Director of Public Policy and Government Affairs*

Comments on the Draft NIH Policy for Data Management and Sharing:

I. Purpose

The American Association of Immunologists (AAI) is the nation's largest professional society of research scientists and physicians who are dedicated to understanding the immune system through basic, translational, and/or clinical research. Founded in 1913, AAI members are scientists at all career stages and from all sectors of research – academic, government, corporate, and non-profit. The vast majority of AAI members receives NIH funding to conduct research on critically important and promising areas of immunology. These discoveries have laid the foundation for extraordinary advances in preventing and treating disease; some recent advances, such as immunotherapies to treat certain cancers, have achieved unimaginable success.

While AAI supports the principle of data sharing and agrees with the need for effective data sharing, preservation, and management, AAI is concerned about some of the details, including those related to implementation, that are described in the draft policy; these concerns are addressed in more detail below. As a preliminary matter, AAI believes it is essential for NIH to describe clearly the specific goals of the policy, including the currently perceived deficit in data sharing. Clarification of the goals will help the research community determine exactly which data should be shared, and at what point it should be shared. NIH should also explain how this policy will interact with preexisting data sharing policies. Furthermore, NIH should describe what metrics will be used to evaluate the policy, especially to determine the value added by this policy and to ensure that there are not unintended consequences. Finally, once the policy is modified as a result of comments submitted, it is crucial that NIH provide stakeholders with the opportunity to comment on the revised proposed policy.

II. Definitions

AAI believes that additional clarification of the terms “scientific data” and “preliminary analyses,” as well as definitions of what would be considered “negative data” and “unpublished data,” are needed. NIH should provide specific examples of what would and would not be covered under each. While AAI agrees that there is value in sharing some negative/unpublished results, there is minimal value in sharing uninterpretable data (for example, experiments in which critical controls failed). NIH should recognize the large variability in the quality of data, and therefore should provide clarification about exactly which data NIH seeks to be included.

III. Scope

AAI agrees with the scope specified: that the policy should apply “to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data.”

IV. Effective Date

AAI appreciates having the opportunity to comment on the draft policy. However, as stated above, AAI urges NIH to issue another request for comments after revisions are made to the draft policy. Additional policy details are needed for the community to be able to provide the most thoughtful and thorough feedback. As a result, it may be that the anticipated effective date (2022) will have to be delayed so that NIH can address these additional comments.

V. Requirements

In developing a final policy, AAI hopes that NIH will address the following questions: 1) how the data will be curated and determined to be useful; 2) how to ensure the quality and monitoring of data (including how it will be updated and/or corrected if needed); 3) how to ensure the data will be accessible; 4) how long data will need to be stored; and 5) how to ensure the needed infrastructure is available (given that established repositories may not be able to handle the volume of data that would be stored as a result of this policy). Further, in order to reduce administrative burden with this new policy, AAI requests that NIH coordinate with, clarify, and disseminate any new policies or expectations by individual Institutes, Centers, and Offices (ICO).

VI. Data Management and Sharing Plans

AAI greatly appreciates that the NIH draft policy provides significant flexibility to investigators as they develop their individual data sharing plans; by so doing, NIH recognizes the diversity of data generated and the need for limitations in sharing due to patient privacy, intellectual property protections (including those rights impacted by the date of publication), biosecurity implications or threats to public health or safety, and other relevant issues. However, to ensure that investigators are well-positioned to submit an appropriate and useful plan, NIH should clarify exactly what data needs to be shared, in what format (e.g., raw v. processed), where (which repositories would be acceptable), and when. Without some level of specificity, NIH could receive a wide range of plans, some of which may not satisfy NIH goals or result in more burden on investigators and program officers.

In addition, AAI is concerned that, as a result of this (appreciated) flexibility, there may be a significant lack of standardization within and across ICOs. If the discretion to approve a plan is given, per the draft policy, to ICO program officers, AAI encourages NIH to train, and set parameters for, program officers, in order to ensure fairness and basic consistency. Furthermore, NIH should make researchers aware of the criteria that program officers will use to evaluate the plans.

AAI also greatly appreciates that NIH will require the plan to be submitted during Just-in-Time for extramural awards; this will minimize the potential increase in administrative burden.

AAI also strongly encourages NIH to consider how to ensure the security of shared data, as well as the implications if plans were to be made publicly available. This is an area that requires additional stakeholder input following NIH's release of a draft revised policy.

VII. Compliance and Enforcement

AAI requests that NIH clarify who will be responsible for monitoring and enforcing the policy. If compliance is expected beyond the award term, clarification is also needed as to how this is to be monitored and enforced.

Comments on the Supplemental Draft Guidance on Allowable Costs:

As there are significant costs associated with data sharing and management, AAI is pleased that NIH has included guidance in this area. However, AAI is concerned that this could essentially become an unfunded mandate if additional funds are not made available for this purpose. Clarification is needed as to whether these costs can be covered by a supplement to a grant, or if they must be supported by the grant itself (especially problematic for modular grants). Before a final policy is issued, AAI encourages NIH to determine the total cost associated with this policy, including the cost to researchers and institutions for data storage and retrieval. If the policy will also be enforced after the award term has ended, AAI requests further clarification as to how costs related to long-term data retention and management will be addressed.

Comments on the Supplemental Draft Guidance on Elements of a Data Management and Sharing Plan:

AAI appreciates that NIH has provided supplemental draft guidance. However, per the comments above, AAI urges NIH to strike a balance between the need for flexibility and the need for adequate guidance to ensure fairness and basic consistency.

Other Relevant Considerations:

AAI believes that this policy, if clarified as requested above, implemented with NIH's intended flexibility, and evaluated based on agreed-upon metrics, may enhance reliability in, and the reproducibility of, research findings, and be especially helpful to scientists seeking to re-use hard-to-generate data.

Submission ID: 1313

Date: 1/8/2020

Name: Carol Pulver

Name of Organization: Frontier Science Foundation

Type of Data of Primary Interest: Clinical

Type of Organization: Nonprofit Research Organization

Role: Other

Role - Other: Clinical and Laboratory Data Management

Domain of Research Most Important to You or Your Organization:

Epidemiology

DRAFT NIH Policy for Data Management and Sharing

Section VI: Data Management and Sharing Plans:

The policy suggests that Data Management and Sharing plans would be at the study level. Is it possible to have a more general project level plan and then study-specific plans that refer to the general plan and point out any items unique to that study?

- It was covered in the Webinar that this new policy will take into account other existing NIH policies, such as that on Human Genetic testing and consent. It would be helpful for this policy to directly reference the other relevant NIH policies for clarity.
- Please mention that researchers must ensure that participants are aware and agree to the use of their data as described in the Data Management and Sharing plan (DMSP). Informed

Consents are mentioned in the Supplemental Guidance for Elements of a DMSP ("NIH encourages the broadest use of scientific data resulting from NIH-funded or conducted research, consistent with privacy, security, informed consent, and proprietary issues"), but it would be useful to have this concept in the main Policy as well.

Section VII: Compliance and Enforcement:

May guidance be provided as to the expected frequency of reporting on the progress of the project or trial/protocol, as well as guidance on the expected timing and frequency of sharing the data in order to comply with expectations to maintain funding? Per the policy, the Data Management and Sharing plan would be reviewed at least annually by NIH ICO, but more detail on that would be helpful for planning purposes.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The page limit on a plan seems strict. Will referencing other documentation be allowed or will documents like these be expected to be brief summaries?

- Should the estimates in the Data Type section be based on protocols in development and should these estimates of the data planned to be collected be modified on some interval as participants enroll or as the study reaches milestones when amount of data becomes clearer?
- Should the data elements be only from primary or secondary endpoints and exclude exploratory objectives that team may not end up having funding to collect?
- Please clarify what constitutes specialized tools. Does a Statistical Data Management Center need to include the tools used for data management in the Data Management and Sharing plan?
- Oversight individuals can change over time. Can the types of roles be listed, or does the plan need to include the actual names of the individuals at the time and get maintained over the course of the study whenever there is a staffing change?

Submission ID: 1314

Date: 1/8/2020

Name: Meghan Faherty

Name of Organization: Jean Mayer USDA HNRCA at Tufts University

Type of Data of Primary Interest: Clinical

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization: Nutrition and Aging

DRAFT NIH Policy for Data Management and Sharing

Section VI: Data Management and Sharing Plans:

In order to facilitate proper credit for and sharing of research data, we recommend that the NIH guidelines incorporate recommendations for proper licensing of the published research data.

One common licensing strategy for research data is the Creative Commons Zero copyright license. This will provide other researchers with guidelines for proper use of publicly available data.

In addition to encouraging the use of established repositories for preserving and sharing scientific data, it is critical that investigators submit their data to the most appropriate repositories, such as dbGaP for genotype-phenotype interaction data or The Cancer Imaging Archive for medical images of cancer. Nature, has a list of recommended databases for scientific data that can serve as a good reference(<https://www.nature.com/sdata/policies/repositories#nuc>).

In situations where broad data sharing is not appropriate as listed in the RFI, we recommend the ability to be transparent in regard to what specific restriction prohibits or limits data sharing. This will provide clear rationale to the greater research community for why the data are not publicly available.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Lastly, in regard to allowable costs, it may be helpful if there is guidance on the minimum expected length of time data needs to be publicly available so that any associated costs can be properly determined and included in proposals.

Other Considerations Relevant to this DRAFT Policy Proposal:

The respondents for this RFI are a collection of data specialists and statisticians that have extensive experience in developing and overseeing data management, sharing, and retention requirements for this research center. They also have extensive experience developing and implementing data management plans for funded research that meet industry, foundation, and government (USDA, NIH and others) requirements. We appreciate that the NIH has provided this opportunity for individuals and research groups to provide feedback to the proposed Data Management and Sharing requirements for NIH submissions.

Attachment:

NIH_DataRFI_Responses_HNRCA Tufts University_01.08.2020.docx

Description:

Responses from Jean Mayer USDA HNRCA at Tufts University

Title of RFI: Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance Draft of NIH Data Management and Sharing
<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-013.html>

Institution: Jean Mayer USDA Human Nutrition Research Center on Aging (HNRCA) at Tufts University

The respondents for this RFI are a collection of data specialists and statisticians that have extensive experience in developing and overseeing data management, sharing, and retention requirements for this research center. They also have extensive experience developing and implementing data management plans for funded research that meet industry, foundation, and government (USDA, NIH and others) requirements. We appreciate that the NIH has provided this opportunity for individuals and research groups to provide feedback to the proposed Data Management and Sharing requirements for NIH submissions.

In order to facilitate proper credit for and sharing of research data, we recommend that the NIH guidelines incorporate recommendations for proper licensing of the published research data. One common licensing strategy for research data is the Creative Commons Zero copyright license. This will provide other researchers with guidelines for proper use of publicly available data.

In addition to encouraging the use of established repositories for preserving and sharing scientific data, it is critical that investigators submit their data to the most appropriate repositories, such as dbGaP for genotype-phenotype interaction data or The Cancer Imaging Archive for medical images of cancer. Nature, has a list of recommended databases for scientific data that can serve as a good reference(<https://www.nature.com/sdata/policies/repositories#nuc>).

In situations where broad data sharing is not appropriate as listed in the RFI, we recommend the ability to be transparent in regard to what specific restriction prohibits or limits data sharing. This will provide clear rationale to the greater research community for why the data are not publicly available.

Lastly, in regard to allowable costs, it may be helpful if there is guidance on the minimum expected length of time data needs to be publicly available so that any associated costs can be properly determined and included in proposals.

Submission ID: 1315

Date: 1/8/2020

Name: Meriel Patrick, on behalf of Research Data Oxford

Name of Organization: University of Oxford

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: A wide range of data types are used

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

Research covers a wide range of domains

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

Surprisingly, there's no explicit mention of a minimum preservation period. (There is a sentence in the next section which says 'NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public' – is this perhaps supposed to say 'for as long'? Though that would still be pretty vague.)

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

- The preamble to the 'Elements of a data sharing and management plan' document states 'NIH does not expect researchers to share all scientific data generated in a study', but doesn't go into any more detail about what they do and don't expect researchers to share.
- Are there any plans to provide applicants with a template for a DMP? At the moment, there just seems to be a list of things that researchers should consider talking about. (Experience suggests people are less likely to skip over key sections if there's a specific space on a form where they're prompted to talk about this.)
- This list of things to be covered is pretty long: it extends to over three pages. But the plan itself should be 'two pages or less': this is a lot to pack into not very much space. (We find this is actually a fairly common frustration for researchers across a wide range of funding bodies: it's very hard to cram all the information requested into the space available.)
- The structure of the guidance for a plan has some oddities: in particular, a lot of stuff gets covered under the 'Data Type' heading, including rationale for preservation/sharing decisions, metadata, and plans for protecting sensitive data. There's a significant risk this will lead to some or all of those elements simply getting overlooked – especially given the above point about lack of space.

- There isn't much in the way of explicit reference to preserving data underpinning publications or, as this is for medical applications, data which might underpin clinical guidelines or 'best practice' (e.g. any American equivalent to the UK's NICE guidelines). Including this might help researchers identify sections of their data which need preserving/sharing as priority.
- Section 3 of the 'Elements of a data sharing and management plan', which covers standards, could be enhanced by adding a reference to FAIRsharing (<https://fairsharing.org>). Pointing to this would help researchers to find the standards relevant to them, and those that are also implemented by the repositories. (FAIRsharing is a community-driven and widely endorsed curated, informative and educational resource on data and metadata standards, inter-related to repositories and data policies. FAIRsharing guides consumers to discover, select and use these resources with confidence, and producers to make their resources more findable, more widely adopted and cited. The FAIRsharing operational team is based at the University of Oxford in the UK; its Advisory Board, adopters and user community is international: <https://fairsharing.org/communities>.)
- Although there is an 'Oversight of Data Management' section at the end of 'Elements of a data sharing and management plan', there isn't any explicit mention of a 'data curator' role or similar. After a project has ended and all the researchers have gone off to do other things, who would be tasked with managing queries or requests for access to the research data?

Other Considerations Relevant to this DRAFT Policy Proposal:

The above comments are a combined response from the Research Data Oxford team, who provide the University of Oxford's cross-departmental research data management guidance service.

Submission ID: 1316

Date: 1/8/2020

Name: Rebecca Osthus

Name of Organization: American Physiological Society

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: Professional Org/Association

Role: Other

Role - Other: These comments are submitted on behalf of the American Physiological Society.

Domain of Research Most Important to You or Your Organization:

Physiology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The American Physiological Society (APS) shares the NIH's goal of making the results of publicly-funded research available to the public. As noted in the APS response to the previous RFI, as a publisher of 15 scientific journals, the society's publications policies (1) already encourage authors to "make data that underlie the conclusions reported in the article freely available via public repositories or available to readers upon request."

The proposed policy will increase administrative burden on funded investigators. NIH should make efforts to harmonize requirements across institutes and centers (ICs) and with other federal agencies. A significant concern of the society is minimizing the amount of administrative burden associated with preparing, submitting, and seeking approval for data management plans, as well as preparing and submitting data into appropriate repositories. Under the provisions of the draft plan, investigators will also be required to comply with any additional requirements imposed by the funding institute or center, as well as requirements from other federal funding agencies or entities that support the research project under consideration. Some of this potential burden could be reduced if NIH would consider developing a data management plan template that could be used for a variety of data modalities.

Investigators already face a significant level of administrative and regulatory burden associated with federal grants and imposing additional requirements will further limit the amount of time they can spend focused on engaging in cutting-edge research. NIH should consider these possible consequences as policies are developed and implemented and to the extent possible, harmonize requirements across ICs, as well as with other federal agencies.

(1) <https://www.physiology.org/author-info.data-repositories>

Section II: Definitions:

The inclusion of negative data in the policy is an important step for the scientific community. APS recommends that NIH offer a definition of what constitutes negative data and provide clarity as to what types of negative data are included under the draft policy. The term "negative data" is sometimes used to describe the results of an experiment that disproved a hypothesis, while other times negative data is used to describe an experiment that failed due to experimental error or a bad reagent.

Section III: Scope:

How will the quality of non-peer reviewed data be assessed?

The broad scope of the draft policy includes all data generated, regardless of whether they have been used to support a publication and, therefore, peer reviewed. In cases where shared data has not been published and undergone peer review, how will the quality of the data be assessed? As data from all NIH-funded projects begins to accumulate, the ability of the scientific community to examine and provide meaningful review will be limited by the volume of data available.

One possible solution to the issue of sharing non-peer reviewed negative data would be developing a journal-like platform that would ensure that submissions are valid by asking researchers in the field to review the data before sharing.

Section IV: Effective Date(s):

New repositories and tools are needed before the policy can be implemented.

Many types of data generated and used in physiology are complex and not easily standardized for deposition into a currently available general data repository. NIH should work with investigator communities to determine what types of repositories, templates and standards are needed to facilitate sharing of data within a particular discipline. These resources should be developed, tested and available before requirements for sharing are fully implemented. These resources should include recommendations for data file formats and meta-tagging.

Section V: Requirements:

NIH should seek to harmonize the requirements of the policy across ICs.

APS encourages NIH to make every effort to harmonize the Data Management and Sharing Policy with the requirements of individual institutes and centers in order to minimize the administrative burden imposed on individual investigators. NIH should ensure that the requirements across the agency include standard recommendations for data file formats and meta-tagging.

Section VI: Data Management and Sharing Plans:

Inclusion of Plans with Just-in-time materials is appropriate. APS appreciates the proposed inclusion of Data Management and Sharing Plans (Plans) with other Just-in-Time materials. This will minimize the administrative burden associated with preparing Plans for projects that are unlikely to be funded.

Application of the policy should be fair across ICs. Programmatic assessment of the Plans by NIH staff should allow researchers the ability to create a plan that meets their own particular needs. At the same time, training and oversight should be implemented to ensure that the policy is applied fairly across the NIH community. Posting of successful or example Plans may help investigators develop their own Plans, and understand what constitutes an acceptable Plan.

What recourse do investigators have if they cannot reach agreement on their proposed Plan with NIH staff? NIH should consider what options will be available to investigators who cannot reach agreement on the elements of a Plan with the NIH staff handling their project.

The ability to update Plans is important and should be made possible with minimal administrative burden. The ability to update Plans is an important component of the draft policy to accommodate the often unpredictable nature of scientific research projects. The process for updating the Plan should be designed to minimize administrative burden and NIH should consider use of an existing platform such as eRA Commons.

Section VII: Compliance and Enforcement:

NIH should address costs for long-term storage and maintenance of data, beyond the end of the award period. If the terms and conditions of the award will be enforced even after the award period has expired, how will investigators or institutions be expected to handle long-term costs for data storage and maintenance? Long-term needs may include ongoing personnel costs to maintain servers and curate content.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

NIH should address costs for long-term storage and maintenance of data, beyond the end of the award period. NIH should consider a plan for handling costs that are incurred for long-term data storage and access. After a project's funding expires, any costs associated with long-term maintenance of the data will be unfunded.

NIH should consider how to make necessary tools available to access and use deposited data. NIH should also consider the costs associated with making tools available for investigators to be able to access and use deposited data sets.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Flexibility is appreciated but there are questions that need to be addressed. APS appreciates the flexibility built into the draft policy, but such a flexible approach inevitably leads to questions about how to develop a Plan that complies with the requirements. Additional guidance should be considered for the following questions:

Basic research can generate a significant volume of data. Any given experiment may generate raw data, reduced data (averaged over a time period), summarized data and corrected data. How should researchers determine which data needs to be managed and shared? How should that data be labeled and described such that others can locate and put it to meaningful use? How can associated data sets be linked?

What metadata are required? To what extent will a methods description need to be included with data for the purposes of replicating experimental results?

Researchers rely on their data remaining confidential so that they can publish their findings, prepare future grant applications, and in some cases, commercialize the results of their research. How long will researchers be allowed to keep data confidential for those purposes?

How will issues related to intellectual property be addressed? NIH should ensure that the new policies will allow researchers and institutions to retain the ability to bring discoveries out of the laboratory and into the marketplace where they will benefit the public. If data sharing requirements prohibit researchers from being able to develop and patent new therapies and technologies, scientific progress will be slowed.

Submission ID: 1317

Date: 1/8/2020

Name: Benjamin Haibe-Kains

Name of Organization: University Health Network

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Bioinformatics, Clinical Genomics, Pharmacogenomics, Machine Learning, Cancer

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Overall, I find that the draft NIH Policy for Data Management and Sharing and Supplemental Guidance provide highly relevant guidelines for much needed data management and sharing in biomedical research. NIH is leading the way regarding these important issues in an era where technologies are progressing rapidly and more data are being generated everyday. It is indeed of the utmost importance to "enable the validation of scientific results, allowing analyses to be strengthened by combining data, facilitating reuse of hard-to- generate data, and accelerating future research." However, I have comments regarding certain aspects of the new policy that may trigger updates for the final version of the guidelines.

Section VI: Data Management and Sharing Plans:

How to enforce and/or monitor compliance to the policy?

An important point that the draft policy does not address is its actual implementation, especially its enforcement and monitoring. Researchers have witnessed the updates of many Journal policies to include data sharing, yet many manuscripts are published without clear description of the data and how to access them. This is due to the fact that, although the guidelines are well described in the Journal website, the Editors lack either the expertise or the resources to actually check whether these guidelines are rigorously followed. I am concerned that the NIH Policy for Data Management and Sharing will suffer from the same limitations without a plan to actually enforce and monitor the plan put forward by the researchers after a grant application has been approved for funding. Given the plethora of data standards, repositories and technologies, it is not trivial to check whether the data have been shared according to the plan and whether the data are sufficiently well annotated to enable full reproducibility of the published results. This requires NIH to allocate financial resources and gather the necessary expertise for this new policy to be effective.

Sharing data may be complex but it is no excuse for preventing access.

Sharing patient health information must be done in accordance with the data governance policy in place where the research is being conducted. This may be complex, involving many departments (legal, ethics, privacy, etc.), following a process that must be clearly defined. The complexity of the process is no excuse for preventing access to the data though and it must be clear to the grantees that the importance of data sharing overrules the perceived complexity of the task. If the process is not clearly mapped as part of the data management and sharing section of the grant application, funding should not be approved as the impact of the research will be seriously undermined. If a data access committee is put into place to "control" access to data, the procedure and maximum turn-around time must be explicitly stated in the grant application. Too often, requests for data access are lost or delayed, or rejected for undisclosed or poorly justified reasons, effectively preventing any (some?) researchers to leverage the dataset of interest.

Timing of data release.

NIH should provide specific guidelines regarding the timing of data release. The data release policy of the NIH-LINCS project is a prime example of a policy enabling the timely release of data that underwent quality checks while being made accessible early to the broad scientific community (<http://www.lincsproject.org/LINCS/data/release-policy>):

- LINCS Centers will release primary and processed data on a quarterly basis as described on the data release page. LINCS data will be released as soon as logistically possible after QA/QC has been completed but not later than 3 months after QA/QC and no more than 6 months after data generation.

- In general, LINCS data are released without any restrictions except correct citation. However, LINCS Centers may petition the NIH LINCS Project Team for permission to post data with a request that large-scale analysis not be performed on a dataset until a primary publication has been submitted; in no case will this "embargo" period be longer than 6 months.

Minimal set of data to be shared.

The statement "Note, NIH does not expect researchers to share all scientific data generated in a study." must be clarified. The draft policy is stating "NIH recognizes that while all scientific data need to be managed, not all data generated in the course of research may be necessary to validate and replicate research findings.". An affirmative sentence stating that "all the data necessary to validate and replicate research findings must be shared according to the policy" is required to avoid any ambiguity and loophole in the policy.

The burden and inefficiencies of "data available upon requests" and data access committees.

Obtaining data via direct requests to the authors or data access committees could be tedious, lengthy, therefore wasting precious mantime and financial resources that could be allocated to research. Assuming that the authors or data access committees are uniquely positioned to decide whether the researcher(s) requesting the data are legit and whether the proposed use of the data is scientifically sound and relevant is fallacious. I advocate for the creation of a "college of researchers" whose membership will rely on (1) strict definition on what a eligible researcher is (affiliation to a recognized research institution and no evidence of misconduct for instance); and (2) the obligation for the eligible researcher to follow mandatory training on the acceptable use of biomedical data for research. Data access should be automatically granted for members of the college of researchers, streamlining access to data and saving on precious resources.

Data cannot be copyrighted.

Data are considered "facts" under U.S. law. They are not copyrightable because they are discovered, not created as original works. This must be made clear in the NIH policy so that copyrights cannot be used as a justification not to share data.

Budget for data management and sharing.

I agree that "plans [...] submitted at "Just-In-Time" and reviewed by NIH program staff" is likely to reduce the applicant burden (only those applicants likely to be funded would submit a plan). However, one should not underestimate (as one currently does) the financial resources required to properly implement a sound plan for data management and sharing plan. Such a plan will likely involve expert curators, ontologists, and data scientists to generate a well-formatted dataset and a documented way of accessing the data and replicating all the analysis results. It is therefore important that costs of data management and sharing are included in the budget early on.

Are only large projects required to adhere to the policy?

Th "first NIH Data Sharing Policy to set the expectation that final research data would be shared from awards requesting \$500,000 or more in direct costs in any single year. " I do not think that restricting the scope to large (expensive) research projects is sending the right message. For practical purposes, it might be advisable to first enforce and monitor the compliance of large grants but the policy should be applicable to all projects, regardless of their budget.

Leading by example.

While some researchers are well versed in the technicalities of data management and sharing, many are not. It would be highly beneficial for the future grant applicants if NIH releases a series of well articulated data management and sharing plans using real-world use cases. Applicants will then be able to adapt such examples for their own applications while learning of the various standards and platforms that can be used to share data. NIH could also build a community resources where applicants voluntarily share their plans for the benefits of the broader community. I would even argue that data management and sharing plans of projects approved and funded by the NIH should be made public to ensure full transparency.

Attachment:

NIH Policy for Data Management and Sharing_BHK_2019.pdf

Comments to be submitted via <https://osp.od.nih.gov/draft-data-sharing-and-management>

Overall, I find that the draft NIH Policy for Data Management and Sharing and Supplemental Guidance provide highly relevant guidelines for much needed data management and sharing in biomedical research. NIH is leading the way regarding these important issues in an era where technologies are progressing rapidly and more data are being generated everyday. It is indeed of the utmost importance to “enable the validation of scientific results, allowing analyses to be strengthened by combining data, facilitating reuse of hard-to- generate data, and accelerating future research.” However, I have comments regarding certain aspects of the new policy that may trigger updates for the final version of the guidelines.

How to enforce and/or monitor compliance to the policy?

An important point that the draft policy does not address is its actual implementation, especially its enforcement and monitoring. Researchers have witnessed the updates of many Journal policies to include data sharing, yet many manuscripts are published without clear description of the data and how to access them. This is due to the fact that, although the guidelines are well described in the Journal website, the Editors lack either the expertise or the resources to actually check whether these guidelines are rigorously followed. I am concerned that the NIH Policy for Data Management and Sharing will suffer from the same limitations without a plan to actually enforce and monitor the plan put forward by the researchers after a grant application has been approved for funding. Given the plethora of data standards, repositories and technologies, it is not trivial to check whether the data have been shared according to the plan and whether the data are sufficiently well annotated to enable full reproducibility of the published results. This requires NIH to allocate financial resources and gather the necessary expertise for this new policy to be effective.

Sharing data may be complex but it is no excuse for preventing access.

Sharing patient health information must be done in accordance with the data governance policy in place where the research is being conducted. This may be complex, involving many departments (legal, ethics, privacy, etc.), following a process that must be clearly defined. The complexity of the process is no excuse for preventing access to the data though and it must be clear to the grantees that the importance of data sharing overrules the perceived complexity of the task. If the process is not clearly mapped as part of the data management and sharing section of the grant application, funding should not be approved as the impact of the research will be seriously undermined. If a data access committee is put into place to “control” access to data, the procedure and maximum turn-around time must be explicitly stated in the grant application. Too often, requests for data access are lost or delayed, or rejected for undisclosed or poorly justified reasons, effectively preventing any (some?) researchers to leverage the dataset of interest.

Timing of data release.

NIH should provide specific guidelines regarding the timing of data release. The data release policy of the NIH-LINCS project is a prime example of a policy enabling the timely release of

Comments to be submitted via <https://osp.od.nih.gov/draft-data-sharing-and-management>

data that underwent quality checks while being made accessible early to the broad scientific community (<http://www.lincsproject.org/LINCS/data/release-policy>):

- LINCS Centers will release primary and processed data on a quarterly basis as described on the data release page. *LINCS data will be released as soon as logistically possible after QA/QC has been completed but not later than 3 months after QA/QC and no more than 6 months after data generation.*
- In general, LINCS data are released without any restrictions except correct citation. However, LINCS Centers may petition the NIH LINCS Project Team for permission to post data with a request that large-scale analysis not be performed on a dataset until a primary publication has been submitted; *in no case will this “embargo” period be longer than 6 months.*

Minimal set of data to be shared.

The statement “Note, NIH does not expect researchers to share all scientific data generated in a study.” must be clarified. The draft policy is stating “NIH recognizes that while all scientific data need to be managed, not all data generated in the course of research may be necessary to validate and replicate research findings.” An affirmative sentence stating that “all the data necessary to validate and replicate research findings must be shared according to the policy” is required to avoid any ambiguity and loophole in the policy.

The burden and inefficiencies of “data available upon requests” and data access committees.

Obtaining data via direct requests to the authors or data access committees could be tedious, lengthy, therefore wasting precious mantime and financial resources that could be allocated to research. Assuming that the authors or data access committees are uniquely positioned to decide whether the researcher(s) requesting the data are legit and whether the proposed use of the data is scientifically sound and relevant is fallacious. I advocate for the creation of a “college of researchers” whose membership will rely on (1) strict definition on what a eligible researcher is (affiliation to a recognized research institution and no evidence of misconduct for instance); and (2) the obligation for the eligible researcher to follow mandatory training on the acceptable use of biomedical data for research. Data access should be automatically granted for members of the college of researchers, streamlining access to data and saving on precious resources.

Data cannot be copyrighted.

Data are considered "facts" under U.S. law. They are not copyrightable because they are discovered, not created as original works. This must be made clear in the NIH policy so that copyrights cannot be used as a justification not to share data.

Budget for data management and sharing.

I agree that “plans [...] submitted at “Just-In-Time” and reviewed by NIH program staff” is likely to reduce the applicant burden (only those applicants likely to be funded would submit a plan). However, one should not underestimate (as one currently does) the financial resources required to properly implement a sound plan for data management and sharing plan. Such a plan will likely involve expert curators, ontologists, and data scientists to generate a well-formatted

Comments to be submitted via <https://osp.od.nih.gov/draft-data-sharing-and-management>

dataset and a documented way of accessing the data and replicating all the analysis results. It is therefore important that costs of data management and sharing are included in the budget early on.

Are only large projects required to adhere to the policy?

The “first NIH Data Sharing Policy to set the expectation that final research data would be shared from awards requesting \$500,000 or more in direct costs in any single year.” I do not think that restricting the scope to large (expensive) research projects is sending the right message. For practical purposes, it might be advisable to first enforce and monitor the compliance of large grants but the policy should be applicable to all projects, regardless of their budget.

Leading by example.

While some researchers are well versed in the technicalities of data management and sharing, many are not. It would be highly beneficial for the future grant applicants if NIH releases a series of well articulated data management and sharing plans using real-world use cases. Applicants will then be able to adapt such examples for their own applications while learning of the various standards and platforms that can be used to share data. NIH could also build a community resources where applicants voluntarily share their plans for the benefits of the broader community. I would even argue that data management and sharing plans of projects approved and funded by the NIH should be made public to ensure full transparency.

Submission ID: 1318

Date: 1/9/2020

Name: Lynda Marie Emel

Name of Organization: Fred Hutchinson Cancer Research Center

Type of Data of Primary Interest: Clinical

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your

Organization: HIV prevention clinical trials

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

As part of a clinical trials network that conducts multiple trials, I understand from our NIH ICO that submission of a Data Management and Sharing Plan would be required to submit a plan for each study, as opposed to a more general plan for the network. It would be helpful to make that clear in the final policy.

Section VI: Data Management and Sharing Plans:

As part of a clinical trials network with a cooperative agreement type funding running multiple studies each year, it would be helpful to know if the expectation is that one plan be submitted for each study, and if yes, at what point the plan should it be submitted.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Regarding the data description and data standards part of the plan, there are already multiple requirements for documentation of data management in the Trial Master File of clinical trials at the network Data Management Center (i.e., Data Management Plan, Data Base Specifications, Edit Check Plan, etc.). It would be helpful to state that it would be acceptable to indicate in the Data Management and Sharing Plan that the information about the type and amount of data collected and who has oversight of those data is in the Trial Master File.

Other Considerations Relevant to this DRAFT Policy Proposal:

It would facilitate truly public data sharing if our ICO (NIAID DAIDS) provided a repository dedicated to prevention clinical trial and behavioral data. As it is now, data sharing from our network occurs through publication websites and our own science portal, which is not very visible.

Submission ID: 1319

Date: 1/9/2020

Name: John Noel

Name of Organization: Sleep Research Society

Type of Data of Primary Interest: Clinical

Type of Organization: Professional Org/Association

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Sleep and Circadian Rhythms

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The Sleep Research Society agrees that data management and sharing practices that are consistent with FAIR data principles constitute best research practice and facilitate timely sharing of data. However, given the rapid pace of technological advancements, proposed prospective data plans need to be flexible enough to accommodate these rapid changes. Overly detailed and/or prescriptive data management/sharing requirements may have the unintended consequence of stifling innovations in data science and the adoption of newly developed approaches. Additional institute-specific requirements, if not carefully coordinated among the institutes and centers, may inadvertently increase administrative barriers for highly interdisciplinary fields such as sleep and circadian biology.

Section II: Definitions:

We recommend that the definition for "Metadata" emphasize that some basic, minimal level of information is needed to make the data usable for secondary analysis.

Section III: Scope:

This appears to cover all NIH research. However, the scope does not specifically mention how joint projects with NIH and other partners will be handled. These partnerships could include other US government agencies, commercial entities such as pharmaceutical companies, or international partners. It is important to consider such collaborations and rules governing their data sharing, in the event that NIH guidance is incongruent with data sharing policies that govern collaborators. Data governance issues have become more complex, and international rules regarding data sharing, such as the recent European Union data sharing regulations on international collaborations, need to be considered.

Section IV: Effective Date(s):

We recommend a phased roll-out, starting first with the largest most data and resource-rich grants. This will allow time for any implementation issues to be identified and addressed, and will help ensure earlier access to the larger data sets.

Section V: Requirements:

The broad goals outlined seem reasonable, especially focusing on the minimum level of metadata to be provided. However, the full life-cycle of a 5-year NIH grant, from initial proposal submission to completion, often exceeds 7 years. Over this timeframe significant technical advances may occur, and may alter the original Data Plan proposal. We strongly encourage the NIH to be flexible in managing updated Data Plans. This consideration is especially important for fields such as sleep and circadian science that are relatively early in the process of developing standardized terminologies and identifying the most appropriate metadata to include for studies. Furthermore, sleep and circadian science investigations use a wide variety of data types ranging from questionnaires to electrophysiological measures. Data such as neuro-cardiorespiratory measures are especially challenging with regard to curation and the variety of platforms available for data collection. Additionally, given the increasing prevalence of wearable devices and ongoing efforts to standardize their validation for sleep and circadian studies, we anticipate more flux in our data needs than may be the case in fields with more stable data types. Therefore, more prescribed requirements may inadvertently place a large administrative burden on both researchers and NIH personnel tasked with evaluating data management and sharing plans that require updates. Finally, the duration and degree to which a researcher would be responsible for providing support for data sharing activities is unclear.

Section VI: Data Management and Sharing Plans:

General components of the data management and sharing plans seem reasonable and necessary. However, we support flexibility in the updating of plans, especially early on in the project life-cycle. A flexible approach recognizes the need for carefully-considered data management and sharing plans throughout the project, but equally recognizes that overly-stringent requirements could impede the adoption of novel methods in data science. We also advocate a focus on basic approaches to best practices in data security.

Section VII: Compliance and Enforcement:

Policy compliance should work to ensure that data sharing is timely, promotes transparency, and encourages data reuse. Enforcement during ongoing data collection should be flexible and allow substantial leeway for researchers to rapidly incorporate advancements in data science.

Ideally, raw, preprocessed data should be made available whenever possible, even if "typically" processed data are also made available. The availability of raw data will facilitate short- and long-term data reuse.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

This document recognizes that data sharing is often associated with additional costs not covered in traditional research budgets. We strongly recommend that enforcement of data sharing be associated with appropriate levels of funding for this specific purpose. Furthermore, we encourage NIH to consider making additional funds available throughout the grant life-cycle, because current technology, availability of new repositories, and costs associated with data repositories are likely to evolve substantially throughout the course of the project. In many cases, it may be difficult to accurately predict the cost of the data sharing plan at project initiation. If additional funds are not available, then researchers should have the option of modifying their Data Management Plan. Without such a provision, increasing data sharing costs during the course of a study with a fixed budget could compromise other parts of the project.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We support outlining the minimum data types that will be collected and shared, and the number of subjects. However, we also recognize that some parts of the data management plan, such as anticipated data size and data management tools, software, and code, may be difficult to project during the early phases of a project. The amount of data collected depends on the cost to store data, and data processing depends on computing power. Tools and algorithms proposed in a data plan at the start of a project may need to be replaced, potentially multiple times, by the time the project is completed. For these reasons, it seems more productive and efficient to focus the Data Plan on minimum levels of data, metadata, and general approaches to data analysis rather than specific tools and algorithms. Detailed descriptions of tools, algorithms, and software should be required at the time of final data release to ensure that data can be reused, and scientific results reproduced. Requiring detailed information in early phases of the research is less likely to be useful, and may increase the administrative burden on researchers and staff evaluating these plans.

We also urge a high level of flexibility with regard to data standards. While many standards are being developed in sleep and circadian science, such standards continue to evolve, even in more mature scientific fields. For example, NIH Clinical Data Elements exist for sleep and circadian science, but many of them require further development to serve as reliable data standards.

As one example, entering "insomnia" into the search function of the NIH Clinical Data Repository brings up 10 different items, none of which contain the key diagnostic element required for a diagnosis of insomnia according to International Classification of Sleep Disorders, Third Edition. This example highlights the need for further development of sleep and circadian related CDEs for use in research.

Submission ID: 1320

Date: 1/9/2020

Name: Robert M Cook-Deegan

Name of Organization: Arizona State University

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Biothiologist/Social Science Researcher

Domain of Research Most Important to You or Your Organization:

Cancer genomics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

It is terrific that NIH intends to strengthen its policies in managing and sharing data for all awards, not just those over \$500k. The effort to retain flexibility while also establishing and clarifying enforcement mechanisms is welcome.

Section VI: Data Management and Sharing Plans:

I see two problems with this section. One concern is that the review of data management plans will be by NIH staff only. The justification for that is consistency. But there is little consistency among NIH units. But the larger problem is that some NIH-funded work is explicitly intended to support "community resource projects" or "common pool resources" of the sort that gave rise to the Bermuda Principles for daily data-sharing of DNA sequence data at high-throughput centers, the Fort Lauderdale and Toronto statements. When that is the purpose of a project, the data management plan should be included in peer review, and moving it to just-in-time means a funding decision will have largely been made in the absence of such review, yet the very purpose of the work is to create and manage data. That makes no sense. Perhaps this makes sense with standing Program Announcements, but when Institutes, Centers and Offices put out announcements about funding opportunities or Requests for Applications for common pool resources, data-sharing should be a criterion of grant evaluation through peer review, and program offices should have the ability to designate when such projects meet the criteria for common pool resources or community research resources.

Section VII: Compliance and Enforcement:

The draft policy fails to state whether NIH will make noncompliance with data management plans public. The enforcement mechanisms listed focus on termination of current awards, considerations for future funding, and the option of affecting funding from the institution. It does not clearly state that NIH can list noncompliant investigators, institutions, or other contractors. It should at least make explicit that this is an option. Funding is powerful, but so is publicity, and publicity has the additional virtue of making the information to other investigators and institutions. NIH needs a data management plan for how it handles data about compliance with data-management plans.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

It is terrific that NIH is making explicit that costs associated with making data available to others is a fundable activity. This should enable more uniformity in decisions about budgets, and give investigators a tool to ensure their data are widely useful.

Other Considerations Relevant to this DRAFT Policy Proposal:

NIH taking the rights and interests of Tribal sovereignty and tribal concerns seriously, and devoting a section of the draft policy to that cluster of issues is commendable and important. This is an area where more specific guidance will almost surely be needed.

This cluster of concerns is not restricted to tribal nations with formal governance, however. In our work, we are also finding indigenous populations and population constituencies that do not have formal recognition or governance structures, and yet share the same well-justified concerns about potential for group harm and the history of data misuse. This is a domain where provenance over data is paramount, and the norms and values of the groups really are in tension with "open science" norms of unfettered use of publicly available data.

The process for developing that guidance needs to be, as NIH is well aware, more ground-up from the affected constituencies than top-down from the NIH bureaucracy and research institutions that form NIH's core constituency. I have no specific process suggestions for managing this problem, except to suggest it will take time and needs to be done with great care for respect, for sovereignty, inclusion, and careful listening; it should not be driven by federal deadlines or central processes. The foremost danger here is that NIH and investigators it supports are perceived (sometimes rightly) as seeing population groups as sources of data to be harvested and shared, but without full awareness of the sensitivities and legitimate rights and interests of the populations that are the source of the data.

Description:

This document is a replica of the comments above, with a preamble explaining some of the sources I am drawing on in my comments.

Submission ID: 1321

Date: 1/9/2020

Name: Tom Cheever

Name of Organization: NIAMS/NIH

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All forms of Data

Type of Organization: Government Agency

Role: Government Official

Domain of Research Most Important to You or Your Organization:

Arthritis, musculoskeletal, and skin diseases

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Goal is highly laudable. Please provide a template or fillable form for applicants to complete as part of their application. A HUGE part of the problem now is researchers/applicants FREQUENTLY misunderstand even the current NIH sharing policy (e.g. - they don't think sharing requirements apply if an application is for less than \$500K for example; they don't know about the model organism and unique resource policy, which to be fair is VERY hard to find. The only way I know how to really find it is in a notice to the Federal register which is not very user friendly at all). Provide text boxes or headings for each field that you want an investigator to address. Maybe something like:

-Model Organisms

-Other Unique Resources

-Data Sharing

-Large Scale Genomic Data

And provide instructions for what applicants should be thinking about or addressing for each of these categories. Make it easy for applicants to follow this incredibly important policy! Part of the problem now is it's very hard for even NIH staff to be able to find resources about the NIH data and sharing policies. How can we expect the external community to be able to follow these policies? Let's make it easy for them!

Section II: Definitions:

Include a template for the plan that applicants can work with

Data management definition isn't terribly intuitive

Section III: Scope:

Maybe make it clear that this policy applies to EVERY application regardless of budget. Many, many, many in the research community have it stuck in their head that sharing only applies to applications over \$500K. I hear this countless times in study sections. Need to help them break out of that.

Section V: Requirements:

Include a template with clear and easy to follow instructions and guidance

How will compliance be monitored? If it's going to be on program officers, you need to give them tools and guidance for how to do this. For example, in my experience 99.9999% of RPPRs I review say "nothing to report" or "N/A" for the section on sharing of resources. I think this needs to change. The RPPR format and instructions should be re-written to be more clear and useful. Free text responses should be minimized so that data can be more easily extracted and analyzed. Grantees should be clearly instructed to comment in the section on sharing about how they're complying with the plan.

Section VI: Data Management and Sharing Plans:

STRONGLY DISAGREE with the submission of the Data Management and Sharing Plan as part of JIT. There are many issues with this:

- 1) Peer reviewers are the people who will ACTUALLY be taking advantage of the data and resources shared. They know best what plans will actually be feasible and useful to the community. Removing their input on this will lead to inferior plans. Sure program officers are intelligent people, but by definition our positions are removed from the actual research fields, and we have less ability and recent experience to know if plans are being followed and if plans would actually work in the real world.
- 2) One of the reasons peer review of data and sharing plans is so ineffective now is that the sharing plans come in 1×10^6 variations since there is no standard template. Reviewers don't really know what is acceptable. And applicants don't know what to comment on. In addition, the instructions for the current policy are so hard to find and follow that again, reviewers really don't know what to comment on.

Again, STRONGLY DISAGREE with the review of data and sharing plans being only programmatic for extramural awards. I understand peer review is asked to do a lot, but this is one item that can only be best considered by people in the actual field. If you want to decrease reviewer burden, have program officers review for consideration of sex as a biologic variable. Reviewers get this wrong 99% of the time anyway in my experience (they don't understand the difference between reporting data by sex and sex differences research). Or maybe authentication of key research resources. Again, I think researchers in the field would be most capable to do this, but if you're not going to have peer reviewers review data and resource sharing plans, why is the authentication of key research resources section any different?

Section VII: Compliance and Enforcement:

How is program supposed to enforce this? Again, see above about how little actual information program currently gets on what is being shared. HIGHLY recommend re-writing RPPR instructions and format/template if possible to make it more clear and understandable to grantees what is expected to be reported. 99% of the time they say "nothing to report" or "Not applicable" in the RPPR section on sharing

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Make this very clear on NIH website, application instructions and forms

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

What about resource sharing? Highly recommend NOT splitting this out. Make it a Data and Resource Sharing Plan. We need to start making things easier for applicants, not continually adding more sections and more policies for them to try and keep track of. I generally completely agree with the rationale and need for most policies, but we need to do a better job thinking about how to help the people impacted follow and implement these policies. Let's use behavioral economic principles - if we want people to do something - let's make it easy for them by including templates, guided instructions, etc.

A form with clear instructions should be created for this. Perhaps even a text box for each bullet point so applicants address each point or enter n/a for non-relevant points. Leave it to peer reviewers to determine if it's truly n/a (peer reviewers will be better equipped to do this than POs as it's challenging for POs with broad portfolios to keep track of field standards when they may cover several of them).

Other Considerations Relevant to this DRAFT Policy Proposal:

The current data and resource sharing policies (including GDS) are some of the most challenging for applicants/grantees to follow and program staff to implement. I could not be more supportive of the ideas in this policy, but I urge you to consider how this would be implemented in the construction of this policy - both from the applicant/grantee side, and from NIH extramural staff side. Give applicants/grantees tools to implement the policy faithfully, and give program staff tools to help them review documents and monitor compliance (the answer to this cannot be more vaguely worded checklist questions - that just doesn't cut it in reality).

Submission ID: 1322

Date: 1/9/2020

Name: James H Jose MD

Name of Organization: Children's Healthcare of Atlanta

Type of Data of Primary Interest: Clinical

Type of Organization: Health Care Delivery Organization

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

Health Outcomes Research

DRAFT NIH Policy for Data Management and Sharing

Section V: Requirements:

1. Data Enclaves can open up less restrictive access to analysis of PHI

Methods should be explored which can allow researchers to analyze PHI in data enclaves under the usual rules applied to de-identified data not subject to HIPAA. This could attract researchers to a more secure method of data sharing and promote standardization.

In 2010 the HHS published an OCR generated "Guidance Regarding Methods for De-identification of Protected Health" in which they commented on the "expert determination method." §164.514(b.) This de-identification method contrasts with the commonly used "safe harbor" method that consists of simply stripping the standard 18 identifiers. Although the expert pathway usually refers to use of statistical methods to render identifiers "ambiguous" the guidance document provides helpful advice on the use of data custody strategies and contracts to secure patient data privacy. Data use rules of "deidentified data" thus apply for data secured in an enclave that includes PHI for analysis as long as the method of access only exposes aggregate results.

2. "De-identification and release strategies"

"De-identification and release," which may be characterized as release of de-identified data sets with no contractual controls on administration and custody, should be curtailed by requiring organizations to develop an exception policy process justifying its use in each case. Increasingly sophisticated de-anonymization algorithms coupled with persistent aggregation of unregulated databases over the decades to come represents a threat that should be of concern, particularly for children. Administrative custody controls for data sets do not simply "add" to the long-term reliability of de-identification schemes – they make them possible.

Section VI: Data Management and Sharing Plans:

Method for Data Enclaves "converting" a PHI data set to de-identified data set rules.

- a. Data is released to a custodian under a Business Associate Agreement or Data Use Agreement. This data may include Personal Health Information. The custodian under this contract may not directly engage in research, publish findings or analyze data other than for database maintenance.
- b. Authorized researchers have remote access to the data. They cannot download the data but are able to perform analysis with tools on the custodian's web site or cloud service. Queries are restricted so that results are masked once results decreased to < 11 rows of data (to prevent re-identification if too few subjects are returned in a query.) While a data use agreement or BAA is required for custodians, it is not for researchers.
- c. The custodian allows access to summary results of many PHI data fields – researchers see only group statistics. Some individual data can be plotted with specific algorithms that limited the allowable queries (such as geolocation approximations.)
- d. Even though PHI is included in the data set being analyzed, no line item PHI is exposed to researchers, and thus there is no "exposure" of PHI. Researchers do not need a Data Use Agreement or IRB to test a hypothesis. Data can be shared among institutions without project specific data use agreements beyond what they deem compatible with organization's mission or are limited by other regulatory constraints.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Successful data enclave strategies should be supported with enhanced overhead support because they can provide wider sharing of data with greater protection of patient privacy.

Submission ID: 1323

Date: 1/9/2020

Name: Tobin Magle

Name of Organization: Research Data Access and Preservation Association

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All datatypes

Type of Organization: Professional Org/Association

Role: Other

Role - Other: Data Specialist, President of RDAP

Domain of Research Most Important to You or Your Organization:

The Research Data Access and Preservation (RDAP) Association offers its feedback on the Draft NIH Policy for Data Management and Sharing along with supplement draft guidance. To put this response in context, RDAP is a community of data professionals who work in a variety of roles and disciplines. Our goal is to support an engaged community of information professionals committed to creating, maintaining, advancing, and teaching best practices for research data management, access, and preservation

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Our organization has no comment on this section.

Section II: Definitions:

Our organization has no comment on this section.

Section III: Scope:

Our organization has no comment on this section.

Section IV: Effective Date(s):

Our organization has no comment on this section.

Section V: Requirements:

Our organization has no comment on this section.

Section VI: Data Management and Sharing Plans:

While many other funding agencies require plans at the grant application stage, the NIH's "just-in-time" approach to data management and sharing has advantages and disadvantages. It streamlines the grant submissions process by not requiring plans for projects that may never receive funding. This approach also prevents peer reviewers, most of whom are not data management experts, from needing to assess plans. However, it also allows funded researchers to avoid peer review of their plans and avoid explicitly considering the implications of data management and sharing when designing their research. This lack of peer review and planning could lead to missing infrastructure or lack of budgetary support for data management and sharing. For instance, what happens if a researcher doesn't budget for long-term data storage and access but later realizes that they are required to do so for an extended period of time? Additionally, the decision to have NIH staff evaluate data management plans as a "just-in-time" portion of a proposal assumes these staff members have the expertise to do so. Ensuring that the NIH staff who are reviewing the plans have adequate data management training and experience is critical for the success of this just in time approach.

Because of the nature of the research that NIH funds, data confidentiality is of the utmost concern. We appreciate the sentiment that the NIH would like research data to be shared as widely as possible, but we strongly suggest that NIH provide guidance about ensuring that personal health information and other sensitive data are properly de-identified before sharing. What standards should researchers use to ensure the data is de-identified to a proper level? Who will help researchers with de-identification? Researchers and repositories will not be able to comply without better guidance and standards. More explicit advice about which data can be publicly shared versus data that should not be shared would be most useful for compliance with this new policy. Additionally, the two-page limit may prove to be inadequate for a five-year grant given the complexities of working with human-subject and health information.

Section VII: Compliance and Enforcement:

RDAP recognizes that another advantage to receiving data management and sharing plans in a "just in time" manner, like Institutional Review Board (IRB) and Institutional Animal Care and Use Committee (IACUC) approvals, is the opportunity to update plans during the annual reporting process. This approach encourages researchers to treat their data management and sharing plan as a "living document" and keeps them engaged with data management throughout the research and granting process. While we strongly support the provision that the plan becomes a term and condition of the grant that affects the continuation of funding and the success of future funding applications, we further recommend wording that the NIH expects researchers to update their plans as their research project changes. In general, we recommend NIH provide more guidance on how this accountability will be assured and encourage NIH to require staff training for those who are responsible for this part of the workflow.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The supplemental guidance on allowable costs for data management and sharing is appreciated. The supplement draft does a good job of addressing two underappreciated areas of data management: using existing data standards, and naming who will be responsible for data management tasks. However, the list of activities in this document are comprehensive and require a substantial amount of human intervention. As data professionals, RDAP members request that the NIH state explicitly that grant funds may be spent on research data management personnel. Such a statement would show that NIH understands the sophisticated level of expertise necessary to properly manage research data and indicate to researchers that it is a worthy expenditure.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

With regard to the supplemental guidance on elements of a data management and sharing plan, RDAP is glad that it aligns with the requirements of other funding agencies. The inclusion of code and software as shareable units emphasizes the NIH's commitment to supporting reproducible research. However, we are concerned by the allowance for researchers to include "to be determined" aspects in their plan. We don't expect researchers to have every detail set in stone from the beginning but we recommend researchers begin their work with a "Plan A" and not "to be determined". We also question why plural language about repositories is used in this section, as it implies that an individual dataset should be submitted to more than one place. This wording goes against current best practices regarding data sharing as there should only be one "version of record" and can cause versioning problems if all copies are not maintained and updated simultaneously.

There are further aspects of the draft guidance that could use additional detail. We recommend providing more detail in the following areas:

If the NIH intends to share the plans, more information regarding what circumstances and how these resources will be shared with the public should be provided. We particularly encourage the sharing of exemplary data management and sharing plans, as this type of resource is commonly requested by researchers and is often exceedingly difficult to procure from either researcher or research administration offices.

More information about what the NIH considers "established repositories" is also needed as the current wording places emphasis on age rather than best practice and quality. RDAP would be happy to participate in advising on specific criteria in this area.

More information on future plans for data catalogs, which will greatly increase the findability, and hence value, of these datasets is also needed. Questions around who will be creating the catalogs, what will be eligible for inclusion, and who can contribute should all be addressed.

Finally, the draft document does not address the challenges faced when sharing large datasets. What plans does NIH have to facilitate "big data" sharing? The amount of data is growing, not shrinking, so clarity on these issues is needed in the near future.

Other Considerations Relevant to this DRAFT Policy Proposal:

As the NIH finalizes its guidance in this area, the RDAP association suggests that the NIH consider whether the final policy sets researchers up for success. Namely, does complying with the NIH policy for data management and sharing, as well as other research compliance policies, lead to better research and more effective return on investment of NIH funding? We recommend that the NIH provide educational training on research data management to researchers and research support staff. While RDAP members can assist researchers in meeting new requirements not all researchers who apply for NIH funding are at institutions with adequate research data support, such as staff and infrastructure. The NIH must plan to provide support to researchers who fall into one of these service gaps when making policies. RDAP strongly encourages the NIH to consider future needs: how will this policy remain relevant and how will it be updated in a timely manner?

Thank you for the opportunity to review the draft; the Research Data Access and Preservation Association looks forward to this policy's inception and its subsequent positive impacts on data management and sharing for the NIH research portfolio.

Attachment:

2019_ResponseToNIH_DataMgmtPolicy.pdf

Description:

PDF of response

The Research Data Access and Preservation (RDAP) Association offers its feedback on the *Draft NIH Policy for Data Management and Sharing* along with supplement draft guidance. To put this response in context, RDAP is a community of data professionals who work in a variety of roles and disciplines. Our goal is to support an engaged community of information professionals committed to creating, maintaining, advancing, and teaching best practices for research data management, access, and preservation. Many of us are actively engaged in assisting researchers with writing and complying with data management plans from NIH and other funding agencies including the National Science Foundation. Collectively we possess a wealth of knowledge on how to support data management and sharing as well as expertise on how to ensure that research data remain accessible.

Section I: Purpose (limit: 8000 characters)

Our organization has no comment on this section.

Section II: Definitions (limit: 8000 characters)

Our organization has no comment on this section.

Section III: Scope (limit: 8000 characters)

Our organization has no comment on this section.

Section IV: Effective Date(s) (limit: 8000 characters)

Our organization has no comment on this section.

Section V: Requirements (limit: 8000 characters)

Our organization has no comment on this section.

Section VI: Data Management and Sharing Plans (limit: 8000 characters)

While many other funding agencies require plans at the grant application stage, the NIH's "just-in-time" approach to data management and sharing has advantages and disadvantages. It streamlines the grant submissions process by not requiring plans for projects that may never receive funding. This approach also prevents peer reviewers, most of whom are not data

management experts, from needing to assess plans. However, it also allows funded researchers to avoid peer review of their plans and avoid explicitly considering the implications of data management and sharing when designing their research. **This lack of peer review and planning could lead to missing infrastructure or lack of budgetary support for data management and sharing.** For instance, what happens if a researcher doesn't budget for long-term data storage and access but later realizes that they are required to do so for an extended period of time? Additionally, the decision to have NIH staff evaluate data management plans as a "just-in-time" portion of a proposal assumes these staff members have the expertise to do so. **Ensuring that the NIH staff who are reviewing the plans have adequate data management training and experience is critical for the success of this just in time approach.**

Because of the nature of the research that NIH funds, data confidentiality is of the utmost concern. We appreciate the sentiment that the NIH would like research data to be shared as widely as possible, but **we strongly suggest that NIH provide guidance about ensuring that personal health information and other sensitive data are properly de-identified before sharing.** What standards should researchers use to ensure the data is de-identified to a proper level? Who will help researchers with de-identification? Researchers and repositories will not be able to comply without better guidance and standards. More explicit advice about which data can be publicly shared versus data that should not be shared would be most useful for compliance with this new policy. Additionally, the two-page limit may prove to be inadequate for a five-year grant given the complexities of working with human-subject and health information.

Section VII: Compliance and Enforcement (limit: 8000 characters)

RDAP recognizes that another advantage to receiving data management and sharing plans in a "just in time" manner, like Institutional Review Board (IRB) and Institutional Animal Care and Use Committee (IACUC) approvals, is the opportunity to update plans during the annual reporting process. This approach encourages researchers to treat their data management and sharing plan as a "living document" and keeps them engaged with data management throughout the research and granting process. While we strongly support the provision that the plan becomes a term and condition of the grant that affects the continuation of funding and the success of future funding applications, we further recommend wording that the NIH expects researchers to update their plans as their research project changes. In general, **we recommend NIH provide more guidance on how this accountability will be assured and encourage NIH to require staff training for those who are responsible for this part of the workflow.**

[Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing](#) (limit: 8000 characters)

The supplemental guidance on allowable costs for data management and sharing is appreciated. The supplement draft does a good job of addressing two underappreciated areas of data management: using existing data standards, and naming who will be responsible for data management tasks. However, the list of activities in this document are comprehensive and require a substantial amount of human intervention. As data professionals, **RDAP members request that the NIH state explicitly that grant funds may be spent on research data management personnel.** Such a statement would show that NIH understands the sophisticated level of expertise necessary to properly manage research data and indicate to researchers that it is a worthy expenditure.

[Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan](#) (limit: 8000 characters)

With regard to the supplemental guidance on elements of a data management and sharing plan, RDAP is glad that it aligns with the requirements of other funding agencies. The inclusion of code and software as shareable units emphasizes the NIH's commitment to supporting reproducible research. However, we are concerned by the allowance for researchers to include "to be determined" aspects in their plan. We don't expect researchers to have every detail set in stone from the beginning but **we recommend researchers begin their work with a "Plan A" and not "to be determined"**. We also question why plural language about repositories is used in this section, as it implies that an individual dataset should be submitted to more than one place. This wording goes against current best practices regarding data sharing as there should only be one "version of record" and can cause versioning problems if all copies are not maintained and updated simultaneously.

There are further aspects of the draft guidance that could use additional detail. **We recommend providing more detail in the following areas:**

- If the NIH intends to share the plans, more information regarding what circumstances and how these resources will be shared with the public should be provided. We **particularly encourage the sharing of exemplary data management and sharing**

plans, as this type of resource is commonly requested by researchers and is often exceedingly difficult to procure from either researcher or research administration offices.

- More information about what the NIH considers “established repositories” is also needed as the current wording places emphasis on age rather than best practice and quality. **RDAP would be happy to participate in advising on specific criteria in this area.**
- More information on future plans for data catalogs, which will greatly increase the findability, and hence value, of these datasets is also needed. **Questions around who will be creating the catalogs, what will be eligible for inclusion, and who can contribute should all be addressed.**
- Finally, the draft document does not address the challenges faced when sharing large datasets. What plans does NIH have to facilitate “big data” sharing? **The amount of data is growing, not shrinking, so clarity on these issues is needed in the near future.**

Other Considerations Relevant to this DRAFT Policy Proposal (limit: 8000 characters)

As the NIH finalizes its guidance in this area, the RDAP association suggests that the NIH consider whether the final policy sets researchers up for success. Namely, does complying with the NIH policy for data management and sharing, as well as other research compliance policies, lead to better research and more effective return on investment of NIH funding? **We recommend that the NIH provide educational training on research data management to researchers and research support staff.** While RDAP members can assist researchers in meeting new requirements not all researchers who apply for NIH funding are at institutions with adequate research data support, such as staff and infrastructure. The NIH must plan to provide support to researchers who fall into one of these service gaps when making policies. **RDAP strongly encourages the NIH to consider future needs: how will this policy remain relevant and how will it be updated in a timely manner?**

Thank you for the opportunity to review the draft; the Research Data Access and Preservation Association looks forward to this policy’s inception and its subsequent positive impacts on data management and sharing for the NIH research portfolio.

Submission ID: 1324

Date: 1/9/2020

Name: Richard Platt, Adrian Hernandez, Lesley Curtis, Kevin Weinfurt (see Purpose for full list)

Name of Organization: Harvard Pilgrim Health Care Institute and Harvard Medical School; Duke University School of Medicine

Type of Data of Primary Interest: Clinical

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Pragmatic Clinical Research Embedded in Health Care Systems

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Statement by Individual Leaders and Investigators Involved in Pragmatic Clinical Trials Embedded in Healthcare Systems

Signatories:

Richard Platt (Harvard Pilgrim Health Care Institute and Harvard Medical School);

Adrian Hernandez (Duke University School of Medicine);

Lesley Curtis (Duke University School of Medicine);

Kevin Weinfurt (Duke University Department of Population Health);

Gregory Simon (Kaiser Permanente Washington Health Research Institute);

Laura Adams (Rhode Island Quality Institute);

Mahnour Ahmed (National Academy of Medicine);

Kristine Martin Anderson (Booz Allen Hamilton);

David Westfall Bates (Brigham and Women's Hospital);

Barbara Bierer (Brigham and Women's Hospital/Harvard Medical School);
Elizabeth Chrischilles (University of Iowa);
Jennifer Christian (Center for Advanced Evidence Generation);
Gail D'Onofrio (Yale School of Medicine);
Deborah Estrin (Cornell University);
Beverly B Green (Kaiser Permanente Washington Health Research Institute);
Sarah Green (HCSRN Executive Director);
Michael Ho (University of Colorado School of Medicine);
Susan Huang (University of California Irvine);
Jeffrey Jarvik (University of Washington);
James Jose (Children's Healthcare of Atlanta);
Richard Kuntz (Medtronic);
Eric B. Larson (Kaiser Permanente Washington Health Research Institute);
Keith Marsolo (Duke University);
Edward Melnick (Yale School of Medicine);
Vincent Mor (Brown University);
Rachel Richesson (Duke University School of Nursing);
Russell Rothman (Vanderbilt University);
Lucy Savitz (HCSRN Governing Board);
Stacy Sterling (Kaiser Permanente Northern California);
Elizabeth Turner (Duke University);
Miguel A. Vazquez (University of Texas Southwestern Medical Center);
Joel S Weissman (Health Brigham and Women's Hospital/Harvard Medical School);
Doug Zatzick (University of Washington Medicine);
Song Zhang (University of Texas Southwestern)

Executive Summary

We offer these comments in response to the Department of Health and Human Services (HHS) request for comments on 84 FR 60398: DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance. We, the above listed respondents, are stakeholders involved in pragmatic clinical trials embedded in healthcare systems. We include investigators and leadership from the National Institutes of Health (NIH) Health Care Systems Research Collaboratory, participants in the National Academy of Medicine (NAM) Clinical Effectiveness Research Innovation Collaborative of the Leadership Consortium for Value and Science-Driven Health Care, and leaders of the Health Care Systems Research Network (HSCRN). We emphasize that we offer these comments as our opinion as individuals and not that of the NIH, NAM, HSCRN.

The topics addressed in these comments are:

- **Support for the goals of this policy:** We applaud this policy and the requirement that all research funded by the NIH provide a data management and sharing plan.
- **Assessing and mitigating re-identification risk:** Embedded pragmatic research occurs in a different context than traditional research. It uses routinely collected data from electronic health records and claims databases, and may involve detailed data on large populations, often including hundreds of thousands of patients. In many cases, these studies are conducted with waiver of informed consent. Before sharing data, investigators may need to do more than simply remove or alter explicit identifiers; they may also need to remove or alter data elements that could enable re-identification through data linkage.
- **Protecting secondary subjects:** Embedded pragmatic trials require different considerations to protect the privacy and confidentiality of those involved, who include not only the participants in the trial, but also friends and family members of participants, providers, healthcare systems, and members of vulnerable classes.
- **Use of data enclaves:** Health systems are often voluntary participants in embedded research with the goal of answering specific questions. They may not be willing to bear the risk for use of sensitive organizational information to address unrelated topics. Their providers are often unable to opt out of embedded research in which their delivery system participates. The potential for disclosure of sensitive information regarding providers or health systems could be substantial, with commensurate harm. Data archives and enclaves are acceptable data sharing mechanisms in routine use that can help mitigate these risks. The Centers for Medicare and Medicaid Services Virtual Research Data Center is an example of a research enclave. It permits investigators to conduct research on approved topics by working with the data in the enclave, and only aggregated data can be removed from the enclave. This has proven to provide a good balance between access and protection of patients' privacy.

- Credit those who share data: As stated Credit Data Generators for Data Re-use we need to develop and mandate the use of a data set ID that will link the use and published analysis from a data set back to the original researchers.

We refer HHS to an opinion paper, Data Sharing and Embedded Research. This document provides a rationale for how data sharing plans for pragmatic research embedded in health care systems are from a different context than traditional randomized trials, and therefore, require different considerations. Our comments below summarize major topics in this opinion document, as well as additional recommendations, that we believe merit attention as the NIH Policy for Data Management and Sharing is finalized. We additionally provide examples of data sharing statements from the NIH Collaboratory.

PURPOSE

We applaud the NIH's policy and commitment to making the results and outputs of the research it funds and conducts available to the public. We enthusiastically support data sharing and agree with the principles of this policy. However, we believe more detail is warranted about the different types of research (i.e., embedded pragmatic research) the associated protections, and acceptable mechanisms for sharing data, such as public and private archives and enclaves.

1 Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. *Nature* 2019;570(7759):30–2. Available from: <http://www.nature.com/articles/d41586-019-01715-4>

2 Simon GE, Coronado G, DeBar LL, et al. Data Sharing and Embedded Research. *Ann Intern Med* 2017;167(9):668. Available from: <http://annals.org/article.aspx?doi=10.7326/M17-0863>

Section VI: Data Management and Sharing Plans:

Assessing and mitigating re-identification risk

The draft policy mentions that de-identification or other protective measures may be necessary to protect privacy and confidentiality: "Researchers proposing to generate scientific data derived from human participants should outline in their Plans how human participants' privacy, rights, and confidentiality will be protected, i.e., through de-identification or other protective measures."

It is important to acknowledge that simple removal of explicit identifiers may not offer adequate protection. Probabilistic re-identification may be possible when research data include data elements also found in other data sources, such as electronic health records, insurance claims, financial records, location records, or genomic data. Prior to sharing research data, investigators may need to remove or alter data elements that could enable re-identification via linkage.

Protecting secondary subjects

The draft policy mentions potential harms to members of Tribal Nations in this statement: For instance, NIH recognizes that sovereign Tribal Nations may have unique data sharing concerns and the Agency has engaged these communities through Tribal Consultation sessions across the U.S. to consider their potential needs in the formation of this DRAFT Policy.

Similar concerns apply to other groups of secondary subjects (i.e., people who were not original subjects of research). People in these groups could be harmed by inference (including invalid inference) from research data. Other types of secondary subjects may include health care providers or organizations delivering care to research participants, family members of research participants, or members of other identifiable vulnerable classes.

Use of data archives and enclaves

Investigators may sometimes access sensitive data via data enclaves (computing environments that allow investigators to execute queries or statistical programs without direct access to or control of individual-level data). Examples include the CMS Virtual Data Research Center and the NIH All of Us Research Hub (Table 1). Investigators cannot share data they neither hold nor control. Instead, investigators may be expected to identify the specific resources used and share the technical tools used to create and analyze research datasets.

Potential structures for data sharing (ranging from least to most restrictive) include the following:

Public archive

Use: Any interested user may download and analyze data without restriction

Examples: Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP)

Private archive

Use: Approved users may download and analyze data, sometimes subject to restrictions, often operationalized in a data use agreement

Examples: The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Repository

Yale University Open Data Access (YODA) Project

Centers for Medicaid and Medicare (CMS) Limited Data Sets

Public enclave

Use: Any interested users may submit queries and receive aggregate results

Examples: The NIH All of Us Research Hub

Centers for Medicaid and Medicare (CMS) Virtual Research Data Center (VRDC)

Private enclave

Use: Approved users may submit queries and receive aggregate results (often subject to review and approval of individual queries)

Examples: U.S. Food and Drug Administration (FDA) Sentinel Distributed Data Set

Data Enclaves can open up less restrictive access to analysis of PHI

Methods should be explored which can allow researchers to analyze PHI in data enclaves under the usual rules applied to de-identified data not subject to HIPAA. This could attract researchers to a more secure method of data sharing and promote standardization.

In 2010 the HHS published an OCR generated "Guidance Regarding Methods for De-identification of Protected Health" in which they commented on the "expert determination method." §164.514(b.) This de-identification method contrasts with the commonly used "safe harbor" method that consists of simply stripping the standard 18 identifiers. Although the expert pathway usually refers to use of statistical methods to render identifiers "ambiguous" the guidance document provides helpful advice on the use of data custody strategies and contracts to secure patient data privacy. Data use rules of "deidentified data" thus apply for data secured in an enclave that includes PHI for analysis as long as the method of access only exposes aggregate results.

"De-identification and release strategies"

"De-identification and release," which may be characterized as release of de-identified data sets with no contractual controls on administration and custody, should be curtailed by requiring organizations to develop an exception policy process justifying its use in each case. Increasingly sophisticated de-anonymization algorithms coupled with persistent aggregation of unregulated databases over the decades to come represents a threat that should be of concern, particularly for children. Administrative custody controls for data sets do not simply "add" to the long-term reliability of de-identification schemes – they make them possible.

Credit those who share data

Citing data sets allows academic researchers to get credit for their work and establishes that data are a valuable scientific output. Pierce et al suggest PIDs, which could be linked to individual ORCID IDs and the DOIs of published manuscripts, allowing the ability to track data and give recognition for the generation of useful data.

Action Needed Regarding Policy on Data Management and Sharing Plans

While we applaud the draft policy, we believe the addition of information regarding different types of research and acceptable mechanisms for data sharing will make it stronger. Therefore, we suggest the following:

- Acknowledge in the Policy that simple removal of explicit identifiers may be insufficient to protect the needs of stakeholders. Prior to sharing research data, investigators may need to remove or alter data elements that could enable re-identification via linkage.
- Examine and acknowledge the unique data sharing concerns of other stakeholders, including secondary subjects, who may include health care providers or organizations delivering care to research participants, family members of research participants, or members of other identifiable vulnerable classes.
- Add information regarding different acceptable data sharing mechanisms to the policy. Indicate that when using data enclaves or other restricted-access data environments, although the data itself cannot be shared, the specific resources and the technical tools used to create and analyze research datasets can be shared.
- Develop mechanisms to link data sets to data generators and track data re-use

Example data sharing statement from the Collaboratory

1. Data sharing statement for the Active Bathing to Eliminate (ABATE) Infection Trial:

"The ABATE Infection trial dataset involves data on over half a million patients. Data sharing requests will be addressed through a supervised data enclave, which will be maintained behind HCA's [Hospital Corporation of America's] firewall on HCA servers for 3 years after the primary publication date. Requests are subject to approval based on planned use of the data, protection of privacy, and scope consistent with the outcomes of the ABATE Infection trial. Only aggregate data (e.g., counts, distributions) will be returned. No individual patient-level results will be released. A processing fee will be assessed to cover this service. Request forms are available."

From: Huang SS, Septimus E, Kleinman K, et al. Chlorhexidine versus routine bathing to prevent multidrug-resistant organisms and all-cause bloodstream infections in general medical and surgical units (ABATE Infection trial): a cluster-randomised trial. *Lancet* 2019;393(10177):1205–15.

Attachment: Collaboratory_Response_to_Draft_NIH_Data_Sharing_Policy_Jan9_2020.pdf

Description:

Statement by Individual Leaders and Investigators Involved in Pragmatic Clinical Trials Embedded in Healthcare Systems



Statement by Individual Leaders and Investigators Involved in Pragmatic Clinical Trials Embedded in Healthcare Systems

Richard Platt (Harvard Pilgrim Health Care Institute and Harvard Medical School);
 Adrian Hernandez (Duke University School of Medicine);
 Lesley Curtis (Duke University School of Medicine);
 Kevin Weinfurt (Duke University Department of Population Health);
 Gregory Simon (Kaiser Permanente Washington Health Research Institute);
 Laura Adams (Rhode Island Quality Institute);
 Mahnoor Ahmed (National Academy of Medicine);
 Kristine Martin Anderson (Booz Allen Hamilton);
 David Westfall Bates (Brigham and Women's Hospital);
 Barbara Bierer (Brigham and Women's Hospital/Harvard Medical School);
 Elizabeth Chrischilles (University of Iowa);
 Jennifer Christian (Center for Advanced Evidence Generation);
 Gail D'Onofrio (Yale School of Medicine);
 Deborah Estrin (Cornell University);
 Beverly B Green (Kaiser Permanente Washington Health Research Institute);
 Sarah Green (HCSRN Executive Director);
 Michael Ho (University of Colorado School of Medicine);
 Susan Huang (University of California Irvine);
 Jeffrey Jarvik (University of Washington);
 James Jose (Children's Healthcare of Atlanta);
 Richard Kuntz (Medtronic);
 Eric B. Larson (Kaiser Permanente Washington Health Research Institute);
 Keith Marsolo (Duke University);
 Edward Melnick (Yale School of Medicine);
 Vincent Mor (Brown University);
 Rachel Richesson (Duke University School of Nursing);
 Russell Rothman (Vanderbilt University);
 Lucy Savitz (HCSRN Governing Board);
 Stacy Sterling (Kaiser Permanente Northern California);
 Elizabeth Turner (Duke University);
 Miguel A. Vazquez (University of Texas Southwestern Medical Center);
 Joel S Weissman (Health Brigham and Women's Hospital/Harvard Medical School);
 Doug Zatzick (University of Washington Medicine);
 Song Zhang (University of Texas Southwestern)

Executive Summary

We offer these comments in response to the Department of Health and Human Services (HHS) request for comments on 84 FR 60398: DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance. We, the above listed respondents, are stakeholders involved in pragmatic clinical trials embedded in healthcare systems. We include investigators and leadership from the National Institutes of Health (NIH) Health Care Systems Research Collaboratory, participants in the National Academy of Medicine (NAM) Clinical Effectiveness Research Innovation Collaborative of the Leadership Consortium for Value and Science-Driven Health Care, and leaders of the Health Care Systems Research Network (HSCRN). We emphasize that we offer these comments as our opinion as individuals and not that of the NIH, NAM, HSCRN.

The topics addressed in these comments are:

- **Support for the goals of this policy:** We applaud this policy and the requirement that all research funded by the NIH provide a data management and sharing plan.
- **Assessing and mitigating re-identification risk:** Embedded pragmatic research occurs in a different context than traditional research. It uses routinely collected data from electronic health records and claims databases, and may involve detailed data on large populations, often including hundreds of thousands of patients. In many cases, these studies are conducted with waiver of informed consent. Before sharing data, investigators may need to do more than simply remove or alter explicit identifiers; they may also need to remove or alter data elements that could enable re-identification through data linkage.
- **Protecting secondary subjects:** Embedded pragmatic trials require different considerations to protect the privacy and confidentiality of those involved, who include not only the participants in the trial, but also friends and family members of participants, providers, healthcare systems, and members of vulnerable classes.
- **Use of data enclaves:** Health systems are often voluntary participants in embedded research with the goal of answering specific questions. They may not be willing to bear the risk for use of sensitive organizational information to address unrelated topics. Their providers are often unable to opt out of embedded research in which their delivery system participates. The potential for disclosure of sensitive information regarding providers or health systems could be substantial, with commensurate harm. Data archives and enclaves are acceptable data sharing mechanisms in routine use that can help mitigate these risks. The Centers for Medicare and Medicaid Services Virtual Research Data Center is an example of a research enclave. It permits investigators to conduct research on approved topics by working with the data in the enclave, and only aggregated data can be removed from the enclave. This has proven to provide a good balance between access and protection of patients' privacy.

- **Credit those who share data:** As stated *Credit Data Generators for Data Re-use* we need to develop and mandate the use of a data set ID that will link the use and published analysis from a data set back to the original researchers.¹

We refer HHS to an opinion paper, **Data Sharing and Embedded Research.**² This document provides a rationale for how data sharing plans for pragmatic research embedded in health care systems are from a different context than traditional randomized trials, and therefore, require different considerations. Our comments below summarize major topics in this opinion document, as well as additional recommendations, that we believe merit attention as the NIH Policy for Data Management and Sharing is finalized. We additionally provide examples of data sharing statements from the NIH Collaboratory.

¹ Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. *Nature* 2019;570(7759):30–2. Available from: <http://www.nature.com/articles/d41586-019-01715-4>

²Simon GE, Coronado G, DeBar LL, et al. Data Sharing and Embedded Research. *Ann Intern Med* 2017;167(9):668. Available from: <http://annals.org/article.aspx?doi=10.7326/M17-0863>

² Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. *Nature* 2019;570(7759):30–2. Available from: <http://www.nature.com/articles/d41586-019-01715-4>

²Simon GE, Coronado G, DeBar LL, et al. Data Sharing and Embedded Research. *Ann Intern Med* 2017;167(9):668. Available from: <http://annals.org/article.aspx?doi=10.7326/M17-0863>

PURPOSE

We applaud the NIH's policy and commitment to making the results and outputs of the research it funds and conducts available to the public. We enthusiastically support data sharing and agree with the principles of this policy. However, we believe more detail is warranted about the different types of research (i.e., embedded pragmatic research) the associated protections, and acceptable mechanisms for sharing data, such as public and private archives and enclaves.

DATA MANAGEMENT AND SHARING PLANS

Assessing and mitigating re-identification risk

The draft policy mentions that de-identification or other protective measures may be necessary to protect privacy and confidentiality: *“Researchers proposing to generate scientific data derived from human participants should outline in their Plans how human participants' privacy, rights, and confidentiality will be protected, i.e., through de-identification or other protective measures.”*

It is important to acknowledge that simple removal of explicit identifiers may not offer adequate protection. Probabilistic re-identification may be possible when research data include data elements also found in other data sources, such as electronic health records, insurance claims, financial records, location records, or genomic data. Prior to sharing research data, investigators may need to remove or alter data elements that could enable re-identification via linkage.

Protecting secondary subjects

The draft policy mentions potential harms to members of Tribal Nations in this statement: *For instance, NIH recognizes that sovereign Tribal Nations may have unique data sharing concerns and the Agency has engaged these communities through Tribal Consultation sessions across the U.S. to consider their potential needs in the formation of this DRAFT Policy.*

Similar concerns apply to other groups of secondary subjects (i.e., people who were not original subjects of research). People in these groups could be harmed by inference (including invalid inference) from research data. Other types of secondary subjects may include health care providers or organizations delivering care to research participants, family members of research participants, or members of other identifiable vulnerable classes.

Use of data archives and enclaves

Investigators may sometimes access sensitive data via data enclaves (computing environments that allow investigators to execute queries or statistical programs without direct access to or control of individual-level data). Examples include the CMS Virtual Data Research Center and the NIH All of Us Research Hub (Table 1). Investigators cannot share data they neither hold nor control. Instead, investigators may be expected

to identify the specific resources used and share the technical tools used to create and analyze research datasets.

Potential structures for data sharing (ranging from least to most restrictive) include the following:

Table 1. Data Sharing Mechanisms and Examples

Mechanism	Use	Examples
Public archive	Any interested user may download and analyze data without restriction	Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP)
Private archive	Approved users may download and analyze data, sometimes subject to restrictions, often operationalized in a data use agreement	The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Repository Yale University Open Data Access (YODA) Project Centers for Medicaid and Medicare (CMS) Limited Data Sets
Public enclave	Any interested users may submit queries and receive aggregate results	The NIH All of Us Research Hub Centers for Medicaid and Medicare (CMS) Virtual Research Data Center (VRDC)
Private enclave	Approved users may submit queries and receive aggregate results (often subject to review and approval of individual queries)	U.S. Food and Drug Administration (FDA) Sentinel Distributed Data Set

Data Enclaves can open up less restrictive access to analysis of PHI

Methods should be explored which can allow researchers to analyze PHI in data enclaves under the usual rules applied to de-identified data not subject to HIPAA. This could attract researchers to a more secure method of data sharing and promote standardization.

In 2010 the HHS published an OCR generated “Guidance Regarding Methods for De-identification of Protected Health” in which they commented on the “expert determination method.” §164.514(b.) This de-identification method contrasts with the commonly used “safe harbor” method that consists of simply stripping the standard 18 identifiers. Although the expert pathway usually refers to use of statistical methods to render identifiers “ambiguous” the guidance document provides helpful advice on the

use of data custody strategies and contracts to secure patient data privacy. Data use rules of “deidentified data” thus apply for data secured in an enclave that includes PHI for analysis as long as the method of access only exposes aggregate results.

“De-identification and release strategies”

“De-identification and release,” which may be characterized as release of de-identified data sets with no contractual controls on administration and custody, should be curtailed by requiring organizations to develop an exception policy process justifying its use in each case. Increasingly sophisticated de-anonymization algorithms coupled with persistent aggregation of unregulated databases over the decades to come represents a threat that should be of concern, particularly for children. Administrative custody controls for data sets do not simply “add” to the long-term reliability of de-identification schemes – they make them possible.

Credit those who share data

Citing data sets allows academic researchers to get credit for their work and establishes that data are a valuable scientific output. Pierce et al suggest PIDs, which could be linked to individual ORCID IDs and the DOIs of published manuscripts, allowing the ability to track data and give recognition for the generation of useful data.

Action Needed Regarding Policy on Data Management and Sharing Plans

While we applaud the draft policy, we believe the addition of information regarding different types of research and acceptable mechanisms for data sharing will make it stronger. Therefore, we suggest the following:

- Acknowledge in the Policy that simple removal of explicit identifiers may be insufficient to protect the needs of stakeholders. Prior to sharing research data, investigators may need to remove or alter data elements that could enable re-identification via linkage.
- Examine and acknowledge the unique data sharing concerns of other stakeholders, including secondary subjects, who may include health care providers or organizations delivering care to research participants, family members of research participants, or members of other identifiable vulnerable classes.
- Add information regarding different acceptable data sharing mechanisms to the policy. Indicate that when using data enclaves or other restricted-access data environments, although the data itself cannot be shared, the specific resources and the technical tools used to create and analyze research datasets can be shared.
- Develop mechanisms to link data sets to data generators and track data re-use

EXAMPLES OF DATA SHARING STATEMENTS FROM THE COLLABORATORY

1. Data sharing statement for the Active Bathing to Eliminate (ABATE) Infection Trial:

“The ABATE Infection trial dataset involves data on over half a million patients. Data sharing requests will be addressed through a supervised data enclave, which will be maintained behind HCA’s [Hospital Corporation of America’s] firewall on HCA servers for 3 years after the primary publication date. Requests are subject to approval based on planned use of the data, protection of privacy, and scope consistent with the outcomes of the ABATE Infection trial. Only aggregate data (e.g., counts, distributions) will be returned. No individual patient-level results will be released. A processing fee will be assessed to cover this service. Request forms are available.”

From: Huang SS, Septimus E, Kleinman K, et al. Chlorhexidine versus routine bathing to prevent multidrug-resistant organisms and all-cause bloodstream infections in general medical and surgical units (ABATE Infection trial): a cluster-randomised trial. *Lancet* 2019;393(10177):1205–15.

2. Data sharing statement for the NIH Collaboratory Distributed Research Network paper on statin use in the elderly:

“Data Availability Statement: The data we used belonged to, and remained in the possession of third parties, i.e., the private health plan that created and maintain the data. The lead author did not have special access privileges. Per our agreement with the health plans, a health plan based investigator became an author of this report after meeting ICMJE criteria. Others would be able to solicit participation by these organizations in the same manner. Others would be able to conduct analyses on these data by submitting the programs available as a Supporting Information file to the third party organizations within two years of this publication date. These third party organizations voluntarily participated in this study and would need to participate voluntarily in any subsequent study. They would participate in related follow-up studies proposed by other investigators, subject to the same bandwidth, resource, and collaboration requirements. Interested persons can contract the NIH Collaboratory Distributed Research Network Leadership by emailing...”

From: Panozzo CA, Curtis LH, Marshall J, et al. Incidence of statin use in older adults with and without cardiovascular disease and diabetes mellitus, January 2008-March 2018. *PLoS ONE* 2019;14(12):e0223515. Available from: <https://dx.plos.org/10.1371/journal.pone.0223515>.

Submission ID: 1325

Date: 1/9/2020

Name: Brett Harnett

Name of Organization: University of Cincinnati

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Organization: University

Role: Other

Role - Other: Departmental Director and Faculty

Domain of Research Most Important to You or Your Organization:

All clinical and translational research (from an informatics perspective)

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

This is an exceptional outline, very well-conceived and implemented. I suspect many investigators do not give much thought to data mgt/sharing and instead focus on what has been historically successful for them: solid science wins grants. I support wholeheartedly the concepts within – being in clinical research informatics for over 20 years, I am happy to post some brief comments.

Section V: Requirements:

Actions speak louder than grant submission words. I can see templates being developed by institutions that spell out boilerplate text to meet said requirements. I suggest some mechanisms below to actually operationalize this practice.

Consider underscoring the requirements so institutions know this is a serious issue. When the Privacy Rule came out in 2003, it was required as well but it was not until HITECH in 2009 that HIPAA had teeth. Take that as an example as to what is prioritized by PIs and institutions.

Section VI: Data Management and Sharing Plans:

Investigators have people who manage data – but are not necessarily experts in areas and concepts such as: data processing, cleaning and analyzing using industry best practices and progressive informatics tools. Data Documentation Initiative (DDI) metadata specification using XML to create structured documentation compliant with the international standard for the content and exchange of documentation. Structured, XML-based metadata that are ideal for documenting research data because the structure provides machine-actionability and the potential for metadata reuse.

Standardized repositories such as the NIH Common Data Elements (CDE) Repository, and/or datasets to be registered with DataMed/bioCADDIE and made available on the NCBO BioPortal. Utilizing CDEs as much as possible to harmonize the cohorts. Genomic data (data science/big data) will grow exponentially and make this policy even more critical and challenging to implement.

Section VII: Compliance and Enforcement:

Require 'letter of support' from data mgt team/informatics – those who are appointed institutional Honest Brokers of the data. Require specific data architecture descriptions such as the Advanced Research Computing (ARC) initiative HPC Cluster configuration, EMC ISILON Storage or commercial cloud hosting and even colocation configurations. Also, the type of database schema, platform, encryption and who administers the repository (permissions, roles, rights, etc.)

Suggest, grant submissions with plans that do not meet requirements will not be funded despite any score.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Conversations about what can be indexed in a budget justification vs what should be covered by indirects I suspect is a common conversation at institutions (I'm at my third). The point about "...potential categories of allowable NIH costs associated with data management and sharing..." is a huge YES. This is needed.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

This section is comprehensive and where most institutions may struggle to comply. Describing the data types, formats, standards, volume, veracity, etc. will require significant data expertise parallel to the research domain. Encourage use of standards for data capture such as OMOP, PCORNET, etc. Models at outset that reduce friction and cost to enable FAIR principles.

Need a template that requires certain elements. Also consider requiring an institutional meta data repository that all PIs have to subscribe to that ensures policy compliance – once set up, super easy to maintain and relatively inexpensively.

Point 6 cannot be underestimated.

Other Considerations Relevant to this DRAFT Policy Proposal:

As more data is being held in compliant cloud architectures, it makes sense to create leverage common data models such as OMOP that has scale and flexibility at its core, to enable federated searches across data domains – this not ferrying, copying and even worse, downloading data for analysis and hypothesis generation. Cloud systems are designed to enable this type of query logic while not moving tera/exabytes of data that is simply inappropriate.

Finally, institutions will need time to ingest and understand this change in policy.

Submission ID: 1326

Date: 1/9/2020

Name: Mary Ellen K. Davis, Executive Director

Name of Organization: Association of College & Research Libraries (ACRL)

Type of Data of Primary Interest: Other

Type of Organization: Professional Org/Association

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

The Association of College & Research Libraries (ACRL) is the higher education association for academic libraries and library workers. Representing more than 10,000 individuals and libraries, ACRL (a division of the American Library Association) develops programs, products, and services to help those working in academic and research libraries learn, innovate, and lead within the academic community. Founded in 1940, ACRL is committed to advancing learning, transforming scholarship, and ...

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

(con't) ...creating diverse and inclusive communities. We enhance the ability of academic library and information professionals to serve the information needs of students and researchers. For example, through a one-day workshop, ACRL presenters travel to campuses across the U.S. and train librarians in the nuances of disciplinary requirements for research data management in order to educate their faculty and students about data best practices. As reflected in our previous support for governmental policies and legislation that facilitate open access and open education—including the NIH Open Access Policy, the Office of Science and Technology Policy mandate, and the Fair Access to Science & Technology Research Act and Federal Research Public Access Act bills—ACRL is fundamentally committed to the open exchange of information to empower individuals and facilitate scientific discovery. On December 5, 2018, ACRL provided comments in response to the NIH Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research. We appreciate the revisions NIH made, which address concerns we raised at that time; however, we have the following recommendations for NIH to further improve the policy before its implementation.

SECTION I: PURPOSE

We recommend providing a citation to the specific definition of FAIR data principles mentioned at the end of the first paragraph of this section. The following article is cited in the NIH Strategic Plan for Data Science (<https://grants.nih.gov/grants/nih-public-access-plan.pdf>) and provides more details about what Findable, Accessible, Interoperable, and Reusable mean in practice:

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 160018. doi:10.1038/sdata.2016.18.

Section II: Definitions:

Data Management and Sharing Plan: The concept of accessibility has been removed from the definition provided in Section II. We question why it has been removed in this section, as this is in conflict with Section I, which encourages following FAIR data principles. By removing "accessible," NIH opens the possibility of researchers sharing insufficient information, omitting information that is required for data to be fully accessible.

Data Management: We appreciate the definition of Data Management that has been added to this list of definitions. However, this is another point at which FAIR data principles can be included.

Scientific Data: We also appreciate the addition of the clause "regardless of whether the data are used to support scholarly publications" to this definition. Not all experiments result in a formal publication, but data generated may have significant value to other researchers. However, we recommend NIH clarify the definition of Scientific Data to indicate that although the list of examples of what are not considered Scientific Data are excluded from what needs to be shared, they are types of data that should be carefully managed.

Additionally, the definition of Scientific Data includes the statement that "NIH expects that reasonable efforts will be made to digitize all scientific data." "Reasonable efforts" is vague and should be more clearly defined. What criteria will NIH set for what scientific data should be digitized? Additionally, digitization can be expensive. Will costs of digitization be an allowable cost? It is not listed in the Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing.

We recommend that a definition is provided for the term "preservation," which is used liberally throughout this policy but is subject to a multitude of differing definitions.

Section III: Scope:

We recommend that NIH explicitly state to which types of grants the policy will apply. Training grants and career development grants may generate scientific data—are they considered "other funding agreements" and thus subject to this policy?

Section IV: Effective Date(s):

We recommend that this policy be made effective to all calls for proposals released after the publication of this memo, allowing applications in progress to proceed with their current project designs.

Section V: Requirements:

The timeline for requiring the submission of the Data Management and Sharing Plan should be clarified, particularly in light of Sections IV and VI. Is the Plan to be submitted with the grant application or only upon request (e.g., as Just-In-Time material)?

Section VI: Data Management and Sharing Plans:

Throughout this section, we recommend the removal of the word "consider" to require that the Plan include all of the elements described.

The importance of a plan for managing and sharing data cannot be overstated. We believe that researchers should be required to think through the data management and sharing issues related to their work for all NIH-funded research when they are first planning their research and drafting proposals. Designating "Just-In-Time" (https://grants.nih.gov/grants/policy/nihgps/html5/section_2/2.5.1_just-in-time_procedures.htm) as the point in the process at which Plans are submitted to NIH lessens the importance of having such a plan. Data management practices and metadata standards are associated with specific methods, disciplines, and epistemologies. Therefore, effective data management planning begins during project design and is tied to research methodology. We recommend NIH consider clarifying by explicitly stating that a Plan is required for all grant proposals, but that additional information can be included as part of Just-In-Time requests. The Policy should have clear language indicating that the Plan is required as part of submission and will be evaluated as part of the quality of the proposal. Also, the policy should address how much of the plan can remain "to be determined" in the Just-In-Time submission.

One of the most common requests received by librarians who assist researchers with their data management plans is for examples of successful plans. We encourage the NIH, for the benefit of the community, to revise this statement to read: "NIH will make Plans associated with successful grant submissions publicly available."

The statement "Researchers proposing to generate scientific data derived from human participants should outline in their Plans how human participants' privacy, rights, and confidentiality will be protected, i.e., through de-identification or protective measures" should cite best practice documents for de-identification and other types of protective measures; for example, NIST's De-Identification of Personal Information (<https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>). This section should also be revised to lighten the focus on de-identification as the only named measure. We suggest including access security (or similar) in the list of examples provided.

We appreciate that NIH encourages researchers to use established repositories. However, it would be useful to define what NIH means by "established repositories." Would it require that repositories follow the ISO standard for trustworthy digital repositories (<https://public.ccsds.org/pubs/652x0m1.pdf>) and/or have CoreTrustSeal (<https://www.coretrustseal.org/>) certification? Many research institutions have institutional repositories, some of which meet the ISO standard referenced or have acquired CoreTrustSeal certification, which could potentially be used to provide long-term access and storage.

We appreciate that NIH will allow researchers to update plans "during regular reporting intervals if changes are necessary or at the request of the NIH ICO to reflect changes in the previously documented approach to data management and data sharing throughout the research project, as appropriate."

We recommend NIH more thoroughly explain what is meant by the statement that "Plans will undergo a programmatic assessment" for extramural awards. Include explanations of the evaluation process and criteria.

Section VII: Compliance and Enforcement:

We appreciate that the Data Management and Sharing Plan review and update process will be integrated into RPPRs. Plans should be a living document that can be adjusted to address the unexpected turns that research can take. This section states that these reviews will happen during regular reporting intervals, with the implication that the same body reviewing RPPRs is reviewing these. NIH should clarify who will be reviewing/assessing plans.

We appreciate that NIH has included compliance language. We recommend making a stronger statement by replacing "may" with "will" in the statement that not following the Plan "may affect future funding decisions." Strengthening the compliance language associated with the policy requiring the public sharing of publications appears to be what significantly improved the compliance rate for that policy.

Similarly, the statement that "After the end of the funding period, no-compliance with the NIH ICO-approved Plan may be taken into account" should be strengthened by changing "may" to "will" or should include a more definite statement of what "taken into account" means. (E.g., would reports on past compliance levels be considered as part of any future funding request?)

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Per the definition of Scientific Data, will digitization costs be allowed in "Curating data and developing supporting documentation"? If so, this should be explicitly stated.

Item 2, Preserving and sharing data through established repositories, allows fees and charges for repositories. However, some repositories require a recurring fee. How will such fees be addressed? Would applicants be granted no-cost extensions (provided the fee is written into the original grant and a specific retention period is defined) to cover these fees beyond the grant period? We recommend NIH develop explicit rules and procedures for how this will work. An alternative to basing repository selection on fee structure may be the development of a funding and budget model that allows for the maintenance and curation of grant-developed resources.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

As data management and sharing is a requirement for responsible research, the word "consider" should be dropped from throughout this section. The Supplementary Information preceding the draft Plan (page 6) states, "...supplemental DRAFT guidance documents intended to help researchers prospectively integrate Data Management and Sharing Plans into *routine research practices*" (emphasis added). Again, public access policies for publications succeeded when compliance was enforced.

Throughout this section, remove "consider." Again, data management/sharing is a requirement for good research and as a result of federal funding, NIH research is a public good and thus must be properly managed and shared.

We appreciate that elements of a Plan should provide, "a rationale for decisions about which scientific data are to be preserved." Principal Investigators should thoroughly consider and be able to articulate why they do what their plan says.

Section 1, last bullet. The guidance should explicitly require that human participants are given the option of being made aware of how their data will be shared. This is a core ethical principle.

Section 2. Related Tools, Software, and or/Code. NIH should require sharing of code necessary to reproduce results based on shared data.

Section 5, second bullet. We recommend clarifying the phrasing of "Whether the applicant anticipates entering into any agreements that could limit the ability to broadly share scientific data and describe those agreements." It is unclear what this means or what kind of agreements NIH would allow.

Other Considerations Relevant to this DRAFT Policy Proposal:

Notes on SUPPLEMENTARY INFORMATION (pages 2-7) provided before the draft Policy.

While page 3 indicates that "Plans will be included as part of the technical evaluation performed by NIH staff," further guidance on evaluation criteria for data management plans will be needed.

Many university libraries provide data management services, such as planning and/or preservation. Researchers that employ such institutional resources should demonstrate that they have made contact with the relevant program managers, for example, through a letter of support.

Page 6 states that "NIH recognizes that the deliberate flexibility of its DRAFT Policy may require additional implementation guidance." We agree that policies require a certain measure of flexibility, especially in a research area as diverse as health. However, flexibility should not be synonymous with weakness. We recommend the entire Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan be strengthened by removing the word "consider," thus requiring applicants to provide information for each of the elements described.

Submission ID: 1327

Date: 1/9/2020

Name: Erik Deumens

Name of Organization: University of Florida

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: General, all data

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

data management

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

University of Florida supports the comments submitted by CASC Coalition for Academic Scientific Computing.

These comments serve only add details to the "Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan"

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

1. Data Type: The supplemental draft guidance is clear in articulating key requirements for a NIH data management and Sharing Plan. The addition of (1) categories of metadata, (2) definition of supplemental data resulting in generation of scientific data, (3) definition of FAIR data, and (4) following suggestions could potentially further enhance the supplemental draft for researchers.

- There are categories in the identification of metadata, other relevant data, and any associated data documentation such as descriptive, preservation, technical, structural, and administrative. The inclusion of these categories after "Identifying metadata (e.g., descriptive, preservation, technical, structural, and administrative)..." in the third bullet point under Data Type in Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (Plan) may clarify the types of metadata for researchers.
- The inclusion of a definition for supplemental data outputs (i.e. raw data, temporary data, processed, and analyzed data) may clarify types of data for researchers.

i. Supplemental Data: Supplemental data (e.g. raw data, temporary data, processed data, and analyzed data) result in the generation of scientific data. Supplemental data may be represented as non-standard public records, electronic theses and dissertations (ETDs), open source intelligence, unpublished research data (i.e. raw data, temporary data, processed data, analyzed data), web interfaces, laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.

ii. FAIR (Findable, Accessible, Interoperable, and Reusable) Data: FAIR data is data that support the FAIR guiding data principles for scientific data management and stewardship. The 15 data principles covering four categories refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. Data is considered FAIR if reasonable efforts have been made to make the data findable, accessible, interoperable, and reusable. Recommend inclusion of FAIR data examples, scenarios, and use cases to educate researchers. See: <https://www.go-fair.org/fair-principles/>

- Clarifying sharing and deposit of raw data, processed data, and analyzed data factoring format, size, and copyright may clarify deposit of data types. Added guidance may include recommendations for reproducibility best practices such as db reproducible (See: <http://db-reproducibility.seas.harvard.edu/>). The goal is to enable reproducibility of the raw data and relevant plots that the authors used to draw their conclusions. Authors should provide a complete set of scripts to (1) install the system, (2) produce the data, (3) run experiments, and (4) produce the resulting graphs along with a detailed Readme file that describes the process systematically for reproducibility by a reviewer or other researchers.

2. Related Tools/Software and/or Code: Guidance may suggest created code and scripts following an established best practice (i.e. community standard, best practices, recommendation) be shared with data to make the data replicable. See: <http://db-reproducibility.seas.harvard.edu/>

- Recommend the use of API, open source software, collaborative tools, and version control science scenarios and use cases to make data FAIR. See: <https://open.fda.gov/>

- Provide examples of communities of practice, computational tools & services, data & information assets, management, policy & standards, and science inputs with links to successful examples in each category.

3. Standards: Guidance on identifying, understanding, and using appropriate metadata standards (i.e. discipline-specific, general), ontologies, schemas, and semantics for scientific data to be created, aggregated, represented, disseminated, and preserved during the life of the usefulness of the data in the form of examples may be useful. See: <https://www.go-fair.org/fair-principles/>

4. Data Preservation, Access, and Associated Timelines: Added guidelines on recommendations in the selection of acceptable data repositories for data deposit may be useful to researchers.

- Provide list of NIH approved data repositories (i.e. institutional, general, and discipline-specific), including free, fee. If fee, then guidance on proper budget for data deposits.
- i. NIH Data Repositories and Trusted Partners: <http://tinyurl.com/yhkfyxn>
- Recommend a Data Services & Developer Tools guidelines resources (e.g. <https://www.osti.gov/data-services-developer-tools>) to understand the creation of OAI-PMH metadata records (e.g. <https://www.osti.gov/oairecords>) that allows linking of metadata record with data in an external data repository if unable to deposit data in a repository.
- i. Include guidance for a metadata repository linking option to data repository to provide make data FAIR given capacity, infrastructure, and resources.
- Provide clarification on types and trust of repositories. An institutional repository is not a data repository. A data repository is not a trusted data depository.
 - Provide guidance on trusted data repository. See: <https://www.coretrustseal.org/>
 - Provide examples of trusted repositories for education, guidance, and reference.
- i. NIH Data Repositories and Trusted Partners: <http://tinyurl.com/yhkfyxn>
 - ii. Acceptable Digital Repositories for USGS: <http://tinyurl.com/y6e5d35n>
 - iii. Data Repositories Conformant with DOT Public Access Plan: <http://tinyurl.com/yen5fw2z>
5. Data Sharing Agreements, Licenses, and Other Use Limitations: Guidance on development of acceptable memorandum of understanding (MOU) for NIH funded projects involving collaborators, partners, and researchers within and across organizations. Recommend examples, exemplars, and references involving data use, reuse, and data governance policies (e.g. Creative Commons Zero, Open Data Commons).
6. Oversight of Data Management: The socio-technical management of research data at an institution requires collaboration of diverse stakeholders across multiple units involving the allocation of resources, responsibilities, and support. Recommend NIH resources (e.g. exemplars, use cases) on recommended guidelines for developing collaborative partnerships between stakeholders and Libraries (e.g. NIH P42 Data Management and Analysis Core (DMAC)) in coordinating institution-wide initiatives to educate and support researchers in sustainable compliance with data management and sharing practices, policies, and procedures throughout the life of funded project and beyond. See: <https://er.educause.edu/articles/2013/12/starting-the-conversation-universitywide-research-data-management-policy>

Submission ID: 1328

Date: 1/9/2020

Name: Jason Hilton

Name of Organization: Stanford University

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Other

Role - Other: Data Curator

Domain of Research Most Important to You or Your Organization:

data coordination

Other Considerations Relevant to this DRAFT Policy Proposal:

For the entire draft policy, I strongly agree with Michael Hoffman's comments made here:
<https://hoffman.bitbucket.io/2019/nih-data-management.html>.

The NIH has the opportunity for a much more impactful statement, but that is not realized with this draft.

Submission ID: 1329

Date: 1/9/2020

Name: Suzie Allard

Name of Organization: University of Tennessee

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: interested in data management/workflow for interdisciplinary work

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

scioo-technical aspects of data management & team science

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

This purpose statement is concise and thorough however it would be enhanced by adding a couple points:

1. Data itself is a research output that is a valuable commodity and it is important to establish policy to protect both the research investment AND the valuable product itself.
2. Data is the foundation for many of the new frontiers in scientific inquiry using AI-- therefore it is important to have standards for data collection and management that assure the highest fidelity data for these kinds of analyses.
3. Trusted data is an important component of a scientist's legacy in addition to being a driver for career development.
4. trust in our health sciences is important to the public and a solid data policy helps reinforce this trust.

Section II: Definitions:

Perhaps there should be a definition of what NIH means by replication. This will help separate it from reproducibility and could serve as a touchstone across domains.

Section V: Requirements:

The policy should also:

- specifically discuss how metadata standards will be identified and used by different communities of researchers.
- outline how the cyberinfrastructure for maintaining the data will be sustained.
- specifically call out data security in addition to the usual concern for human subject protection.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

There should be:

- wording to encourage machine-readable data in order to facilitate sharing and re-use especially with AI applications
- a section on data security
- a section on sustainability of data beyond the life of the grant period

Other Considerations Relevant to this DRAFT Policy Proposal:

There is a need for a national level commitment to support the cyberinfrastructure for research data. This may be an opportunity to start introducing this concept.

Submission ID: 1330

Date: 1/9/2020

Name: Mary Lee Kennedy, Executive Director

Name of Organization: Association of Research Libraries

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All

Type of Organization: Professional Org/Association

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

All

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Thank you for the opportunity to comment on the draft version of NIH Policy for Data Management and Sharing, and Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan. I submit the following views on behalf of the Association of Research Libraries (ARL), a nonprofit collective of 124 leading research institutions in the United States and Canada.

ARL recognizes the commitment of the NIH and the National Library of Medicine (NLM) to data-powered health advancements and data science, which are dependent on a robust curation and sharing culture.

Section II: Definitions:

The Association of Research Libraries (ARL) recommends that NIH include "well-documented" or "curated" in its definition of data sharing.

Section III: Scope:

The Association of Research Libraries (ARL) applauds the expansion of data sharing to all extramural awards, and the recognition that data sharing is part of good data management and practice.

Section V: Requirements:

In the interests of reducing complexity and burden, the Association of Research Libraries (ARL) encourages (to the extent practicable and scientifically valid) the NIH Institutes, Centers, and Offices (ICOs) to harmonize their supplemental guidance to this policy.

Section VI: Data Management and Sharing Plans:

As members of the Confederation of Open Access Repositories, hosts and administrators of institutional repositories of various types, and data curators, the Association of Research Libraries (ARL) community welcomes the opportunity to partner and consult on the development of desirable criteria for data repositories.

ARL welcomes the draft policy's steps toward the integration of data management and sharing plans (DMPs) within regular reporting intervals. While the draft policy does not call for machine-readable DMPs, recognition that the plan will be revisited with regular grant reporting is an important step toward creating a culture of active DMPs.

ARL welcomes the proposed reduction in faculty administrative burden that would result from "just in time" data management and sharing plans, and suggests that upon submission a plan be considered in draft, with the elements that need to be evaluated for scientific merit; and the full plan delivered upon award, allowing time for critical intra-institutional consultation (with research offices, computing, and libraries, for example).

ARL recommends that DMPs from funded awards be made available within the awardee's institution, if not publicly.

ARL recommends that NIH strongly encourage machine-readable, or "active" DMPs.

ARL recommends that NIH require or strongly encourage the use of data citation principles as well as persistent identifiers (PIDs) such as ORCIDs for data collectors/managers or digital object identifiers (DOIs) for data sets.

ARL recommends that NIH ICOs provide public guidance on good/exemplar data management and sharing plans.

Section VII: Compliance and Enforcement:

The Association of Research Libraries (ARL) welcomes the proposed enforcement of the policy as a term and condition of awards.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The Association of Research Libraries (ARL) thanks the NIH for including data curation and data preservation in allowable costs as these are necessary activities for meaningful data sharing.

ARL recommends that NIH collect and share data on any cost adjustments for data management between submission and award, and over the course of the awards, so that the community can benefit from data on estimated and actual costs.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The Association of Research Libraries (ARL) applauds the inclusion of guidance for the adoption of common data elements and standards for scientific data and associated metadata in data management and sharing plans.

Attachment:

Association of Research Libraries Comments on Draft NIH Policy for Data Management and Sharing.pdf

Description:

Association of Research Libraries Comments on Draft NIH Policy for Data Management and Sharing



Comments of the Association of Research Libraries Regarding “DRAFT NIH Policy for Data Management and Sharing, and Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan”

January 9, 2020

Thank you for the opportunity to comment on the draft version of NIH Policy for Data Management and Sharing, and Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan. I submit the following views on behalf of the Association of Research Libraries (ARL), a nonprofit collective of 124 leading research institutions in the United States and Canada.

ARL offers the following comments:

- ARL applauds the expansion of data sharing to all extramural awards, and the recognition that data sharing is part of good data management and practice.
- ARL recognizes the commitment of the NIH and the National Library of Medicine (NLM) to data-powered health advancements and data science, which are dependent on a robust curation and sharing culture.
- As members of the [Confederation of Open Access Repositories](#), hosts and administrators of institutional repositories of various types, and data curators, the ARL community welcomes the opportunity to partner and consult on the development of desirable criteria for data repositories.
- In the interests of reducing complexity and burden, ARL encourages (to the extent practicable and scientifically valid) the NIH Institutes, Centers, and Offices (ICOs) to harmonize their supplemental guidance to this policy.

- ARL welcomes the draft policy's steps toward the integration of data management and sharing plans (DMPs) within regular reporting intervals. While the draft policy does not call for machine-readable DMPs, recognition that the plan will be revisited with regular grant reporting is an important step toward creating a culture of active DMPs.
- ARL applauds the inclusion of guidance for the adoption of common data elements and standards for scientific data and associated metadata in data management and sharing plans.
- ARL thanks the NIH for including data curation and data preservation in allowable costs as these are necessary activities for meaningful data sharing.
- ARL welcomes the proposed enforcement of the policy as a term and condition of awards.

ARL offers the following additional recommendations for the Draft Policy:

- ARL welcomes the proposed reduction in faculty administrative burden that would result from “just in time” data management and sharing plans, and suggests that upon submission a plan be considered in draft, with the elements that need to be evaluated for scientific merit; and the full plan delivered upon award, allowing time for critical intra-institutional consultation (with research offices, computing, and libraries, for example).
- ARL recommends that NIH collect and share data on any cost adjustments for data management between submission and award, and over the course of the awards, so that the community can benefit from data on estimated and actual costs.
- ARL recommends that DMPs from funded awards be made available within the awardee's institution, if not publicly.
- ARL recommends that NIH strongly encourage machine-readable, or “active” DMPs.
- ARL recommends that NIH require or strongly encourage the use of [data citation principles](#) as well as persistent identifiers (PIDs) such as ORCIDs for data collectors/managers or digital object identifiers (DOIs) for data sets.
- ARL recommends that NIH ICOs provide public guidance on good/exemplar data management and sharing plans.
- ARL recommends that NIH include “well-documented” or “curated” in its definition of data sharing.

Thank you for your consideration of these comments.

Sincerely,
Mary Lee Kennedy
Executive Director
Association of Research Libraries

Submitted electronically: <https://osp.od.nih.gov/draft-data-sharing-and-management/>

Submission ID: 1331

Date: 1/9/2020

Name: Sue Miller

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All

Type of Organization: University

Role: Institutional Official

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

My friends have comments but need more time. N.I.H. should extend this comment period if it truly wants comments. Last year the N.I.H. Office of Science Policy and Dr. Erin Luetkemeier at many meetings said there would be a 90 day comment period when the draft policy comes out. This policy was announced November 6, 2019. January 10, 2020 is only 65 days to comment. There have been many holidays in November and December shortening this period. The Office of Science Policy needs to live up to its word and extend the time to a full 90 days until February 3, 2020 to respond.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

My friends have comments but need more time. N.I.H. should extend this comment period if it truly wants comments. Last year the N.I.H. Office of Science Policy and Dr. Erin Luetkemeier at many meetings said there would be a 90 day comment period when the draft policy comes out.

This policy was announced November 6, 2019. January 10, 2020 is only 65 days to comment. There have been many holidays in November and December shortening this period. The Office of Science Policy needs to live up to its word and extend the time to a full 90 days until February 3, 2020 to respond.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

My friends have comments but need more time. N.I.H. should extend this comment period if it truly wants comments. Last year the N.I.H. Office of Science Policy and Dr. Erin Luetkemeier at many meetings said there would be a 90 day comment period when the draft policy comes out. This policy was announced November 6, 2019. January 10, 2020 is only 65 days to comment. There have been many holidays in November and December shortening this period. The Office of Science Policy needs to live up to its word and extend the time to a full 90 days until February 3, 2020 to respond.

Other Considerations Relevant to this DRAFT Policy Proposal:

My friends have comments but need more time. N.I.H. should extend this comment period if it truly wants comments. Last year the N.I.H. Office of Science Policy and Dr. Erin Luetkemeier at many meetings said there would be a 90 day comment period when the draft policy comes out. This policy was announced November 6, 2019. January 10, 2020 is only 65 days to comment. There have been many holidays in November and December shortening this period. The Office of Science Policy needs to live up to its word and extend the time to a full 90 days until February 3, 2020 to respond.

Submission ID: 1332

Date: 1/9/2020

Name: Tonia M. Masson

Name of Organization: Society of Toxicology (SOT)

Type of Organization: Professional Org/Association

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

toxicology

Attachment:

SOT Comments on NIH Data Sharing Policy FINAL.pdf

Description:

Comments from the Society of Toxicology on the draft NIH Data Sharing Policy

2019–2020 COUNCIL**PRESIDENT**

Ronald N. Hines,
MS, PhD, ATS
Research Triangle Park, NC

VICE PRESIDENT

George P. Daston,
PhD
Procter & Gamble Company

VICE PRESIDENT-ELECT

Myrtle Davis,
DVM, PhD, ATS
Bristol-Myers Squibb Company

SECRETARY

Laurie C. Haws,
PhD, DABT
ToxStrategies, Inc.

SECRETARY-ELECT

Suzanne C. Fitzpatrick,
PhD, DABT
College Park, MD

TREASURER

Anthony M. Ndifor,
PhD
Janssen Research &
Development

PAST PRESIDENT

Leigh Ann Burns Naas,
PhD, DABT, ATS, ERT
Traverse City, MI

COUNCILORS

Virunya S. Bhat,
PhD, DABT
ToxStrategies, Inc.

Michael J. Carvan III,
PhD
University of
Wisconsin-Milwaukee

Anne H. Chappelle,
PhD, DABT
Chadds Ford, PA

Barbara L. F. Kaplan,
PhD
Mississippi State University

Cynthia V. Rider,
PhD, DABT
NIEHS/NTP

Courtney E. W. Sulentic,
PhD
Wright State University

EXECUTIVE DIRECTOR

Tonia M. Masson

TO: National Institutes of Health Office of Science Policy

FROM: The Society of Toxicology (SOT)

RE: Draft “NIH Policy for Data Management and Sharing and Supplemental Draft Guidance”

The Society of Toxicology (SOT) is supportive of the draft “[NIH Policy for Data Management and Sharing and Supplemental Draft Guidance](#).”ⁱ SOT strongly agrees that an open data policy:

- Promotes transparency in science.
- Allows for independent data validation.
- Facilitates the re-purposing of valuable data to support novel research at a cost-savings.
- Allows for the possible combining of data sets to answer questions that historically were difficult to approach because of limited resources.

Consistent with the February 22, 2013, Holdren memorandumⁱⁱ outlining important principles for a federal open data policy, SOT urges NIH to maintain a 12-month embargo on making publications and any underlying data publicly available. Such a practice will be critical for maintaining the long-term successful partnership between the scientific community and the publishing industry.

The Society is pleased that the NIH draft policy recognizes the value and challenges associated with making data from studies with human volunteers publicly accessible and is proposing that costs associated with anonymizing such data be an allowable expense. However, given the complexities and importance of anonymizing such data, SOT urges NIH to consider developing a centralized resource that would be made available to achieve this goal in a reproducible, reliable, and rigorous manner. Developing such a resource as a cross-agency effort may serve a wide scientific audience.

ⁱ “NIH Data Management and Sharing Activities Related to Public Access and Open Science,” National Institutes of Health Office of Science Policy, updated November 2019, <https://osp.od.nih.gov/scientific-sharing/nih-data-management-and-sharing-activities-related-to-public-access-and-open-science/>.

ⁱⁱ John P. Holdren, Director of the US Office of Science and Technology Policy, to the Heads of the Executive Departments and Agencies, memorandum, February 22, 2013, “Increasing Access to the Results of Federally Funded Scientific Research.” https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Submission ID: 1333

Date: 1/9/2020

Name: Mary Langman

Name of Organization: Medical Library Association & Association of Academic Health Sciences Libraries

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of those listed

Type of Organization: Professional Org/Association

Role: Other

Role - Other: health sciences information

Domain of Research Most Important to You or Your Organization:

health and biosciences information

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The Medical Library Association (MLA) and the Association of Academic Health Sciences Libraries (AAHSL) support NIH's commitment to making funded research results and outputs available to the public and advancing biomedical research by enabling the validation of results, combining datasets to strengthen analyses, facilitating reuse, and accelerating future research. We are pleased that the Policy focuses on the importance of reproducibility and reliability of research findings. We laude the purpose of this policy and appreciate that data sharing is the bedrock of the advancement of biomedical research exploration. We make specific recommendations in the following sections for success in implementing and complying with the Policy.

Section II: Definitions:

- We recommend that laboratory notebooks either be included in the definition of scientific data or defined as a separate term. Laboratory notebooks are the primary record of research where data is recorded within context. This information is vital for reproducibility, and it preserves research integrity by increasing transparency of experimental details and by showing provenance. It is important to clarify that this document may not be necessarily shared but nonetheless should be preserved.

- We recommend, based on our comments in later sections, that the definitions section be expanded to include additional relevant definitions. Examples would include what patient data should be considered for sharing.
- We also recommend that references to existing resources be provided to researchers (e.g., the National Network of Libraries of Medicine Data Thesaurus, <https://nnlm.gov/data/thesaurus>).

Section III: Scope:

We applaud NIH for covering all supported research that results in the generation of data in this Policy. However we urge NIH to consider how to better share patient data in a manner that is consented for by each patient, and when patients do not consent, that patient privacy is upheld. We also ask NIH to carefully consider the costs of sharing patient data, including the effort to de-identify it before sharing so that it cannot be later triangulated with other data sets and re-identified, and provide supplemental direct funds for this activity. We recommend clarifying that all scientific data is covered under this Policy, not just data associated with a publication.

Section IV: Effective Date(s):

We believe that any formal policy adopted by the NIH and the Office of Science Policy should allow institutions at least one year to make needed internal changes to policies and procedures in order to ensure compliance. We also urge NIH and the Office of Science Policy to continue to take the deep care it has taken thus far as it advances the adoption and implementation of this policy. We suggest setting effective dates that would allow researchers and evaluators to have as much educational and infrastructure support in place as possible. We expect education and training would take at least one year, and developing adequate infrastructure may take longer. For example, data repositories do not yet exist for many areas of biomedical research. Librarians and information professionals also need training to support the data management and sharing needs of their researchers. Given these parameters, we recommend effective dates no earlier than January 2022, if NIH plans to implement the Policy within the next year.

Section V: Requirements:

We are pleased that NIH is taking steps to implement data sharing requirements for all NIH-funded research. To ensure the Policy meets its goal of increasing the volume and quality of shareable research data, researchers, research administrators, program officers, ICOs, and the public will be well-served by having

- clear guidance, including clear definitions of what constitutes compliance (within both the spirit and letter of the law); and

- clear indications of both incentives for exemplary data sharing practices, and consequences for researchers who remain out of compliance (excluding those whose patients have not consented to having their data publicly shared).

While we appreciate the flexibility given to various ICOs, we are concerned that this lack of formal guidance will lead to inconsistencies in requirements and compliance, and ultimately result in uneven practices of data sharing. We recommend including assessment guidelines for ICOs (e.g. a rubric) to enable transparency in evaluations of Plans and to define what constitutes compliance. We suggest that the Plan elements be used to form the basis for a standard rubric for evaluation.

Section VI: Data Management and Sharing Plans:

- We disagree with Plans being submitted "Just-in-Time." This practice would send a clear message to researchers that planning for data management is not important and is secondary to the research proposal. We recommend that Plans be considered equally important and evaluated during the research proposal review, which would be in line with other federal agencies including the National Science Foundation (NSF).
- We further recommend that Plans be reviewed by the organization's Institutional Review Board (IRB) to ensure that there are appropriate human subjects protections in place, patient data sharing is clearly consented, mechanisms exist for identifying patient data where sharing is not consented, and data management plans for the sharing of patient data is supported at the organizational level, as ultimately institutions will be held responsible and liable. This additional step will create more administrative burden, and thus should be compensated by NIH in the direct funding of each grant.
- We strongly recommend that the Supplemental Guidance on Plan elements be incorporated into this section because it would send a stronger message that all these elements are important to consider.
- The research community also values materials that are produced alongside the scientific data as they are essential to interpret the data within its initial context, to its reuse, and to its replication. We recommend expanding the requirements to include not only scientific data but also any other essential materials (e.g. code, custom software).
- We support making Plans publicly available which would:
 - (1) hold the researcher accountable;
 - (2) allow others to learn and reuse as examples; and
 - (3) provide a set of textual data for mining and study.

- We suggest developing a database similar to PubMedCentral, that also supports updates to the Plans if and when necessary. Such updates should be version controlled so that changes are tracked and viewed.

Section VII: Compliance and Enforcement:

- While compliance is listed as a requirement, the Policy does not provide specific guidance for ICOs concerning how to develop robust and consistent models for defining compliance, or identifying and addressing noncompliance. As mentioned above in Section V, we are concerned that this could lead to an inconsistent approach among ICOs.
- As with the NIH Public Access Policy, we believe NIH must establish clear expectations and pathways to compliance. For example, adding a component within the Annual Research Performance Progress Reports (RPPR) that requires a description of compliance with the Policy would apply to projects across ICOs.
- We also recommend including a section in the Final RPPR for reporting how well the Plan had been completed or addressed (i.e., this is what was proposed and this is how the data was managed and then shared).
- As with publications, we expect that embargoes on data sharing along with information on when and where exactly the data will become available should be included in the Final RPPR.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

- It is unclear if allowable costs are in addition to the award or if these costs should be budgeted at the time of proposal submission.
 - o If the former, we recommend specifying caps on costs.
 - o If the latter, we strongly reiterate that the Plan should be part of the proposal so that researchers can account for costs at the outset.
- We recommend that guidance for allowable costs include the following specifics:
 - (1) salary support for personnel dedicated to data management such as a data librarian or information professional;
 - (2) acceptable repositories and associated costs; and
 - (3) maximum costs for recurring fees associated with long-term preservation and what amount of time is reasonable.

We strongly request that NIH consider that smaller institutions may not have existing research support in place, and that enacting the Policy may come with hidden costs and unforeseeable institutional burdens.

Submission ID: 1334

Date: 1/9/2020

Name: Jorge Contreras and Tammy Frisby

Name of Organization: University of Utah

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: all

Type of Organization: University

Role: Biothnicist/Social Science Researcher

DRAFT NIH Policy for Data Management and Sharing

Section VI: Data Management and Sharing Plans:

The Draft DMS Policy NIH offers no specific guidance regarding the types of NIH-funded data that must be shared. In fact, the principal descriptive statement regarding data types is a negative one: "NIH does not expect researchers to share all scientific data generated in a study" (p. 16). Rather, NIH leaves it entirely to the investigators to propose what types of data will be shared. Likewise, the Policy offers no guidance regarding when data should be shared by investigators: upon generation, upon manuscript acceptance, upon publication, or at some other date. Again, this critical data sharing variable is left entirely to the discretion of the investigators.

Instead of offering specific guidance regarding these key data sharing variables, the Draft DMS Policy requires that investigators describe their approach to data sharing in a Data Management and Sharing Plan (Plan), and that each such Plan undergo "programmatic assessment by NIH staff within the proposed funding NIH ICO" (p. 13). In theory, this assessment procedure could be used by NIH to ensure that investigators abide by NIH's stated goal of broad data sharing for the public benefit. However, nothing in the Draft DMS Policy guarantees that this will be the case. Instead, discretion is left entirely to the funding NIH ICO staff, with no minimum requirements at the DMS Policy level. The Draft DMS Policy offers little indication how such Plans will be assessed, the criteria that will be used in assessment, or the weight that such Plan assessments will have in the overall scoring and funding of extramural grant applications. If the DMS Policy is expected to ensure broad data sharing, then it should include explicit instructions to reviewing NIH ICO staff regarding the minimum requirements that Plans must include in order to be compliant with the DMS Policy.

Likewise, in order to facilitate the development of robust and meaningful Plans, NIH should offer potential applicants more detailed guidance regarding expectations for these key data sharing variables. Without such guidance, there is a risk that submitted Plans will do no more than offer vague and broad statements regarding data sharing with little meaningful content.

Along these lines, we also encourage NIH to provide formalized training to the program staff who will be reviewing data sharing Plans, and to encourage individual NIH ICOs to develop detailed scoring and evaluation criteria for submitted Plans, and to release these publicly. As part of the programmatic assessment of each Plan, specific feedback should be provided to applicants regarding any deficiencies identified in their Plans. Ideally, a formalized mechanism would be developed for publicly disclosing the key features of Plans from funded applications.

Finally, the Draft DMS Policy states that it will replace NIH's 2003 Data Sharing Policy (p. 7). However, in order to reduce ambiguity, NIH should make more explicit the effect of any adopted DMS Policy on other NIH data sharing policies, including the 2016 Cancer Moonshot PADS Policy, the 2014 Genomic Data Sharing Policy, and the like.

Attachment:

NIH DATA SHARING COMMENTS 1-9-20.pdf

Description:

Comment Letter

To: National Institutes of Health

From: Jorge L. Contreras, J.D.
Tammy M. Frisby, Ph.D
University of Utah, Salt Lake City, Utah

Date: January 9, 2020

Re: Comments on DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance dated October 30, 2019

We appreciate the opportunity to provide comments to the National Institutes of Health (NIH) regarding its Draft Policy for Data Management and Sharing (DMS Policy). At the outset, we commend NIH for its care and thoughtfulness in developing the Draft DMS Policy and its commitment to the broad sharing of scientific data.

By way of background, we have collectively studied data sharing in the sciences, and NIH's data sharing plans in particular, for more than a decade. Professor Contreras has published numerous articles discussing the benefits and drawbacks of these policies. A partial list of references is included below. In addition we have recently concluded a study of the implementation of the National Cancer Institute's 2016 Public Access and Data Sharing (PADS) Policy (article under review). Below, we offer some comments on the Draft DMS Policy.

The Draft DMS Policy NIH offers no specific guidance regarding the types of NIH-funded data that must be shared. In fact, the principal descriptive statement regarding data types is a negative one: "NIH does not expect researchers to share all scientific data generated in a study" (p. 16). Rather, NIH leaves it entirely to the investigators to propose what types of data will be shared. Likewise, the Policy offers no guidance regarding *when* data should be shared by investigators: upon generation, upon manuscript acceptance, upon publication, or at some other date. Again, this critical data sharing variable is left entirely to the discretion of the investigators.

Instead of offering specific guidance regarding these key data sharing variables, the Draft DMS Policy requires that investigators describe their approach to data sharing in a Data Management and Sharing Plan (Plan), and that each such Plan undergo "programmatic assessment by NIH staff within the proposed funding NIH ICO" (p. 13). In theory, this assessment procedure *could* be used by NIH to ensure that investigators abide by NIH's stated goal of broad data sharing for the public benefit. However, nothing in the Draft DMS Policy guarantees that this will be the case. Instead, discretion is left entirely to the funding NIH ICO staff, with no minimum requirements at the DMS Policy level. The Draft DMS Policy offers little indication how such Plans will be assessed, the criteria that will be used in assessment, or the weight that such Plan assessments will have in the overall scoring and funding of extramural grant applications. If the DMS Policy is expected to ensure broad data sharing, then it should include explicit

instructions to reviewing NIH ICO staff regarding the minimum requirements that Plans must include in order to be compliant with the DMS Policy.

Likewise, in order to facilitate the development of robust and meaningful Plans, NIH should offer potential applicants more detailed guidance regarding expectations for these key data sharing variables. Without such guidance, there is a risk that submitted Plans will do no more than offer vague and broad statements regarding data sharing with little meaningful content.

Along these lines, we also encourage NIH to provide formalized training to the program staff who will be reviewing data sharing Plans, and to encourage individual NIH ICOs to develop detailed scoring and evaluation criteria for submitted Plans, and to release these publicly. As part of the programmatic assessment of each Plan, specific feedback should be provided to applicants regarding any deficiencies identified in their Plans. Ideally, a formalized mechanism would be developed for publicly disclosing the key features of Plans from funded applications.

Finally, the Draft DMS Policy states that it will replace NIH's 2003 Data Sharing Policy (p. 7). However, in order to reduce ambiguity, NIH should make more explicit the effect of any adopted DMS Policy on other NIH data sharing policies, including the 2016 Cancer Moonshot PADS Policy, the 2014 Genomic Data Sharing Policy, and the like.

Thank you for the opportunity to comment on this important policy initiative.

Selected References

Jorge L. Contreras & Bartha M. Knoppers, The Genomic Commons, 19 *Annual Rev. Genomics & Human Genetics* 429-453 (2018)

Jorge L. Contreras, Leviathan in the Commons: Biomedical Data and the State, in *Governing Medical Knowledge Commons*, Ch. 2 (Katherine Strandburg, Brett Frischmann, Michael Madison eds., Cambridge Univ. Press: 2017)

Jorge L. Contreras, NIH's Genomic Data Sharing Policy: Timing and Tradeoffs, 31 *Trends in Genetics* 55-57 (2015)

Jorge L. Contreras, Constructing the Genome Commons in *Governing Knowledge Commons*, Ch. 4 (Brett Frischmann, Michael Madison, Katherine Strandburg, eds., Oxford Univ. Press: 2014)

Jorge L. Contreras, Bermuda's Legacy: Patents, Policy and the Design of the Genome Commons, 12 *Minn. J.L. Sci. & Tech.* 61-125 (2011)

Jorge L. Contreras, Prepublication Data Release, Latency, and Genome Commons, 329 *Science* 393-394 (2010)

Submission ID: 1335

Date: 1/9/2020

Name: Mara Blake, on behalf of JHU Data Services

Name of Organization: Data Services, Sheridan Libraries, Johns Hopkins University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: General interest in data

Type of Organization: University

Role: Other

Role - Other: Library and information professional

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We, JHU Data Services (<https://dataservices.library.jhu.edu>), fully support NIH's decision to implement a data management and sharing policy. In particular, we are pleased to see the strong encouragement of data preservation and sharing, without which scientific reproducibility and reliability cannot be assessed and improved upon. Also, given our experience over the last 7 years with helping researchers write data management plans for other funders and archive data within our JHU Data Archive (<https://archive.data.jhu.edu>), we support the requirement of a written data management and sharing plan. These plans help researchers hone their data collection practices prior to conducting research, which in turn improves the organization, security, and quality of data, as well as the ease of public data sharing.

Section VI: Data Management and Sharing Plans:

Overall, we see the benefits of the preparing of DMSPs for the just-in-time period as opposed to Plan submission at the proposal stage, which other funders require. We anticipate that investigators would be more motivated to invest thought into how they will successfully manage and share their data knowing that the project is under consideration for funding. However, we hear from our colleagues in IRB and research administration that this will not provide enough time for investigators to comply with institutional requirements. We suggest that the NIH provide the staff who review the plans training and tools to evaluate DMSPs using consistent criteria. However, one concern with just-in-time submission is a possible lack of incentive to share data if there is no merit given to data sharing during the initial grant evaluation. We hope NIH staff will have the power to strongly encourage sharing of all de-identified data, and we encourage NIH to seek suggestions from the community of academic library RDS for best practices on DMSP review workflows and criteria.

While we applaud NIH's efforts to encourage data sharing when possible, we hope that NIH will give clearer guidance to researchers conducting human subject research as to what they can share once the policy is finalized. In our experience, many researchers find de-identification of their research data daunting, and understandably so. We fear that this could lead to researchers avoiding data sharing, even when appropriate de-identification for sharing is possible. While we are fortunate at Johns Hopkins to offer support for disclosure risk screening for clinical data, as well as de-identification training provided by the library, many research institutes do not provide such support. Additionally, JHU does not currently provide disclosure screening for other types of non-clinical data with disclosure risk. We would like NIH to offer formal training and credentialing in data de-identification, as well as guidance for institutions on how to implement an institutional solution to ensuring proper practices with human subject data across disciplines. It would also be helpful to provide guidance to IRBs and individual researchers on appropriate participant consent language that describes data sharing accurately and clearly to human subjects. Furthermore, NIH should ensure that all sanctioned repositories provide clear instructions to depositors on appropriate given conditions for restricted or public access.

The draft policy also states that data should "be made available as long as it is deemed useful to the research community or the public" (p. 3). With this language, it is not clear what data is deemed "useful" and who will determine the usefulness of scientific data. We believe researchers would benefit from a decision support tool for determining what data can be shared unrestricted and what types and levels of de-identification could be beyond feasible scope and justifiably not shared as public access.

Many institutions provide their researchers archives that facilitate data sharing and curation, as well as services that help researchers identify appropriate data sharing outlets. For example, JHU Data Services offers consultation on identifying appropriate repositories for researchers and supports the JHU Data Archive (<https://archive.data.jhu.edu>), an open access institutional data repository. We suggest that NIH highlight such institutional services to researchers.

Section VII: Compliance and Enforcement:

We are pleased that NIH requests researchers to discuss anticipated timelines for data preservation and sharing (e.g., data sharing after manuscript is accepted). Having documented timelines should improve NIH's ability to monitor compliance. In addition to documenting anticipated timelines, we believe NIH should include a new section in annual reports that requires researchers to report their data management and sharing progress. Similar to the anticipated timelines, this will help NIH check if researchers are in compliance with their own data management and sharing plans. We suggest that NIH communicate this information back to academic institutions.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We were happy to see clear guidance on allowable costs for data management and sharing. This is an area that we often get questions about from either research administrators or researchers.

At JHU, we have heard stories circulating among researchers and data managers of requests for extra funds for data management and sharing being denied by NIH and other funders, such as for ClinicalTrials.gov compliance. Even if these are apocryphal cases, such rumors are another disincentive for data sharing. NIH should be prepared to follow through on assessing and approving where appropriate the allowable costs associated with data management and data sharing included in the budget. They should also consider publicizing cases in which extra funding for data sharing led to important accessible data sources. This might be done centrally by NIH or within institutions by research administration or library research data service communications. NIH can also assist these groups with advising investigators on "knowing what to ask for" in such budgeting, as an extension of the "Allowable Costs" supplemental guidance draft. Case-based guidance and decision tools could help investigators mediate the feasibility of their requests.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Overall, we believe that researchers will find the Supplemental Guidance on plan elements helpful as they are writing their plans. Of note, we appreciate that it mentions several resources for finding standards. While this information in the draft on documentation/metadata is helpful, we suggest adding something about the purpose of providing the documentation, such as "data and any accompanying code/other materials should contain all metadata and documentation required for someone else to understand and use your shared data." Even though specifics vary across standards, it would help to make the rationale behind all documentation clearer and more prominent.

In addition to discussing standards, we also appreciate how related tools, software, and code are explicitly mentioned in the guidance as necessary components for reproducibility and reuse. In our data archiving experience, we find that not all researchers think of these as important elements to document and preserve as part of their research. However, across our field and in our participation in the Data Curation Network, we see increased demand from research to archive and share code associated with research data.

However, we have concerns about the range of questions in the proposed NIH data management and sharing plans and the two-page limit. We expect some investigators may find it challenging to fit all of the details within two pages and may have to truncate important information. We would like to see NIH encourage investigators to create more detailed DMSPs for their internal use. Ideally, they should treat the DMSP as a "living document" possibly with the option to include a more detailed DMSP as part of the record for active grants.

Other Considerations Relevant to this DRAFT Policy Proposal:

The JHU Data Services data management group (<https://dataservices.library.jhu.edu>), established in 2011, was among the first academic library services providing direct consultative support for the NSF data management plan as a dedicated unit or to offer an institutional data archive. With three full-time and one half-time staff in this role, and another data librarian at the JHU Medical Library, we remain one of the largest academic RDM services in the United States. JHU Data Services features a service model supporting three main areas: consultations to researchers on data management, sharing, and research workflows; an educational and training program; and support for data sharing and the JHU Data Archive. Given that JHU applies for and receives more NIH grants than most academic institutions, our staffing and workflow model may demonstrate the upper limits of demand and capacity for DMSP support services. Going forward, it may be helpful for us to present our current support plans for expanded support for NIH DMSPs. These will likely need to adapt to contingencies when NIH launches the policy, but we would be happy to serve as a benchmark going forward, along with various support strategies by the academic RDM community.

JHU Data Services, part of the JHU libraries serves both the medical campus and several other Johns Hopkins schools as a central resource for data management plan support. As such, we work closely with Research Administration, IRBs and other compliance offices at JHU to be the point of referral for guidance for DMSP preparation and implementation, focusing on data sharing requirements. These offices link to the Data Services website resources on DMSPs and our central email account, dataservices@jhu.edu. Data Services also provides regular outreach to departments and faculty about available DMP support.

Investigators preparing DMSPs can contact Data Services via the central email to receive assistance from a Data Management Consultant. This email exchange can include inquiries about the type of data and data sharing options investigators are considering, allowing us to suggest data repositories, resources and content prior to their drafting the DMSP. We also refer investigators to the DMPTool.org interface, on which we have created customized guidance. The DMPTool allows researchers to send completed forms directly to Data Services consultants for review and feedback. At this time, we can offer additional guidance on data sharing repositories, including the JHU Data Archive when appropriate (typically when NIH or field-specific repositories are not available). Investigators are encouraged to contact us with questions throughout the process of drafting the DMSP. We are available for direct consultation in-person or by phone or video chat.

We maintain internal tracking of consultations, including indications of the repositories the researcher plans to use. We currently do not systematically track who receives grants or do any direct follow-up with investigators, apart from those who subscribe to our newsletter.

In addition to individual and group consultations, JHU Data Services offers a robust series of trainings and workshops to support JHU researchers. Topics includes such as best practices in data management, preparing data management and sharing plans, and de-identifying human subjects data for sharing. We recently collaborated with the JHU Center for Clinical Research Data Acquisition (CCDA) to develop and deliver a certification program for their adjunct staff. In collaboration with colleagues from the Data Curation Network, we preparing to offer a special workshop on data curation for NIH/NLM. JHU Data Services participates in field-wide initiatives such as the Data Curation Network (<https://datacurationnetwork.org>), a Sloan funded, multi-institution effort to leverage curation expertise, provide training and resources for data curation, and pursue research on data curation and re-use. We strongly encourage NIH to pursue more training options such as these.

Submission ID: 1336

Date: 1/9/2020

Name: Hae Kyung Im

Name of Organization: The University Chicago

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization: statistical genetics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose: I agree with the comments of Michael Hoffman

Section II: Definitions:

I agree with the comments of Michael Hoffman

Section III: Scope: I agree with the comments of Michael Hoffman

Section IV: Effective Date(s):

I agree with the comments of Michael Hoffman

Section V: Requirements:

I agree with the comments of Michael Hoffman

Section VI: Data Management and Sharing Plans:

I agree with the comments of Michael Hoffman

Section VII: Compliance and Enforcement:

I agree with the comments of Michael Hoffman

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

I agree with the comments of Michael Hoffman

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

I agree with the comments of Michael Hoffman

Other Considerations Relevant to this DRAFT Policy Proposal:

I agree with the comments of Michael Hoffman

Submission ID: 1337

Date: 1/9/2020

Name: Scott Edmunds

Name of Organization: GigaScience

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Genomic, Imaging, Mass Spectrometry

Type of Organization: Other

Type of Organization - Other: Data Journal

Role: Other

Role - Other: Executive Editor & Lecturer in Data Management at HKU

Domain of Research Most Important to You or Your Organization:

All areas of biomedical research creating large scale data

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I generally agree with the comments of Michael Hoffman

At the very latest data should be shared at the time of publication, and there should be encouragement and credit for doing this earlier. The Toronto workshop did a good job of this: <https://www.nature.com/articles/461168a>

Section II: Definitions:

I agree with the comments of Michael Hoffman

As Michael says promote and use the FAIR principles (I was one of the authors on the original paper <https://www.nature.com/articles/sdata201618>).

Section III: Scope:

I agree with the comments of Michael Hoffman

Section IV: Effective Date(s):

I agree with the comments of Michael Hoffman

Section V: Requirements:

I agree with the comments of Michael Hoffman

I would also strongly recommend the use of interoperable licenses such as creative commons CC-BY or CCO.

Section VI: Data Management and Sharing Plans:

I agree with the comments of Michael Hoffman

I could also add that machine actionability of these would be useful. See:
<https://doi.org/10.1371/journal.pcbi.1006750>

Section VII: Compliance and Enforcement:

I agree with the comments of Michael Hoffman .

The draft policy's requirements are too weak and it is meaningless without enforcement.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

I agree with the comments of Michael Hoffman .

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

I agree with the comments of Michael Hoffman .

Other Considerations Relevant to this DRAFT Policy Proposal:

I agree with the comments of Michael Hoffman .

Submission ID: 1338

Date: 1/9/2020

Name: Anshul Kundaje

Name of Organization: Stanford University

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization: Genomics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I agree with the comments of Michael Hoffman <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section II: Definitions:

I agree with the comments of Michael Hoffman <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section III: Scope:

I agree with the comments of Michael Hoffman <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section IV: Effective Date(s):

I agree with the comments of Michael Hoffman <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section V: Requirements:

I agree with the comments of Michael Hoffman <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section VI: Data Management and Sharing Plans:

I agree with the comments of Michael Hoffman <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section VII: Compliance and Enforcement:

I agree with the comments of Michael Hoffman <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Submission ID: 1339

Date: 1/10/2020

Name of Organization: International Society for Biological and Environmental Repositories (ISBER)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All data generated from research on biospecimens

Type of Organization: Professional Org/Association

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

All research collecting, storing, distributing and using biospecimens

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose: See attached document.

Section II: Definitions: See attached document.

Section III: Scope: See attached document.

Section IV: Effective Date(s): See attached document.

Section V: Requirements: See attached document.

Section VI: Data Management and Sharing Plans: See attached document.

Section VII: Compliance and Enforcement: See attached document.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

See attached document.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

See attached document.

Other Considerations Relevant to this DRAFT Policy Proposal:

See attached document.

Attachment:

ISBER Comments on the Draft NIH Policy for Data Management and Sharing_FINAL.pdf

Description:

ISBER Comments on the Draft NIH Policy for Data Management and Sharing



ISBER Comments on the Draft NIH Policy for Data Management and Sharing

The following comments are respectfully submitted from the International Society for Biological and Environmental Repositories (ISBER) in response to a request for comments on the Draft NIH Policy for Data Management and Sharing (November 2019).

ISBER is an international organization addressing the technical, legal, ethical, and managerial issues relevant to repositories of biological and environmental specimens (see www.isber.org for additional information). Although the great majority of ISBER members focus on providing human biospecimens and associated data for research, ISBER membership is open to all types of biorepositories. ISBER membership and expertise in the area of human biospecimens and associated data used for research is extensive, longstanding, ongoing, and representative of best practices in the field. ISBER's thought leaders in this area are worldwide.

ISBER is committed to ensuring that data are shared as widely as possible to further advances in scientific research while at the same time protecting the privacy of research participants and the confidentiality of their data. While biospecimens are specifically excluded from the draft Policy, the policy does cover research data generated from biospecimens. As such, we have a keen interest in NIH's Draft Data Sharing Policy.

General Comments:

ISBER strongly supports the overall goal of the policy to make the results and outputs of the research that it funds and conducts broadly available. However, we have some general concerns about the draft policy as written, as well as more specific ones, as detailed further below.

One major issue relates to how to share data obtained from human research participants broadly while still respecting their rights and welfare and complying with Federal, Tribal, state and local laws and regulations. There are major unresolved issues in this regard. For example, the revised Common Rule lacks clarity regarding secondary data sharing. In addition, privacy regulations in some jurisdictions (such as the EU General Data Protection Regulations) are making it extremely difficult for researchers to share data broadly. The Policy states that NIH recognizes that "there may be unique circumstances in which broad data sharing may not be appropriate (i.e., particularly sensitive data or data restricted by certain Federal, Tribal, state, and local laws, regulations, etc.)" and has asked for input on strategies for promoting responsible data management and sharing practices in these circumstances. NIH could play a lead role in helping to clarify and develop policies and best practices around consent and other legal and other ethical issues related to broad data sharing. In addition, it will be important for NIH to play a key role in addressing the challenges of data-sharing resulting from the EU-GDPR. This will be essential in order to facilitate implementation of this Policy and achieve the Policy's goals.

The Policy does not afford sufficient attention to the need for local governance or oversight of data-sharing to ensure that the rights and welfare of participants are protected. Without sound governance and oversight, broad secondary data sharing, inappropriate data-sharing and secondary uses of participant data could lead to loss of public trust in the research enterprise. Little guidance currently exists from the regulators or NIH on this important point. It will be important for NIH to develop (and disseminate) best practices for responsible data-sharing under this policy.



Additional guidance will also be needed in other areas to help investigators implement the Policy. For example, the Policy refers to sharing with existing data repositories. It would be helpful to provide a listing of such resources with web links, descriptions of the data the repositories contain, instructions for data deposition, etc. In addition, while the draft policy includes supplemental guidance regarding elements of data sharing plans, it would be helpful to provide specific examples of acceptable data sharing plans and guidance on the ways in which researchers can make their data available under the Policy.

Specific Comments:

Policy:

The Policy makes no mention of data quality and control and the need to QC the data before data are made available for broad sharing. Sufficient attention will be needed for careful QC before sharing to avoid further confusion in research findings and publications.

p. 1, second paragraph: The Policy states that “Shared data should be made accessible in a timely manner for use by the research community and the broader public”. However, it is not clear what is meant here with regard to the use of data by the broader public. No distinction is made in the policy between individual-level data and general research findings. No indications are provided on how this data should be made accessible and what oversight should be in place in providing research data in particular to the general public. Clarification of this point would be helpful to avoid misunderstanding about public access to individual level data.

p.1, II, Definitions: We suggest that “Data Management and Sharing Plan (Plan)” should include verification of data, deletion of data and error handling as part of management.

p. 2, IV, 4th bullet: The Policy mentions other funding agreements, such as Other Transactions. It would be helpful to clarify what “Other Transactions” are.

p.3, VI: Data Management and Sharing Plans. “Plans should also identify strategies or approaches to ensure data security and compliance with privacy protections are in place throughout the life of the scientific data”.

This point should include mention of the concept of ‘data governance’, or ‘guardianship’. Data governance groups, particularly in the case of some indigenous populations, are often responsible for access and best use of data for the duration of its life which includes broad secondary-use. The “Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan” alludes to data oversight in points 4, 5 and 6, however it does not adequately address the encompassing nature of guardianship.

p. 3, VI, 4th bullet: Data Management and Sharing Plans. The Policy should specifically address data management after the life cycle of the project other than “as long as it is deemed useful to the research community or to the public” as this could be a challenge post project funding.

This point should also include mention of applicable international laws and regulations.



p. 4. Post Funding or Support Period – This section would benefit from further clarification. How long does the responsibility to share data last? What happens at the end of the project period? Is the researcher still expected to make data available? If so, this would amount to an unfunded mandate.

Supplemental Draft Guidance: Allowable Costs for Data Management and Sharing:

It may be very difficult for researchers to accurately estimate the costs of preparing data for data-sharing in their initial budget request for a number of reasons. The demand for such data may be unknown (or even difficult to estimate) at the time of submission of the plan. Also, it may be difficult to anticipate the volume of data generated during any given funding period.

In addition, certain aspects of the supplemental draft guidance on allowable costs for data management and sharing are unclear. Will there be a line item in the budget for data-sharing? Will the budgets for allowing costs for data management and sharing be submitted Just in Time along with the data management and sharing plan or will they be submitted as part of the total budget submitted along with the initial application? If it is the latter, how will reviewers assess the budget for data management compliance if the Plan is submitted Just in Time?

Item 2. The supplemental draft guidance indicates that if the Plan proposes use of multiple repositories, the researchers may consider including costs associated with the use of each proposed repository. However, having the same data in multiple repositories could be problematic unless the repository is very clear about the source of the data so that researchers using data from multiple repositories understand this to avoid introducing further confusion into the literature.

On behalf of the International Society for Biological and Environmental Repositories (ISBER).

Sincerely,

Debra Leiolani Garcia
ISBER President
January 7, 2020

Submission ID: 1340

Date: 1/10/2020

Name: Jo Anne Goodnight

Name of Organization: The Jackson Laboratory

Type of Data of Primary Interest: Genomic

Type of Organization: Nonprofit Research Organization

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

Mammalian genetics and human genomics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

As stated in the Draft NIH Policy for Data Management and Sharing (herein, the Draft), the purpose of the NIH Policy for Data Management and Sharing is to reinforce NIH's long-standing commitment to making results and outputs of the research it funds available to the public. After listening to the video and reviewing the slides of December 16, 2019, we understand that the purpose of this policy is to: 1) include, in addition to a data sharing plan, a description of a detailed data management plan, and 2) be applied to all NIH grants, not just those asking for more than \$500K. We support the inclusion of a detailed data management plan as well as the application of a new Data Management and Sharing Policy to all NIH-supported grants.

Section II: Definitions:

The Definitions outlined in the Draft are appropriate. An additional definition might include:

- Persistent Unique Identifier: Persistent unique identifiers provide a means of long-lasting identification of digital objects that are global, standardized, and widely used in the digital environment and can provide information on the object, regardless of where the object is located. Assigning persistent unique identifiers to data helps to provide a method to locate data in the vast amounts of research data generated on a daily basis.

As explained below, we believe the use of persistent unique identifiers for investigators, projects and data are essential, and thus the term should be defined.

Section III: Scope:

We applaud and encourage the NIH to apply the Data Management and Sharing Policy to all NIH-funded research, including extramural grants, contracts, intramural research projects, or other funding agreements regardless of NIH funding level or funding mechanism.

Section IV: Effective Date(s):

We understand that implementation of deadlines are dependent upon feedback on this proposal. We further understand that we are encouraging the NIH to work with standards organizations and the European communities to learn from what they have already put in place (see below), which may take time. Accordingly, we urge the NIH to factor in sufficient lead time for implementing a Data Management and Sharing Plan to allow for adequate training and understanding of the requirements. We suggest that the "effective date" be no earlier than October 1, 2020 (i.e., fiscal year 2021).

Section V: Requirements:

We appreciate the importance of NIH implementing a policy that requests investigators to share data in a "timely manner for use by the research community." We are concerned, however, that the policy as proposed (see Data Preservation, Access, and Associated Timelines) specifies that scientific data should be made available "independent of award period and publication schedule." We urge NIH to give consideration to issues like timely review, analysis and data curation by those who generate the data. We also request that consideration be given to the potential for creating a competitive disadvantage that could result from the early release of data prior to project completion and publications. We agree that data sharing must be timely, however, "timely" should not be driven by administrative mandates. Rather the science should drive the timing and timing should be set based on agreement between the investigator(s) performing the research and the NIH Institute/Center/Office (ICO). Factors such as time needed for proper data curation and time to publish results should factor into the schedule. Moreover, the proposed policy does not discuss a rebuttal or appeal process should the investigator consider the timing established by the ICO to be unreasonable or unattainable. We encourage NIH to consider including language that addresses this possible scenario.

Section VI: Data Management and Sharing Plans:

We support the Draft plan that researchers with NIH-funded or -conducted research projects resulting in the generation of scientific data are required to submit a Data Management and Sharing Plan. Such plans should explain how scientific data generated by a research study will be managed and which of these scientific data will be shared, and how they will be shared.

We strongly recommend that the NIH adopt the principles laid out in the European 2018 FAIR Data Action Plan. The FAIR Data Action Plan states that open and documented formats for standards and code should be employed and that while minimum metadata and documentation is necessary to accompany these core data bits, enabling basic data discovery, richer information and provenance is necessary to understand why, when and by whom the data were created and accompanied with an appropriate data usage license (attached). We also believe that the NIH should educate the researchers on how to adhere to these principles and provide tools to facilitate the creation of the data management plans for their research.

For research initiatives that generate large volumes of data and require coordination among multiple laboratories (e.g., NIH Human Microbiome Project, ENCODE, IMPC, PDX Net, etc.), the trend at NIH has been to fund Data Coordination Centers (DCCs). While DCCs have, for the most part, been very successful at coordinating activities across different centers and delivering data to the community in standard formats, there is a danger that the application of different annotation standards at different DCCs will result in data silos. There are excellent examples of institutes and centers using a central data repository with excellent standards and formats, the IMMPORT data repository, <https://import.niaid.nih.gov> is a Nature Scientific Data recommended repository for cytometry and immunology. It holds the Core Seal Trust mark for trustworthiness since 2017.

In addition, once large initiatives are no longer active, the plans and funding for transitioning data and resources to community data resources for longer term stewardship are lacking. We also recommend that the NIH require DCCs to partner with community data resources such as the Model Organism Databases and other consortia, to ensure that the data analysis pipelines and annotation practices used by DCCs are in alignment with community standards and that a long term data stewardship plan is organized at the start of large-scale data generation projects. When large-scale data generation initiatives are funded, the budget for such projects should include appropriate provisions for long term data stewardship and not just for short term coordination activities provided by the DCCs.

To facilitate broader data sharing, the NIH may also consider funding innovation projects that propose to develop new platforms for facilitating the comprehensive 'discovery' of where data or knowledge of interest exists when such data cannot be openly exposed (e.g., private patient data). This way the scientific question is answered in an electronic assay where data are not compromised.

Section VII: Compliance and Enforcement:

We generally agree with the compliance and enforcement rules laid out in the Draft. However, we emphasize that development of, and compliance with, a detailed data management plan is a considerable burden for a researcher to manage on their own. Not all research institutes have the infrastructure to support data management plans. How will the NIH provide real, tangible support? We strongly suggest that the NIH bolster its investment in biomedical research by providing tools, tutorials, lessons and training in best practices for secure, efficient and ethical data management.

If the data are deposited in a FAIR manner, it is possible that automatic methods could be developed to check compliance and enforcement of a data management plan. For example, a testing algorithm could be developed and applied to challenge each specific data management plan and flag any errors for correction before submission. Enforcement will then be self-evident after researchers undergo training and are provided with the appropriate support to adhere to the rules and regulations of management. With more and more resources cloud based, services could be made available to the researcher to capture their data in a secure, private and consistent manner. To leave this compliance to the individual researcher is risky to the NIH without providing appropriate support that will enable the researcher to comply. It may not be possible to implement all manners of support simultaneously; we suggest that the priority might be on metadata standards and ontology tagging.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Based on the webinar and the wording in this Supplemental DRAFT Guidance, it is clear that there is some recognition of the need to allow for additional costs required to process and store collected data objects. We recommend that NIH ICOs allow costs for data management to be included during the Just-in-Time process based on the final data management plan that is negotiated with the investigator. This flexibility should also be provided during the Research Performance Progress Report (RPPR) period as it is possible that investigators could incur additional costs for data management.

The NIH is aware and it is usually well-known what the costs of particular services may be, such as sequencing costs, mass spectrometry costs, microscopy, flow cytometry, and the like. When asking for a data management plan, there must be additional funds permitted on top of the grant to allow for these data to be collected, processed, managed and stored for the duration of the grant. Finally, allowances should be given for archival services to ensure there is a golden copy for long-term access.

More and more frequently, sequencing centers are uploading their raw sequencing data into cloud buckets. When using a particular cloud provider, there are usually no costs for moving

the data within the same region (AWS or Google Cloud), but there are costs associated with moving the data to another region or downloading these data. If this is the case, NIH should consider allowing for the storage of data in the cloud and as well as for computing against these objects for the duration of the award. The NIH should consider following what the NCI did with the credits that were awarded to particular applications that be used against these storage costs. Credits are like cash permitting the use of a particular vendors platform (e.g. Amazon Web Services, Google Cloud, Azure Cloud). Bulk negotiation at the funders level will far outweigh what an individual research would be able to negotiate. Finally, for the particulars of what types of data objects that are generated, we recommend that the NIH provide clear guidelines for the long-term storage of these data objects and allow for the costs needed for the deposition of these data objects into long term storage. This would then complete the lifecycle of those data objects.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We applaud the effort by the NIH to provide guidance for the development of a Data Management and Sharing Plan. We would like to emphasize the following points for data management and sharing practices:

1. Data must have persistent unique identifiers. We strongly support the use of persistent unique identifiers for both individuals generating and depositing data, and for projects to streamline data management. Investigators generating and depositing data should be identified with unique identifiers (such as ORCIDs) while projects should be identified by their Research Activity identifier (RAiD). The Australian Research Data Commons (<http://raid.org.au/raid-faqs>) indicates that persistent unique identifiers help eliminate the administrative burden by facilitating automation and raise the visibility of the research encouraging cooperative practices. The FAIR Data Action plan emphasizes the need for persistent unique identifiers as they enable stable links to objects and provide support for citations and reuse. Persistent unique identifiers promote data interoperability and the ability for all data published to be readable by both human and machine as laid out by the W3C Standards.
2. The NIH must provide specific guidelines as to how data objects will be represented and found. We believe that NIH ICOs should, whenever practicable, use common data element formats and data standards for collecting data and information that is applicable across ICOs. We also believe that researchers should receive specific guidance and encourage the NIH to:
 - Define where specific data could or should be deposited.
 - Define best practice for metadata capture.

- Support the development of appropriate linkages and tools that allow the uniform and efficient upload of data and metadata.
- Promote the development of appropriate testing and quality control workflows to be embedded in the upload process.
- Encourage the use of time lines for adoption of standards.

We encourage the NIH to explore concepts introduced by the NCI Data Commons Pilot and other Data Commons initiatives, such as awarding cloud credits to facilitate not only the raw data upload but also appropriate data capture. If we are to move beyond realms of documents that are not really accessible by machines, and hence by machine learning, we will not be at the point, which we believe is the objective of the NIH, to have the ability to leverage on previously accessed and reasoned data.

3. The NIH should require that the data management plan define specific details of data production.

- What was the specific project question?
- What were its aims?
- How will the data be collected?
- What were the machines used to produce these data?
- What software was used to produce these data?
- What version of the software was used?
- What reference data were used?
- What version of the reference data were used?
- What were the specifics of the sample that was collected?
- How was the sample collected?
- What were the treatments, etc.

All of the metadata/annotation standards must be codified using the appropriate ontological terms (e.g., OBO Foundry, <http://www.obofoundry.org/>, is a repository for open-source ontologies). It is important to note the version of the ontological terms – as these are updated regularly. The appropriate way to interact with any of these services is through a service that

queries the terms. A particular conclusion made through an analysis published in a particular publication should make note of this.

4. Data management plans must address the issue of reproducibility. The code should be in a version-controlled repository, and the workflow should be presented in a common workflow language. The analysis should be in a trackable system to ease reproducibility.
5. A long-term data stewardship plan should be supported. To ensure the highest impact of data generated by both small scale and large scale data generation initiatives, the need for funding of core community data resources for long-term data stewardship must be factored in to the planning process.

Further guidance should be provided on how specific data sets should be deposited and how specific projects and samples should be encoded. We strongly encourage NIH to post sample data management and sharing plans similar to how it has provided example plans for "Sharing of Model Organisms and Related Resources." Rather than permitting free text, adherence to specific ontologies should be encouraged. The deposition of data into a repository should provide sufficient information regarding details behind how the data were collected, what machines were used, and the versions of software used to process the data before deposition. For example, with long read RNA-sequencing data, where should the circular consensus data be deposited and how should it be annotated? To support the use of specific controlled vocabularies for machines used for data collection, the NIH should consider providing tools to facilitate the practice. Data sharing is best done if complete metadata is captured including where necessary the library preparation steps. Ideally, these steps are codified as well and kept behind specific namespaces. To facilitate the creation of specific data management plans, we would encourage the NIH to provide tools to facilitate the creation of a data management plan. For example, a Data Stewardship Wizard was developed by institutes in the Czech Republic, Prague and the Netherlands (Elixir Czech Republic, the Institute of Organic Chemistry and Biochemistry of the CAS, Prague, Centre for Conceptual Modeling and Implementation Research Group, "Smart Data Management Plan", <https://ds-wizard.org/>) to help researchers develop their data plan. This was in response to the European Union FAIR data action plan published in June 2018 (see attached FAIR Data Action Plan) which outlined the core bits of information that should be collected on data to make data meaningful.

Description:

FAIR Data Action Plan

Submission ID: 1341

Date: 1/10/2020

Name: Dylan Roskams-Edris

Name of Organization: Canadian Open Neuroscience Platform and the Tanenbaum Open Science Institute

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All Neuroscience Data

Type of Organization: Other

Type of Organization - Other: The CONP is an association of Canadian research institutes and TOSI is the unit within the Montreal Neurological Institute dedicated to advancing the MNI's open science commitment.

Role: Institutional Official

Domain of Research Most Important to You or Your Organization: Neuroscience and Open Science

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose: See attached letter.

Section II: Definitions: See attached letter.

Section III: Scope: See attached letter.

Section IV: Effective Date(s): See attached letter.

Section V: Requirements: See attached letter.

Section VI: Data Management and Sharing Plans: See attached letter.

Section VII: Compliance and Enforcement: See attached letter.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

See attached letter.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

See attached letter.

Attachment:

Commentary on NIH Policy for Data Management and Sharing - CONP and TOSI - 1.9.2020.pdf

Description: Commentary from the Canadian Open Neuroscience Platform and the Tanenbaum Open Science Institute

To: National Institutes of Health

Re: Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance documents.

We begin by applauding the NIH's efforts to encourage the sharing of scientific data. This Policy, and the guidance documents that supplement it, represent a firm step towards ensuring that data generated by researchers delivers maximum benefit to society.

Of particular merit are:

- (1) the acknowledgement that sharing data can have associated costs and explicitly allowing reasonable costs to be offset with provided funding,
- (2) highlighting the need to share sufficient information for others to access and use the software tools needed to replicate findings using shared data, and
- (3) the clear statement that compliance with data sharing plans will be a key consideration in future funding decisions at *an institutional level*.

There are several areas where we believe further attention and development would increase the impact and utility of the Policy and guidance materials.

- (1) Imbedding links to relevant external material. Some examples: to the FAIR Principles (<https://www.go-fair.org/fair-principles/>) in the purpose section and to NIH's Sample Data Sharing Plan (<https://www.niaid.nih.gov/research/sample-data-sharing-plan>). Doing so should increase the clarity of intention and usability of the Policy and guidance documents.
- (2) One area of consistent confusion for researchers is the applicability of copyright to research data. Providing clear guidance on when copyright applies to data and how various licenses can be used would help alleviate some of this confusion. While we would advocate for an open science approach - meaning preferably a public domain dedication, but at most an attribution-only license – it is less important for the NIH's Policy that a particular approach is valorized than that some clear guidance is given.
- (3) Building on section 2 of the supplemental "Elements..." document we believe it would be appropriate to encourage the sharing of resources beyond software tools, including protocols, equipment, and materials. The definition of "Scientific Data" found in the Policy makes it clear that a major concern is sharing what is "necessary

to validate and replicate research findings...”. While software is of obvious importance, sharing protocols, materials, and equipment may be just as critical for validation and replication. Required sharing of information sufficient to find, access, and use these resources, written in a way similar to the current language around software, would strengthen the Policy from a replication standpoint and build on the NIH’s current policy concerning “Sharing Model Organisms.” (<https://www.niaid.nih.gov/research/sharing-model-organisms>).

- (4) Greater emphasis should be put on the importance of associating shared data with a unique persistent digital identifier (e.g. a DOI). Ensuring that such identifiers are associated with shared data increases the long-term accessibility of the data, eases the task of pointing to data within papers or other research outputs, and lays the ground for the generation of use-metrics.
- (5) The Policy and the scientific research community would benefit from greater clarity concerning the de-identification of human data. Accurate replication may, for example, require the inclusion of quasi-identifying characteristics such as a participant’s date of birth. While these characteristics alone are not in themselves identifying, the risk of re-identification increases as more metadata is shared. Providing guidance on standards which balance de-identification and inclusion of information sufficient for replication, either within the Policy itself or in supplemental materials, would be of significant value. Further, guidance on what standards of de-identification are required (i.e. coded, anonymized, or irretrievably de-linked) would be beneficial. Hand in hand with such guidance should be a clear statement of the NIH’s approach to users who attempt to use shared data to reidentify participants.
- (6) Finally, it should be made clear whether data sharing plans will be made openly available online and, if so, where and how to access them. Ideally there would be a dedicated space, perhaps within the NIH Figshare instance, where data management and sharing plans would be made openly available (taking into account any privacy concerns relating to information gathered from research participants). Doing so would help significantly with making shared data findable, identifying common practices within fields, and allowing public participation in the effort to ensure compliance with the Policy.

We look forward to the adoption of this Policy. We would be happy to collaborate with the NIH to help incorporate any of the suggestions above. We would, moreover, be pleased to assist in fostering the development of the Policy more generally.

Best Wishes,

Canadian Open Neuroscience Platform

Naser Muja, Executive Director

CONP Ethics and Governance Committee

Ann Cavoukian, Privacy by Design Centre of Excellence, Ryerson University

John Clarkson, Ontario Brain Institute

Jennifer Flynn, Division of Community Health and Humanities, Faculty of
Medicine, Memorial University

Richard Gold, Faculty of Law, McGill University

Judy Illes, Division of Neurology, Department of Medicine, University of British
Columbia

Bartha Knoppers, Centre of Genomics and Policy, McGill University (Chair)

Roland Nadler, Peter A. Allard School of Law, University of British Columbia

Walter Stewart, Walter Stewart and Associates

Adrian Thorogood, Centre of Genomics and Policy, McGill University (Manager)

Tanenbaum Open Science Institute

Vivian Poupon, Chief Operating Officer

Dylan Roskams-Edris, Open Science Alliance Officer

Composed by Mr. Dylan Roskams-Edris on behalf of the Canadian Open
Neuroscience Platform Ethics and Governance Committee and the Tanenbaum
Open Science Institute.

Submission ID: 1342

Date: 1/10/2020

Name: Keith Webster

Name of Organization: Carnegie Mellon University

Type of Data of Primary Interest: Imaging

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization:

Cognitive Neuroscience

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose: Please see attached

Section II: Definitions: Please see attached

Section III: Scope: Please see attached

Section IV: Effective Date(s): Please see attached

Section V: Requirements: Please see attached

Section VI: Data Management and Sharing Plans: Please see attached

Section VII: Compliance and Enforcement: Please see attached

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Please see attached

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Please see attached

Other Considerations Relevant to this DRAFT Policy Proposal:

Please see attached

Attachment:

NIH_Feedback Carnegie Mellon.pdf

Description:

Institutional response



Carnegie Mellon University
Hunt Library
4909 Frew Street
Pittsburgh, PA 15213-3890

Andrea Jackson-Dipina, Dr.PH,
Director of the Division of Scientific Data Sharing Policy,
Office of Science Policy,
NIH,
6705 Rockledge Drive,
Suite 750,
Bethesda, MD 20892

Re: Comment on NIH's DRAFT Data Management and Sharing Policy and
Supplemental DRAFT Guidance

Dear Dr Jackson-Dipina

On behalf of the Carnegie Mellon University (CMU) research community, the University Libraries has collated feedback from our research community and institutional leadership which responds to NIH's Draft Data Management and Sharing Policy and Supplemental Draft Guidance. This response is based on CMU's data sharing practices, our experience and institutional support in the data sharing arena, and specific feedback from those in receipt of NIHfunding.

We applaud the NIH for taking this important step in supporting data management and sharing, and we encourage the organization to follow through with the implementation of this policy providing clear guidelines for researchers, and appropriate enforcement of the policy. As an academic institution already supporting the future of scientific research that is interdisciplinary, collaborative, reproducible, and reusable, we are excited to have the opportunity to comment on the draft NIH Policy for Data Management and Sharing and Supplemental Draft.

By way of introduction, our principle feedback is that:

- i. We encourage public dissemination of Data Management Plans.
- ii. We recommend a more generous definition of scientific data that reflects the expansion of the scholarly record to include laboratory notebooks, code, protocols, and other research outputs.
- iii. We would welcome clarification on how the broader usefulness of scientific data is to be determined.
- iv. We would welcome further information on mechanisms that might be used to encourage and monitor compliance with the final policy and supplementary guidance.
- v. We encourage the earliest possible implementation of the final policy; we note that institutions, research libraries, and data management professionals have been building appropriate infrastructures for some time.

The new policy appears to support data management plans (hereafter DMPs) as living documents through a compliance period factoring in plan updates, which is an important step in encouraging researchers to regularly engage with their DMPs and ensure their research is following the protocol identified in the plans. As the draft policy states DMPs may be made publicly available, we believe this is a good practice in supporting compliance, education on DMP development, and facilitating broader best practices for a culture of open science.

As a research-intensive academic institution, CMU has identified several areas of opportunity in the draft policy, which the NIH may wish to consider when implementing the final version of the Policy for Data Management and Sharing. These areas are organized into five thematic sections, in the order of (1) definitions of scientific data and DMP guidelines, (2) data sharing, (3) costs, (4) compliance and enforcement, and (5) effective dates.

(1) Definitions of scientific data and DMP guidelines. The policy's *scientific data* definition notes laboratory notebooks are not considered data and do not need to be digitized. As an institution, we consider research products including laboratory notebooks to be valuable even if they are not considered data in this context, as they support reproducibility and reusability when included alongside data. While we understand the NIH does not consider these to be scientific data, we believe the NIH should encourage researchers to share relevant documentation along with data when possible. More broadly, we consider code, analysis environments, protocols, metadata schema, stimuli analysis pipelines, and other documentation to be essential

accompaniments to data, and implore the NIH to include language in the final policy encouraging local institutions to make these ancillary outputs of research a part of the scholarly record available alongside the data.

The current policy document states the researcher should limit their DMP to two pages. However, we encourage the NIH not to enforce a page limit, as projects will require differing levels of information depending on the type of data and the field of research.

(2) **Data sharing.** Within this draft policy, NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public. We would welcome clarification on how the decision is made to determine the broader usefulness of data. At CMU, we encourage our researchers to err on the side of sharing data, as we cannot predict all the future scenarios in which our data will be useful. As an institution in which a large proportion of our scholarly excellence and innovations are rooted in secondary reuse (computation, re-analysis, modeling) of scientific data, we deem it incredibly important to produce datasets for not only dissemination, but also reuse within and outside of our own research communities. We encourage the NIH to include language in the final policy document encouraging researchers to de-identify and share data when possible, and include language that clarifies budget allowance on related costs (further discussed in section 3). We also suggest encouraging researchers to share intermediate data when it is needed to ensure reproducibility of the funded project. We recommend data must be shared within 12 months of project end date. In Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (Plan), more information on what constitutes “findable” and “trackable” would be helpful for researchers, as would a statement on ethical data use and governance. We would encourage the NIH and funded researchers to consider the FAIR (findable, accessible, interoperable, and reusable) principles that allow for broad reuse and aggregation of data outside of the original discipline including making de-identified data discoverable, machine-readable, and combinable. On a related theme, we note the relationship between research data and the tools and software used in their generation. We encourage the NIH to consider making recommendations around standardization and/or curation and emulation of specialist software that may be required fully to utilize data shared under this policy.

(3) **Costs.** As one of many academic institutions with an institutional repository, we are unclear on the cost structure and allowable costs surrounding the use of these repositories. In reference to the Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing, would large data storage in CMU’s institutional repository, KiltHub, which is hosted on the Figshare platform, be considered an allowable repository cost or could this be considered institutional infrastructure that

should be covered by overhead? KiltHub allows storage of up to 1TB per project free of charge to CMU researchers, but additional storage needs require cost-sharing. Could our researchers include these additional costs within their funding proposal? Would this support be subtracted from research funding, or would this be considered separate from and therefore in addition to research funds? Similarly, we also would welcome clarification on the kinds of curation and de-identification services researchers can include within their budgets, including hiring a third-party curation service and/or using their institutional library's curation services. We also suggest providing language on additional costs the researcher(s) should consider in cases of data reuse. Will tools needed to run the data be usable or accessible in 10 years? Cost considerations for software migration, software preservation, etc. should be highlighted to the researcher(s) and encouraged for inclusion in the DMPs. In general, we encourage the NIH to include more detailed information on data archiving and allowable costs for the researcher.

We have witnessed a general trend of steadily declining costs of storage. Therefore, it is reasonable in the long run that data would be preserved in perpetuity. We encourage the NIH to determine appropriate responsibility for payment of long-term stewardship. Our recommendation is that we focus on institutional stewardship of data for a fixed period (10 years), at which point there is a review process through which data are dark archived or discarded.

(4) Compliance and enforcement. Regarding compliance and enforcement, we would like to see more information on concrete, trackable metrics that could be placed in the policy to encourage compliance, such as supplying a citation with a DOI or permanent URL for all datasets produced in grant reports. We are also unclear on how non-compliance will affect future funding decisions for the institution, including what constitutes non-compliance and which stakeholders will track compliance. As it currently stands, the policy seems to suggest an audit risk to the institution at large if researchers are not compliant with their plans. More clear information on non-compliance in the final policy would be useful to both researchers and their host institutions; for example, would changing the metadata schema used for the data from what is proposed in the DMP be considered non-compliance, or does this refer to larger efforts such as not appropriately sharing the required data? We also encourage the NIH to clarify how compliance with the DMPs and overall policy will be enforced. In support of discoverability, we suggest the NIH implements a system for creating a discovery layer across trusted/established repositories in which stakeholders can efficiently verify the location of shared data, which would also require the organization to encourage researchers to use appropriate metadata within their datasets to increase discoverability.

(5) **Effective dates.** Regarding Section IV (Effective Date(s)), we encourage the earliest possible implementation of the final data management and sharing policy. We believe the scientific community has had ample time to prepare for these data management and sharing mandates (given the 2013 OSTP data sharing memorandum), and institutions, research libraries, and data management professionals have been building appropriate infrastructures and policies to support these coming mandates.

Carnegie Mellon University welcomes the dissemination of the NIH's DRAFT Data Management and Sharing Policy and Supplemental DRAFT Guidance, and we look forward to the publication of the final policy and supplemental guidance in due course. Please do not hesitate to contact us should you have any questions or require clarification on any points made in this response.

Yours sincerely

A handwritten signature in black ink that reads "Keith Webster". The signature is written in a cursive, slightly slanted style.

Keith Webster
Dean of University Libraries
Director of Emerging and Integrative Media Initiatives
Carnegie Mellon University

email: kwebster@andrew.cmu.edu

Submission ID: 1343

Date: 1/10/2020

Name: Timothy J. Triche, Jr.

Name of Organization: Van Andel Institute

Type of Data of Primary Interest: Genomic

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Cancer (risk assessment, risk reduction, and treatment)

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Barring legal, ethical, or insurmountable practical barriers, publicly funded research data underlying a publication **MUST** be made available no later than the time of publication. To allow otherwise erodes the public trust in science, scientists, and the scientific process, directly contravening the mission of the NIH. Optional or unenforced guidelines will not advance the goals of data transparency, reuse, or stewardship of taxpayer monies invested into NIH research. Consequently, the only acceptable modification to the above expectation is for funding opportunities designed to create a shared resource; these must demand a clearly specified embargo date after which data will be disseminated even in the absence of a publication. For such opportunities the data is of greater importance than the publication, thus a failsafe must be included to require the former even in the absence of the latter. This is a special case of "no later than the time of publication" and is to be interpreted as "no later than the time of publication OR the expiration of embargo, WHICHEVER COMES FIRST".

The default, enforced policy **MUST** mandate sharing of data unless it is illegal, unethical, or impractical to do so. Taxpayers and the public are best served by transparent access to the results of research projects into which all taxpayers have jointly invested, unless the harms wrought by such access in a specific situation would outweigh the benefits. The onus must fall upon the researcher or project seeking an exemption to show that this is likely to be the case.

The current language is weak and must be remedied. With explicit guidelines for exceptions generally required by the Declaration of Helsinki or similar ethical standards, and with a proviso for well-justified exceptions required by novel investigations on a case-by-case basis, an opt-out (rather than opt-in) default for data sharing could be quickly implemented, uniformly enforced, and easily explained to members of the general public. This is the preferable path forward.

Section II: Definitions:

This section must include clear definitions of what constitutes FAIR data, and the 15 FAIR principles.

Section III: Scope:

This must clarify that the policy continues to apply for scientific data produced by funding in whole or in part from NIH after the NIH funding period is over. For example, in cases like the MESA project where a shared data resource is constructed under contract and then used for publications by individual investigators on a continuing basis, the expiration of funding for the project would reasonably be construed as the latest date by which data could permissibly be embargoed, rather than the date of individual publications employing the raw data generated under the contract.

Section IV: Effective Date(s):

The absence of an effective data management and sharing policy, and lack of enforcement for current guidelines, creates a serious negative impact on public health by hindering the dissemination of research. This, in turn, enables an ongoing waste of public funds. The noncommittal draft implementation date is unacceptable. The final policy must have a "no later than" date for implementation, ideally 12 months after issuance of the final policy.

Section V: Requirements:

Any data described as collected in a progress report must be deposited independently and an accession code or digital object identifier (DOI) supplied. Except when specified by the funding opportunity announcement, researchers may embargo this data until publication. Grant opportunities specifically designated to create a shared resource should specify a date by which data must be available even in the absence of a publication. Numerous repositories for this purpose exist. For example, large datasets and research artifacts (unpolished code, uncleaned data, and the like) are readily deposited in Zenodo, a CERN-maintained resource designed to house enormous datasets from physics experiments (the vast majority of biological and clinical experiments generate tiny amounts of data compared to that generated by particle physics). The terms of assignment and reuse for Zenodo are exemplary, and a DOI is generated by deposition. Any alternative repository must have similarly acceptable policies and "findability" for research artifacts to meet the needs of NIH, researchers, and the public.

It must be made clear that these requirements apply not just to research project grants and contracts, but also other forms of requests for support that will lead to the creation of scientific data. This includes cooperative agreements, career grants, fellowships, scholarships, and training grants.

Absent a compelling reason otherwise, contract solicitations must specify that collected data is the property of the NIH. They should also include specific requirements that data should be made publicly available in a third-party repository as a periodic deliverable, upon which further funding can be conditioned.

There are a large number of digital repositories with different policies. Acceptable digital repositories must not allow recipients to unilaterally change or delete deposited data. The repositories, may, however, allow adding new versions of data advertised in metadata for the original dataset.

It is important to protect human participant privacy but it is also important that concerns about human participant privacy not be abused to eliminate appropriate data sharing. It is especially worth considering that many human participants expect that data from their participation will be shared with other qualified researchers. Ineffective sharing of the resulting data (assuming appropriate protective measures such as de-identification are in place) is unethical as it wastes human participants' contributions to research and may result in more patients being exposed to harm. Therefore it must be an explicit goal of this policy and any submitted Data Management and Sharing Plans to maximize access in a manner consistent with participant consent and intent.

Section VI: Data Management and Sharing Plans:

A concise default (opt-out) data sharing plan for common data types, with clearly specified preferable standards and methods of dissemination, modeled after that required for National Science Foundation (NSF) grants, would advance the interests of researchers (reducing administrative burden), the public (setting expectations for research products), and ultimately NIH (by expediting large-scale analysis of the effectiveness and shortcomings of such a framework).

The current and insufficient draft "encourages" data sharing. An effective policy would instead REQUIRE data sharing, absent a compelling demonstration that the benefits of data sharing

would be overshadowed by the likely harms. An effective Data Management and Sharing Plan will increase the overall impact of a grant. An ineffective one will decrease it (in both real terms and, study section participants may be reminded, in the evaluation of a proposal).

It is important that Data Management and Sharing Plans be provided to NIH peer reviewers and ICO advisory council review so they can consider the plan's effect on the application's overall impact, significance, and approach. Guidance to reviewers on how to score review criteria such as significance and approach would do well to include review of the Data Management and Sharing Plan.

One may compare projects such as NHLBI TOPmed, with inefficient and baroque data sharing provisions, against the NCI TCGA or ENCODE projects. The former has generated tens of citations. The latter have generated tens of thousands of citations. It is unambiguously clear that the lasting scientific impact of the latter projects is greater, and this is due in no small part to the clearly stated requirements and effective implementation for data distribution.

Therefore, NIH should require Data Management and Sharing Plans at the regular submission due date for an application, not as Just-in-Time submissions. Overcoming deficiencies in the Data Management and Sharing Plan identified in summary statements could be provided as a Just-in-Time submission. The National Science Foundation (NSF) requires data management plans to be submitted with proposals for reviewer consideration. This is the correct approach.

NIH must also require that data management plans (DMPs) must describe how the researchers address each of the 15 FAIR Principles. A two-page template similar to that provided by the NSF would be sufficient to elicit suitable feedback in this manner, and to guide review and scoring of the resulting proposal. It would be advisable for these two pages to be independent from the overall page limit of grant proposals. An R03 proposal merits up to 2 pages of DMP. An R01 proposal merits up to 2 pages of DMP. Each project proposal within a P01, P50, or U54 might merit its own 2-page DMP, or they might be merged. Regardless of the nature of a solicitation, under no circumstances should a DMP be given less than 2 pages for a project. This will encourage substantially broader impact.

NIH should publish data management plans for funded grants and contracts alongside abstracts in public databases such as RePORTER. This will increase transparency and let other researchers and the public know what the grantees promised to NIH. This is the only way to enforce transparency with regards to individual plan items possible, in a manner that is both practical and consistent with both NIH's mission and that of the ORI. Failure to deliver generated data in a manner consistent with the proposed and funded data management and sharing plans, especially when the generated data is used to support published findings used to guide scientific and public policy, is a violation of research integrity and may merit referral to ORI.

NIH does not have the resources for exhaustive, systematic checks on compliance. However, secondary users of data, clinical or industry partners, and the general public have both the ability and incentive to flag research practices that are inconsistent with stated and funded DMP/DMSP proposals. Currently, data sharing plans are available through Freedom of Information Act requests. This hinders both enforcement and public engagement. Making data management and sharing plans for funded projects available on RePORTER will reduce the burden on data requesters as well as NIH personnel and, ultimately, the ORI.

The draft states that only data "deemed useful to the research community or the public" needs to be shared. This subjective phrase (who deems data useful or useless?) is unacceptable. Data sharing must be opt-out: unless there is a reason not to share data supporting a publication, finding, or resource, it is expected that it will be shared. Any exceptions to the general principle that scientific data must be shared must be justified and funding conditioned on prior approval by an NIH advisory committee of data management experts that includes data scientists and librarians. Failure to share data in a manner consistent with stated proposal aims is a specific type of failure to produce and must be viewed (and scored) accordingly. Data is expensive. This must be recognized.

For intramural research, you must not give a single NIH official (such as Scientific Director or Clinical Director) the ability to assess Data Management and Sharing Plans without oversight. Data Management and Sharing Plans must be reviewed and approved by Boards of Scientific Counselors and ICO advisory councils during the existing periodic peer review and site visit process.

Section VII: Compliance and Enforcement:

It is currently unclear where to turn when NIH data sharing expectations and policies are not followed. In my own experience, program officers associated with grants move on, and the chain of responsibility for enforcement is often broken when researchers refuse to disseminate funded, unencumbered data (i.e., data that has already been generated, is not ethically ambiguous, and distribution of which is consistent with participant wishes) in spite of NIH policy and contract specifications. To be clear, a breach of contract is a violation of civil law, and this oversight leaves injured parties without any means for redress.

Within the judicial system, such a situation would be viewed as a grievous systemic failure. An effective trans-NIH data sharing and management policy will address this shortcoming with equivalent urgency. The injury to secondary users and the taxpaying public accrues via misuse of millions of dollars of research project grant funding. This is a specific flavor of research integrity violation and its negative impact must not be minimized in an any effective policy.

To address this, RePORTER should list, for each grant, contact information to request corrective action for violations of the Data Management and Sharing policy or published Data Management and Sharing plans, to include contact email addresses for the principal investigators/project directors of the grant, contact email addresses for officials representing the grantee institution, and a contact email address at NIH. Specific policies and procedures for escalation must allow redress.

Repeated escalation by the same or collaborating groups suggests the possibility of willful violations of research integrity and is cause for auditing, sanctions, and eventual disqualification from certain contract solicitations. Similar information must be made available for contracts and for intramural research projects.

NIH ICOs may perform random audits to ensure grantees are performing data management as expected. Actionable complaints and escalations, with some form of documentation to identify repeat sources of complaints, whether as users or producers, would logically take priority in terms of enforcement.

Current sanctions listed in the draft policy are incredibly weak and will have no deterrent effect. The policy must mention that failure to follow the Data Management and Sharing policy can be considered research misconduct by NIH. The policy must specify that violating the policy in place at the time of competing award at any time thereafter (including after the end of the award period) will result in sanctions. These sanctions can include publication of a notice describing the violation in the NIH Guide to Grants and Contracts, debarment and suspension from contracting, subcontracting, or financial assistance from the federal government, and prohibition of service to the Public Health Service on advisory committees, boards, or peer review committees, or as a consultant. Because it touches on potential research misconduct, this policy must be reviewed by the HHS Office of Research Integrity.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The guidance should specify that fees that preserve data beyond the funding period are allowed, as are personnel expenses related to data sharing.

The National Science Foundation (NSF) stipulates that submissions must account for, and request, appropriate archival funding or a plausible externally funded and durable repository for secondary research artifacts (data, metadata, intermediary analysis summaries) produced as part of a viable data management plan.

The NIH would do well to adopt large portions of the existing NSF data management framework and its independent page limit in NSF grant proposals, not least because the true costs of responsible archival infrastructure and long-term data sharing are more apparent when they are made explicit.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Any suitable NIH data management and sharing policy elements must REQUIRE, in the absence of compelling reasons to omit, sharing of data generated by NIH funding. Taxpayers contribute billions of dollars per year to the NIH with the understanding that NIH will maximize the eventual scientific impact of this investment. It is both unethical and inexcusable to encourage a status quo (poor/inconsistent data sharing) when widespread and consistent data sharing has generated substantial impacts in academia, industry, and clinical settings (the Gene Expression Omnibus and Sequence Read Archive are sterling examples).

The stated goal of the NIH is to "[turn] discovery into health". The best way to accomplish this is to maximize the reach and impact of discoveries it funds.

An entry of "to be determined" in a Plan is not acceptable. This encourages superfluous Plans and is inconsistent with the purpose of the draft policy.

Statements such as "NIH does not expect researchers to share all scientific data generated in a study" are appalling in a draft data management policy. Instead NIH must make clear that sharing of scientific data is REQUIRED, with certain specific and limited exceptions, any variance from which must be justified by the applicant with prior approval by reviewers, program staff, and an NIH advisory committee of data management experts including statisticians and librarians.

Section 1 describes "consistency with community practices" as a potential rationale for deciding which data are preserved and shared. In many scientific disciplines, community practices lag far behind general best practices and what the public expects for data management and sharing. This language allows certain communities to settle for mediocrity in data management and sharing, defeats the aim of this policy to improve data management and sharing. It should be removed. This also illustrates why decisions to withhold scientific data from sharing should not only be reviewed by study section members trained in the same discipline but also an NIH advisory committee of data management experts that includes data scientists and librarians.

Section 4 says that "if an existing data repository(ies) will not be used, consider indicating why not". This policy should require the use of established repositories, except when exceptions are justified and approved. It should not be up to applicants to unilaterally decide not to use standard established repositories and to not even justify the same.

Section 5 anticipates that applicants may have restrictions on sharing imposed by existing or future agreements. This provides a major loophole in the policy in that applicants may choose to enter into more restrictive agreements than necessary so that they can avoid data sharing. This can be overcome by (1) providing data sharing plans as part of initial peer-review so that peer reviewers can appropriately score any decrease in impact that may come about from restrictions on sharing, and (2) review by an NIH advisory committee that includes data scientists and librarians.

Other Considerations Relevant to this DRAFT Policy Proposal:

It is time for NIH to adopt a clearly stated, unambiguous expectation of opt-out data sharing (i.e., all proposals are expected to share generated data by default, and it is the burden of researchers seeking exceptions to demonstrate why the overall benefit of the exceptions is greater than that of dissemination). The public seeks both transparency and ethical behavior from scientists and institutions. The NIH has an opportunity to demonstrate clear leadership, as the largest funder of health research in the world, in setting high standards for both science and transparency, and justifying public trust in scientific progress. Realizing this opportunity will further burnish the NIH's credentials as the worthy steward for billions of dollars of public funding, hopes, and dreams.

Three incentives merit addition to the section "Incentives for High-Quality Data Management and Sharing" section of the draft policy:

- 1) Add to the NIH biosketch a section for key personnel to describe their most significant contributions to data management and resource sharing (including data, code, reagents, samples, and other materials). This should be separate from other contributions. The past record of the principal investigator and other key personnel should be explicitly added to the scored review criteria.
- 2) NIH must create awards to recognize and cultivate excellence in data management and resource sharing, both at the individual and institutional levels.
- 3) NIH must periodically assess the impact of data sharing incentives upon research and its impact, with adjustments consistent to its overall mission.

The NIH has perhaps the finest reputation for scientific integrity in the world. I trust that the NIH will continue to set the highest of standards for research, and continue to turn discovery into health, as citizens become ever more engaged in their own health and ever more willing to contribute directly to research.

Submission ID: 1344

Date: 1/10/2020

Name: Robert Allaway

Name of Organization: Sage Bionetworks

Type of Data of Primary Interest: Genomic

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Cancer biology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I have read and support Anna Greene, Casey Greene, and John Wilbanks' comments on Section I of the Draft NIH Policy for Data Management and Sharing.

Section II: Definitions:

I have read and support Anna Greene, Casey Greene, and John Wilbanks' comments on Section II of the Draft NIH Policy for Data Management and Sharing.

Section III: Scope:

I have read and support Anna Greene, Casey Greene, and John Wilbanks' comments on Section III of the Draft NIH Policy for Data Management and Sharing.

Section IV: Effective Date(s):

I have read and support Anna Greene, Casey Greene, and John Wilbanks' comments on Section IV of the Draft NIH Policy for Data Management and Sharing.

Section V: Requirements:

I have read and support Anna Greene, Casey Greene, and John Wilbanks' comments on Section V of the Draft NIH Policy for Data Management and Sharing.

Section VI: Data Management and Sharing Plans:

I have read and support Anna Greene, Casey Greene, and John Wilbanks' comments on Section VI of the Draft NIH Policy for Data Management and Sharing.

Section VII: Compliance and Enforcement:

I have read and support Anna Greene, Casey Greene, and John Wilbanks' comments on Section VII of the Draft NIH Policy for Data Management and Sharing.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

I have read and support Anna Greene, Casey Greene, and John Wilbanks' comments on the Supplemental Draft NIH Policy for Data Management and Sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

I have read and support Anna Greene, Casey Greene, and John Wilbanks' comments on the Supplemental Draft NIH Policy for Data Management and Sharing.

Submission ID: 1345

Date: 1/10/2020

Name: Anthony Gitter

Name of Organization: University of Wisconsin-Madison

Type of Data of Primary Interest: Genomic

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization:

Computational biology

Other Considerations Relevant to this DRAFT Policy Proposal:

Overall, I would like to see a data sharing policy proposal that establishes stronger expectations for data sharing along with appropriate incentives and enforcement. My more detailed thoughts are aligned with those Dr. Michael Hoffman articulated at <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Submission ID: 1346

Date: 1/10/2020

Name: Henry Chang, M.D.

Name of Organization:

Type of Data of Primary Interest: Clinical

Type of Organization: Not Applicable

Role: Member of the Public

Domain of Research Most Important to You or Your Organization:

All

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

This and earlier versions of the NIH Data Sharing Policy are based on the premise that data are accurate and well-curated, whereas they often contain biases and flaws that make research unreproducible. The value of sharing is lessened if data are bad, so NIH should facilitate access for curation, as I explain below.

Section VI: Data Management and Sharing Plans:

Errors found in shared databases should be reported back to the NIH and corrected.

Section VII: Compliance and Enforcement:

There must be easier access for NIH staff to confirm the validity of primary data in RPPRs.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

See above and below.

Other Considerations Relevant to this DRAFT Policy Proposal:

I am a physician-scientist who retired from the NIH last year, partly in frustration with the lack of open access to data for the purposes of curation. As a program director, I felt the NIH public access policy limited our staffs just to make sure publications were freely accessible, without being able to confirm data were accurate. In my 40+ year career, I often had questions about scientific papers that could not be addressed easily because of lack of access to primary data.

My awakening to a scientific climate change came in 2005 when Dr. John Ioannidis published a seminal paper on "Why Most Published Research Findings Are False" (PLoS Medicine). About this time, a draft of the human genome had been completed, and concerns about genetic identification were raised, so perhaps this is why NIH leadership favored policies protecting privacy for almost a decade. In 2014, the NIH acknowledged that inaccurate or unreproducible research was a problem (Nature, 2014), and in 2016, a Nature survey of 1,576 researchers found that, "More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments." Even so, data sharing discussions have insufficiently emphasized data curation to improve data quality, because what is the point of sharing bad data?

I began to think about some scientific biases and blindspots that create barriers to data sharing:

1. Academia has many pressures to publish positive results quickly to claim priority, funding, and promotions. Research and papers contain insufficient rigor and are often rushed.
2. Journals pride themselves on impact factors and depend on advertising and subscription revenues. They don't like to publish negative results and retractions that reveal flaws in peer-review. Reviewers also have been sued when they challenged the validity of papers.
3. Sponsors don't like to admit they may have funded poor-quality research. In the case of the NIH, it would risk Congressional budget cuts, so it invokes the "self-correcting" nature of science. Based on the lack of reproducibility cited above, most research seems more of a gamble than an investment.

As an extramural NIH Program Officer, I began to see how these problems hurt the biomedical workforce. Too many scientists had wasted time chasing false leads and in turn, were generating unreliable results. We spent much time doing portfolio analyses of faulty research, so I tried to estimate how much bad data are in big databases, but couldn't get open access, as NIH staff go through the same process as the general public. This barrier exists at other institutes, and even managers of large academic/international collections couldn't tell me how much bad data they had, but spent about 80% of their time "cleaning" data. With the effort and costs of data sharing and storage rising, I didn't find much NIH support for curation, although I believe most patients want bad data corrected in order to prevent harm from medical errors. Instead, the NIH is taking the approach of artificial intelligence, but no one can explain how precision medicine can be achieved without precise data.

Nevertheless, here are my recommendations:

1. The NIH has the largest staff of biomedical workers in the U.S., and should use them fully, expanding their extramural role to more than portfolio analyses or public access to citations. Dr. Ioannidis also wrote a paper entitled "Science Mapping Analysis Characterizes 235 Biases in Biomedical Research" (J. Clin. Epidemiol., 2010). No one person can detect all such biases, so the more eyes on the data, the better. If given open access, NIH staff are less likely to violate privacy, and are unlikely to be sued for critical reviews of publications. Of course, extra training should be given to NIH staff to keep them scientifically engaged.

The user of any database release should be required to report back to the NIH any errors found (which is not the case currently). Curation will generate multiple versions, so plans for annotation, notification, and reanalysis are needed.

NIH grant applicants should get trained in data curation as part of the responsible conduct of research, and provide evidence they can critically analyze background information. For their publications, they should include a URL to primary data for peer reviewers to assess.

Research on data quality by Dr. Ioannidis and others is underfunded by the NIH. This should include studies on uncertainties in genetic, epigenetic, and environmental factors that would make the public favor more data sharing vs. privacy. This work should be a major part of the rigor and reproducibility mandate in Section 2309 of the 21st Century Cures Act or a Congressional line item.

2. Journal reviewers should also get a link to access primary data of the manuscript, and the publishers should keep comments in a moderated archive online. This resource would help readers to maintain critical thinking skills.

3. Academia needs to incentive collaborative research, data sharing, and curation. We have lost a generation of young scientists who wound up pursuing bad leads. Mentors may spend over 40% of their effort on administrative tasks, and have less time to teach and supervise. While collaboration is largely good, coworkers must remain attentive and avoid complacency or groupthink.

This scientific climate change should no longer be denied. The "hockey stick" of bad data in big data is too harmful to sweep under the rug or into cyberspace. It poses a threat to progress and public health.

Submission ID: 1347

Date: 1/10/2020

Name: Abeer Sarker

Name of Organization: Emory University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Free Text from any resource

Type of Organization: University

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Biomedical Informatics

Attachment:

DRAFT NIH Policy for Data Management and Sharing.pdf

Description:

Responses to specific points in the draft

Response to “DRAFT NIH Policy for Data Management and Sharing”

Abeed Sarker, Ph.D.
 Department of Biomedical Informatics
 School of Medicine
 Emory University
 Email: abeed@dbmi.emory.edu
 Twitter: @sarkerabeed

I read the draft with great curiosity and interest, and after finishing my read, I reviewed some of the publicly available comments that have been posted about it. I found a couple that were very much in line with my thoughts [1, 2], so I will just take this opportunity to primarily add my thoughts on top of those.

I outline my thoughts regarding specific statements made in the draft. For future reference, these comments are in response to the draft policy released on November 2019 [3].

I. Purpose

Some of the statements from this section that I would like to be updated are as follows:

“In addition, NIH emphasizes the importance of good data management practices, which provide the foundation for effective data sharing and improve the reproducibility and reliability of research findings.”

“Shared data should be made accessible in a timely manner for use by the research community and the broader public.”

I think the focus/purpose should be more directly about reproducibility and utility. The first sentence suggests that reproducibility and reliability are *desirable* criteria. They should be *essential* criteria. Furthermore, utility of the scientific outputs (‘outputs’ here refers to elements in addition to data; more on this later) of a study/project is not mentioned. Research advances are necessarily incremental. This is particularly true for my field—informatics. Current research in informatics builds on years of incremental progress, and past project outputs that had high utility for future research served as the platform. At the same time, there were funded projects that did not produce any outputs of utility for future research. Thus, there should be an explicit focus on utility.

I agree with Michael Hoffman [1]: the language must be more specific and needs to make explicit the components that *should* be shared and those that *must* be shared (see below for my thoughts about what *must* be shared).

“Data Management: *The process of validating, organizing, securing, maintaining,*

and processing scientific data, and of determining which scientific data to preserve.”

- ➔ There needs to be an explicit focus on reproducibility, particularly for informatics research. In this case, the broadly-defined scope of “Data management” should encompass management of systems developed, source codes and resources developed. The following are some specific examples:
 - Systems—trained machine learning models or rule-based systems that have been executed on the data and for which performance metrics (*e.g.*, accuracy, F1 score, precision, recall) have been reported.
 - Source codes—*python* or *R* scripts, along with parameter configurations and instructions for plug-and-play-type execution.
 - Resources—lexicons developed for natural language processing and external features for machine learning.
- ➔ While it is understandable that not all data can be shared without compromising privacy of the subjects in some way, there is ***no plausible reason*** for not sharing systems, codes and resources that have no influence on privacy.
- ➔ A guideline on data sharing is incomplete, particularly for informatics, if it does not include specifications for system/code/resource sharing.
- ➔ Every informatics researcher knows that *there are too many studies*, in every sub-field, which report performances of unavailable systems/algorithms on private/internal datasets. What purpose do such studies serve in progressing research?

“Scientific Data: *The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. NIH expects that reasonable efforts will be made to digitize all scientific data.”*

- ➔ The definition of *scientific data* needs to be broadened, as discussed above. In this case, I agree with Professor William Hersh [2].

VI. Data Management and Sharing Plans

- “*Researchers with NIH-funded or conducted research projects resulting in the generation of scientific data are required to submit a Plan to the funding NIH ICO as part of Just-in- Time for extramural awards”*

The plan should be for *guaranteeing reproducibility* not only for *sharing*, and it must be submitted with the proposal.

- “*Plans should explain how scientific data generated by a research study will be managed and which of these scientific data will be shared.”*

This is *very* narrow. Scientific data should be inclusive of data and resources. It should also include data that is *not directly generated* by the research project/study. Second, plans *must* explain how the *research methods and results* can be reproduced, particularly if the data used is not shared.

- “*NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public.*”

It should state the opposite. All data must be made public unless there is a compelling reason not to.

[1] Link to blog: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

[2] Link to blog post: <https://informaticsprofessor.blogspot.com/2018/12/response-to-nih-rfi-proposed-provisions.html>

[3] Draft available at link: [https://osp.od.nih.gov/wp-content/uploads/Draft NIH Policy Data Management and Sharing.pdf](https://osp.od.nih.gov/wp-content/uploads/Draft_NIH_Policy_Data_Management_and_Sharing.pdf). [Accessed January 10, 2020]

Submission ID: 1348

Date: 1/10/2020

Name: Holly Murray

Name of Organization: F1000

Type of Data of Primary Interest: Genomic

Type of Organization: Other

Type of Organization - Other: Publisher

Role: Other

Role - Other: Data Project Lead

Domain of Research Most Important to You or Your Organization:

Open research, open data, data management, data publishing, software publishing

DRAFT NIH Policy for Data Management and Sharing

Section II: Definitions:

Scientific data would benefit from mention of software, algorithms and source code (is this being considered as data by the NIH or not?) as well as a clear distinction between digital and non-digital data. Please also define data and indigenous data explicitly here.

Section IV: Effective Date(s):

The proposed timeline for implementation is unclear – but in our experience, this is likely to span years. Whatever the timeline, it must allow for reconciliation with existing policies and increased awareness and capacity building among researchers. Given the novelty of the policy, it would be worth starting with a pilot project, where, for e.g. a subset of NIH research areas are subject to the policy requirements, to allow any potential issues and challenge to emerge and be resolved – prior to introducing the policy more broadly. Similarly, the NIH could consider piloting DMPs before they are implemented across all funding schemes. This is also where collaboration with relevant partners can be tested.

Section VI: Data Management and Sharing Plans:

Key details that should be added:

- Comment on software, algorithms, and source code
- Management of data in the context of indigenous research

Section VII: Compliance and Enforcement:

I'd caution that clearer policy on monitoring and compliance is required for effective data management and sharing. What role, if any, would other stakeholders play in this process? For instance, do institutions have any responsibilities here?

This section would also benefit from a comment on 'credit' for producing a DMP and wider compliance with the policy so as to ensure this is not seen as simply another administrative burden.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

It is imperative that this guidance include management of data in the context of indigenous research as well as sensitive data. For instance, a clear statement that personal data should not be released, unless there is consent or legitimate basis for release, with reference to relevant data protection laws should be included. While this may seem somewhat repetitive, attention and care must be given to the protection of participant confidentiality. In our experience, it is useful to repeat guidance through a policy to avoid doubt.

In a similar vein, this guidance should spell out any acceptable exceptions to data sharing (for instance, where 3rd party data has been used – think multi-site studies). Each exception should include guidance on what the researcher should share (for instance, a metadata record for sensitive data).

Additional detail to be sought from researchers via the DMP:

- How will data be backed up during the research process?
- What quality control practises will be implemented?
- How might data be reused in other contexts?
- Who will be in charge of the plan's: implementation, review, revision?

Other Considerations Relevant to this DRAFT Policy Proposal:

I am concerned that the policy itself does not stand alone in relation to indigenous and sensitive data. Please be clear, how does this policy relate to the management of Indigenous research, knowledge and data? The THRO, USIDSN and CARE principles deserve specific mention. Similarly, how does this policy relate to sensitive data? What exceptions are in place for sensitive data? In what cases should managed access be used, in what cases should only a metadata record be shared?

I also strongly urge that the policy give preference to open formats and licenses (while access to the public is mentioned in the purpose, this seems to be lost throughout the policy). I envisage any data sharing policy provide a clear recommendation on the default license under which data be released; and this license must not unduly restrict text and data mining of research data. The EC2020 Guidelines, for example, state: "as far as possible, projects must then take measures to enable third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) this research data. One straightforward and effective way of doing this is to attach Creative Commons Licences (CC BY or CC0) to the data deposited." There is a real opportunity here for the NIH to back barrier-free data sharing and further their investments.

Submission ID: 1349

Date: 1/10/2020

Name: Sarah Damaske, Incoming Associate Director

Name of Organization: Population Research Institute, The Pennsylvania State University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Demographic, Qualitative, Big Data

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

NICHHD, NIA, NIDDK

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

PRI Response to NIH data sharing plan_IABdraft_v2.docx

Description:

Population Research Institute, Penn State Response to NIH draft sharing plan

There are several aspects of the latest version of the NIH data sharing plan that are greatly appreciated by the research associates and affiliates of the Population Research Institute of The Pennsylvania State University. First, we appreciate that this latest version of the NIH data sharing plan does not mandate data sharing. Second, it is very helpful that the data sharing plan will be due with the JIT portion of the application submission, rather than earlier in the process. Third, we believe that the second set of guidance about allowable costs to provide for local data management costs within the grants' scope will be useful.

Our remaining concerns are fourfold. First, the timing/flexibility of the release of the data must allow adequate time for data collection, data analysis, and data writeup, and publication. For example, if the time period is too short, investigators will face the administrative burden of collecting the data, carrying out the reporting requirements of leading an NIH grant, and complying with data sharing requirements, which may leave them without sufficient time or attention to publishing and disseminating the results of their study. Given that publication pressures are not as acute for established scholars, we are concerned that a tight timeline would be particularly disadvantageous to new investigators, early career scholars, and underrepresented scholars. We are concerned that this puts scholars at risk of not advancing their careers and continues rather than ameliorates the restricted mobility to higher rank in their home institutions. We also note that NSF has directorate specific guidelines that allow for differences within fields.

Second, despite the provision that data sharing is not mandated, it remains unclear how the current plan might shape the recruiting of clinical populations, qualitative populations, or other vulnerable populations. We see particular challenges for complying with IRB oversight, which may be reluctant to approve a project that requires data sharing and meeting these new requirements. (For some populations, any data sharing may put individuals at risk of identification or create situations where institutional gatekeepers to those individuals will not grant access for researchers.)

Third, we expect that standards within fields about best practices for data sharing may be inconsistent or even in contention. This may make it difficult to define what best practices are and create more uncertainty. Further, it suggests different levels of administrative burden for investigators from some fields versus others even when data are quite similar. Moreover, researchers in collaboration may straddle more than one field. Or individual researchers may work across disciplines themselves. In these cases, it is not at all clear which field guides standards for data sharing and security. Although the ambiguity in the guidelines offers flexibility, there could be unintended constraints that may counter efforts to increase collaboration and interdisciplinarity.

Finally, given that some data cannot be shared (due to privacy concerns, restrictions from data providers, etc.), it is important to provide guidance within this policy that emphasizes that shared data is not of greater value than non-shared data. While the majority of our population scientists see the value in sharing data, we also recognize there are limitations that prevent

some data from being shared, but which still have unique and valuable theoretical and empirical import to our fields.

Submission ID: 1350

Date: 1/10/2020

Name: Barbara Stranger

Name of Organization: Northwestern University Feinberg School of Medicine

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Member of the Public

Role - Other:

Domain of Research Most Important to You or Your Organization:

genetics and genomics of health and disease

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I agree with the comments of Michael Hoffman

Section II: Definitions:

I agree with the comments of Michael Hoffman

Section III: Scope:

I agree with the comments of Michael Hoffman

Section IV: Effective Date(s):

I agree with the comments of Michael Hoffman

Section V: Requirements:

I agree with the comments of Michael Hoffman

Section VI: Data Management and Sharing Plans:

I agree with the comments of Michael Hoffman

Section VII: Compliance and Enforcement:

I agree with the comments of Michael Hoffman

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

I agree with the comments of Michael Hoffman

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

I agree with the comments of Michael Hoffman

Other Considerations Relevant to this DRAFT Policy Proposal:

I agree with the comments of Michael Hoffman

Attachment:**Description:**

Submission ID: 1351

Date: 1/10/2020

Name: Duke University Libraries Research Data Working Group

Name of Organization: Duke University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: We support all disciplines and their data

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Team of Research Data Management Consultants, Data Librarians, Repository Manager, and Subject Liaisons from Social, Behavioral, Medical, General Sciences and Humanities

Domain of Research Most Important to You or Your Organization:

We support data management education, data workflow design, and data sharing and preservation across all disciplines

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We appreciate the concise yet clear explanation of the rationale for practicing good data management and prompt data sharing as a means to improve research reproducibility and reliability and facilitate new research. We feel communicating these reasons in this manner sets the tone that this is more than a requirement - it is necessary for good science.

Section II: Definitions:

The definitions are clear and useful. The inclusion of what is "not" scientific data will be very helpful for researchers to identify what type(s) of data they are expected to share. One small suggestion is that the statement "NIH expects reasonable efforts to be made to digitize all scientific data" seems like it might make more impact if moved directly to the Purpose section in the second paragraph following "Shared data should be made accessible in a timely manner for use by the research community and the broader public." So that it reads "Shared data should be made accessible in a timely manner for use by the research community and the broader public. NIH expects reasonable efforts to be made to digitize all scientific data." Those reading that statement can then refer to the definition of scientific data.

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

We think it would be beneficial for NIH to state that while researchers need to take "into account any potential restrictions or limitations" that they also need to address how those restrictions and limitations will be ameliorated for research transparency (i.e. instructions for how to obtain access, what files were used, what variables were analyzed and also including statistical analysis code). While reasonably addressed in the draft guidance it's worth noting that efforts at amelioration will be made.

Section VI: Data Management and Sharing Plans:

There are several key points in this section that we would like to respond to.

1. Reviewing plans at "JIT" - While we think this will help with efficiency in reviewing grant applications, there is some concern that other aspects of timing will be problematic. For example, the budget is due before the data management and sharing plan - what if the budget needs to be adjusted as a result of what they include in the plan? Will NIH catch that? Also, what if the data management and sharing plan is inadequate? How many revisions will be allowed? Will NIH staff assist researchers if they are struggling with their plan and do not have local support?
2. Allowing plans to be updated - this encourages the use of the plan as a living document rather than a box to be checked/form to be completed never to be looked at again.
3. Making plans publicly available - this is a good idea. It helps researchers understand what a good plan looks like, it will hold researchers accountable for sharing the resultant data as they stated in their final plan, and also serves as a means for others to locate the data associated with a particular grant.
4. Use of established repositories - we appreciate NIH making this recommendation. A repository will provide the mechanisms needed for long term preservation and access - which is not something a project website or portal can necessarily do. One question we do have is regarding the statement "NIH encourages shared scientific data to be made publicly available as long as it is deemed useful to the research community or to the public." Who determines that usefulness? If this refers to repository policies to retain data for a certain period of time based on access statistics (weeding) then that is not something that has to be stated outright. Instead, the ending statement of "NIH encourages the use of established repositories to provide long-term access and preservation for scientific data" could be used to replace that statement.

Section VII: Compliance and Enforcement:

It is mentioned that after the funding period, non-compliance with the plan may be taken into account for future funding decisions for the recipient institution. Instead of referring out to a

different policy (though citing it is useful) it would be best to include those potential sanctions here in plain text.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

While we are pleased that the NIH is advising researchers to use established repositories, researchers ideally should choose one repository for the canonical version of their dataset (when possible). Multiple repository use can cause issues with version control as well as raise budget costs for storage and preservation unnecessarily. It would also be advisable to mention that larger datasets (several GB and beyond) can result in higher fees for storage and preservation. While some repositories offer free storage up to a certain amount, researchers should be aware of those limits and account for potential overage. We are also happy to see that curation is explicitly listed as an allowable service to include in the budget. This is important for educating researchers on what data curation entails, and the value in using/applying curatorial standards for their data. Many established repositories (including institutional repository options offered through academic libraries like we have at Duke) have staff involved in performing curation. Some types of specialized curation may incur added costs (i.e. disclosure risk assessment) while others may be available free of charge.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Generally speaking, we found the guidance related to the elements of the data management and sharing plan to be clear and will assist researchers as they prepare their plans, although some sections could use additional guidance and clarification. We have provided detailed suggestions below.

Data Type: The two page limitation may hinder researchers who have data with more complex ethical dimensions, particularly human subjects' data, if they are expected to address issues fully related to confidentiality, de-identification, access conditions, etc.

Related Software and Tools: It would be useful to also guide researchers to include details about any software dependencies and version information for verification and reproducibility purposes.

Standards: While the focus on standards is useful in promoting uptake, if no discipline-specific standards exist the guidance could encourage the use of discipline-agnostic standards for data description such as Dublin Core or the Oxford Common File Layout. It should also be clarified whether a description of the documentation that will be provided alongside the data that does not fall within a standard (i.e., README files, etc.) should be described under "Standards" or "Data Types." Even when a standard is not available, this type of unstructured documentation

can be invaluable in reuse and we are concerned that this type of documentation may be overlooked with the current section structure.

Data Preservation, Access and Associated Timelines: We are greatly encouraged by how explicitly NIH is encouraging the use of established data repositories for data preservation and sharing. However, we think there is an opportunity to further encourage researchers to consult with existing repositories while drafting their DMPs. Repositories can help researchers work through where they should store the canonical version of their data (ideally a distinct dataset should only be assigned one DOI for persistent identification), and questions related to persistent identifiers, indexing, security and integrity, and restricting access (as appropriate).

Some of the language around timelines/timeframes could be clearer. First, how the lead-in sentence is structured it appears that timelines is the key aspect of this section versus the real focus which is determining where data will be preserved and made accessible (ideally in a repository). The last two bullet points could then be simplified to state "The timeframe should include when the anticipated data will be made available through a repository (when it will be deposited), and how long the data will remain available (after 7 years the data will be assessed for utility and potentially closed for access)."

Data Sharing Agreements, Licenses and Other Use Limitations: There is an opportunity in this section after the last bullet point to provide more specific examples of how one might share scientific data (and associated materials) while complying with limitations - namely by describing processes to gain access to restricted data, indicating files used, including scripts or code used to process or analyze the data.

Oversight of Data Management: We would also suggest adding to the example roles someone who oversees disclosure risk assessment as well as data curation experts (who may be based within a repository) as a reminder that these are critical data management roles that are often overlooked.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1352

Date: 1/10/2020

Name: John Wilbanks

Name of Organization: Sage Bionetworks

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Neuroscience, Oncology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Recommendation I.a: Make 'Timely' More Specific

The policy states that "shared data should be made accessible in a timely manner...". Timely should be defined, so that researchers understand the baselines expected of them, and have a boundary beyond which they must share data. These baselines and boundaries should be reflected in the templated DSMPS we recommend elsewhere. More details are provided in our recommendations in the Requirements section, and we further recommend in section VI that such DMSPs are scored elements of applications.

Recommendation I.b: Elevate the Importance of Data Management.

We applaud the mention of data management in both the purpose section and even in the title, but it is not given adequate attention in this section. The Purpose text does not address the lessons learned from data sharing within NIH-funded collaborations. From the Cancer Genome Atlas to Clinical Translational Science Awards to the Accelerating Medicines Partnerships, NIH has committed billions annually to such programs. Data sharing sits at the heart of these

collaborative networks, but our experience indicates that simply "sharing" data sets is not sufficient to meet the stated purpose. Data management is rarely elevated to a role commensurate with its importance in data reuse. As such, we recommend adding the following text to the end of the first paragraph to delineate the importance of management to achieving the purpose of the policy:

"Data management is an ongoing process that starts well before and goes on well after the deposit of a file under FAIR principles, and NIH encourages practices that have been demonstrated to succeed at promoting data sharing and reuse in previous awards."

Section II: Definitions:

Recommendation II.a: Amend the Definition of Data Management and Sharing Plan.

The definition of the Data Management and Sharing Plan does not sufficiently capture how DMS is integral to the research process. This Policy should make it clear that the data sharing is not an add-on or checkbox, but an ongoing management process that is integrated into the scientific research process. We recommend adding the following text to the definition of:

"The plan should describe clearly how scientific data will be managed across the entirety of a research grant and specific descriptions of how and when resulting data will be shared, including descriptions of which NIH approved repositories they will be deposited (or, if depositing outside this group, how the proposed repository will be sufficient to meet the requirements)."

Recommendation II.b: Replace the Definition of Data Management.

The definition of Data Management does not sufficiently reflect the true extent to which data management must permeate the research process, nor why it is important. Data management is a massive undertaking that improves the quality of shared data. We endorse the 2018 AMIA definition of data management and recommend that the NIH adopt it, replacing the current definition text with the following:

"The upstream management of scientific data that documents actions taken in making research observations, collecting research data, describing data (including relationships between

datasets), processing data into intermediate forms as necessary for analysis, integrating distinct datasets, and creating metadata descriptions. Specifically, those actions that would likely have impact on the quality of data analyzed, published, or shared."

Recommendation II.c: Add a Definition for Scientific Software Artifacts

The stated purpose of this policy is "to promote effective and efficient data management and data sharing." Per our recommended additions to the Scope section, below, the policy should make clear that what must be managed and shared are not only the "scientific data" and "metadata" created in the course of research, but also the scientific software artifacts created, such as the code underlying the algorithms and models that process data. Accordingly, we echo AMIA's call for definitions of "scientific software artifacts" and recommend NIH include in this policy the following definition:

"Scientific software artifacts: the code, analytic programs, and other digital, data-related knowledge artifacts created in the conduct of research. These can include quantitative models for prediction or simulation, coded functions written within off-the-shelf software packages such as Matlab, or annotations concerning data or algorithm use as documented in 'readme' files."

Recommendation II.d: Add a Definition for "Covered Period."

Making data available for others to use can pose a significant burden, per the supplemental guidance on Allowable Costs. Investigators will need clear definitions of exactly what will be required of them for data hosting in the short, medium, and long term. As such, we recommend that NIH include a definition in this section for "covered period," providing as much detail as possible on the expectations for the length of time that investigators must make their data available, including differences in requirements for research awards and data sets (including scientific software artifacts) of different scales.

Section III: Scope:

Recommendation III.a: Include Scientific Software Artifacts as an Asset to be Managed/Shared.

The first sentence in this Policy notes "NIH's longstanding commitment to making the results and outputs of the research that it funds and conducts available to the public." Scientific software artifacts (as defined in the response to the Definitions section, above) are outputs as much as data, equally determinative of research findings. Thus, managing and sharing the means of manipulating data from one form to another, transforming raw inputs into valuable outputs, is also important to the end goal of rigorous, reproducible, and reusable science. Furthermore, it is possible to technically share data while withholding key artifacts necessary to make those data valuable for reuse. These key artifacts could then be exchanged for authorship, position on proposals, or other scientific currencies, thus circumventing a major desired outcome of this policy: removing the unfair advantages of already funded investigators. As such, we recommend that the Scope section include the following statement:

"NIH funded research produces new scientific data and metadata, as well as new scientific software artifacts (e.g. the code of algorithms and models used to manipulate data). Software artifacts are outputs of research as much as data, and it is just as important to manage and share them in the interest of rigor, reproducibility, and re-use. NIH's commitment to responsible sharing of data extends to scientific software artifacts. As such, throughout this policy, the use of the term "data" should be understood to include scientific software artifacts, per the definition established in Section II."

Section IV: Effective Date(s):

Section V: Requirements:

Recommendation V.a: Tier the Sharing Date Requirement.

This policy will require cultural and practice changes for most funded researchers, as well as a nimble reaction to the realities of implementations by NIH. Failing to anticipate the implications of those changes could cause a severe backlash to the policy, undermining its purpose. As such, investigators of those projects least able to redistribute resources necessary to abide by this policy should be given more time to do so. We recommend that NIH adopt AMIA's 2018 tiered proposal for establishing sharing date requirements based on the size of funding. Projects funded over \$500,000 per year would have to comply within one year of approval of the DMSP, those between \$250,000-\$500,000 within two years, and those below \$250,000 within three years.

Recommendation V.b: Create DMSP Templates

We do not expect most researchers to know how to structure a Data Management and Sharing Plan. Furthermore, grants structure into different categories: the funding mechanisms behind the Cancer Genome Atlas and the AllofUs Research Program are different than early career researcher grants and most R01s. We therefore recommend the ICs create templates for at least four categories of funding: grants intended to create reference resources for the scientific community, grants that create collaborative networks of multiple laboratories, grants that form "traditional" research but integrate at least two institutions, and grants that only flow to a single institution.

These templates will facilitate understanding of the DSMP obligations by researchers (a form of learning by bootstrapping), as well as facilitate review by standardizing the essential elements and layout of the DSMPs across submissions. Researchers who do not use the standard template would not be penalized, but any DSMP they submit should clearly mark how and where their essential elements map onto the templates provided by NIH. Segmenting these templates by class of resources expected to be shared will make it easier for researchers to understand expectations (and can be tied to kinds of funding mechanism, e.g. U24) and will also make like-to-like evaluation easier for the NIH in evaluation over time.

Section VI: Data Management and Sharing Plans:

Recommendation VI.a: Make Data Sharing a Requirement

This section states, "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." However, the future utility of data is often unknown at the time it would be required for deposit, and it is unclear who would be responsible for deeming data as useful. We recommend that NIH require, not encourage, data to be shared. The NIH should also provide both alternate "sharing" mechanisms and opt-out processes for the situations when data sharing is either impossible or inadvisable (i.e. when sharing data would compromise participant privacy or harm a vulnerable group).

Alternate mechanisms could include a private cloud where users "visit" the data and are surveilled in their uses or "model-to-data" approaches where a data steward runs models on behalf of the community. Opt-outs should be rare but achievable, and patterns of opt-out usage should be tracked at the researcher and institution level to assist in evaluation of their use and impact.

Recommendation VI.b: Distinguish Between Purposes of Sharing

The requirements for data sharing should be different for data whose value to the community is realized in different ways. There is a difference between data that are generated with the explicit intent of creating a shared resource for the research community (e.g. TCGA), and data that are generated within the context of an investigator-initiated research project and are to be shared to promote transparency, rigor, and to support emergent long-term reuse. In the former case, a description of a detailed curation, integration, synthesis, and knowledge artifact plan should be present. In the latter case, a description of file format, simple annotation, and long-term storage should be front and center. We recommend that this section explicitly distinguish between these two purposes of sharing, and that different formats be used for developing and assessing DMSPs with respect to these different purposes.

Recommendation VI.c: Require the DMSP as a Scorable Part of the Application

In this policy, the DMSP will be submitted on a Just-in-Time basis. This signals that the plan is not a valued part of the application and is, in fact, an afterthought. NIH should factor the quality of the DMSP in its funding decision process. We recommend that the DMSP be required as a scorable part of the application so that appropriate sharing costs can be budgeted for at the time of application, and the plan can be included as part of the review process.

Recommendation VI.d: Make DMSPs Publicly Available

This section states that, "NIH may make Plans publicly available." We believe that NIH should ensure transparency with the public who has funded the work, and take advantage of transparency as a means for encouraging compliance. As such, we recommend that this section state that "NIH will make Plans publicly available."

Section VII: Compliance and Enforcement:

Recommendation VII.a: Give Investigators Time to Share

Judging an application based on performance on past DMSPs is only fair if the investigators have had sufficient time to implement that plan. Per Recommendation V.a (above) to tier the sharing date requirement, we recommend that application reviewers begin using evidence of

past data and software artifact sharing starting between one and three years after the adoption of the DMSP, depending on the size of the prior award. Those with a prior award of \$500,000 per year could be judged after one year of approval of the DMSP, those with a prior award between \$250,000-\$500,000 after two years, and those with a prior award of below \$250,000 after three years.

Recommendation VII.b: Use Existing Annual Review Forms for Proof of Compliance

Compliance with this policy should be integrated with current annual review processes for funded research projects. Proof of compliance should not require more than a single line in existing documentation, otherwise proof of compliance, itself, becomes an unnecessary burden of compliance. We recommend that NIH add a URL to a FAIR data file in annual review forms, alongside those lines for publications resulting from the data. This would provide an incentive to encourage a broad array of DMS practices and make it as simple as "filling the blank" on the form. We also recommend that NIH create an evaluation checklist as part of DSMP annual review to be filled out by the investigator and shared alongside the existing annual review forms.

Recommendation VII.c: Certify "Safe Spaces" for DMSP Compliance

Compliance and enforcement will also be significantly easier if NIH develops a process to certify data commons, knowledgebases, and repositories as "safe spaces" for DMSP compliance. Such a process could analyze the long-term sustainability of a database, its capacity to support analytic or other reuse, its support of FAIR principles, and more. Such a network would significantly "raise the floor" for the broad swath of researchers unfamiliar with FAIR concepts, for researchers at institutions without significant local resources to make data FAIRly available, and more. Accordingly, we recommend that this section include language detailing an NIH certification process for these resources.

Recommendation VII.d: Add data sharing and management experts to review panels

The composition of review panels is a key part of using DSMPs in award decisions. Ensuring data sharing and management expertise is represented as part of baseline review panel

competency will increase both initial review and also encourage long-term compliance with the key goals of DSMPs.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Recommendation VIII.a: Detail the Duration of Covered Costs for Preservation

The funding period for a research project is relatively short compared to the period after the research is complete wherein its outputs might be replicated or reused. Ideally, research outputs would be preserved indefinitely, but preservation has costs. The draft guidance does not specify whether costs to preserve data beyond the duration of the funded grant are allowed or encouraged. We recommend that this section provide detail as to whether NIH will cover data preservation costs after the funding period and, if so, for how long.

Recommendation VIII.b: Detail the Covered Costs for Personnel

DMS costs are not limited to the acquisition of tools, infrastructure, and the procurement of services; they also entail the time and effort of research staff internal to the investigating institution. The draft guidance does not specify whether personnel costs are allowable expenses related to data sharing. We recommend that this section provide detail as to whether NIH will cover such personnel costs - data sharing and management, done well, imposes a short term cost in anticipation of longer term benefit. NIH should clarify where that cost comes from as part of the Policy.

Recommendation VIII.c: Detail How Cost Levels Will Affect Funding Decisions

The Policy does not state whether a higher cost for better DMS might penalize (or advantage) a proposal in an IC's funding decisions. If potential recipients A and B propose to do the same research with the same traditional research costs, but A budgets for a robust "Cadillac" DMS plan, whereas B budgets for a bare-minimum "Chevy" plan, which does NIH choose? All things equal, should they choose the costlier, more robust option? Is it OK that it is a "tax" on the research proper? Is there an ideal ratio of traditional research costs to DMS costs? Is there a standard way to compare costs with benefits? We recommend that NIH provide detail in this section regarding how and if DMS costs will affect funding decisions.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Recommendation IX.a: Address Different 'Community Practices' Across Disciplines

Section 1 of this supplemental guidance states that, "Providing a rationale for decisions about which scientific data are to be preserved and made available for sharing, taking into consideration...consistency with community practices." However, different disciplinary fields can have different community standards. Some disciplines have a culture of sharing more, while in others it is less or not at all. Should all disciplines be held to the same DMS standards, or will investigators of different disciplines be expected to adhere to different community practices? If the former, how will this standard be established and what are the ramifications for compliance in disciplines currently outside of this standard? We recommend that NIH provide additional detail in this section (or, if necessary, in separate supplemental guidance) as to what the DMS expectations are within and across scientific disciplines.

Recommendation IX.b: Direct the Use of Existing Repositories

Section 4 of this supplemental guidance states, "If an existing data repository(ies) will not be used, consider indicating why not..." We recommend that the word "consider" be removed. This policy should recommend the use of established repositories and, if this is not feasible, then the investigator should justify their decision with a specific reason. We understand that many scientists are unaware of the infrastructure already in place, so we also recommend that NIH provide a list of existing data repositories with a certification of compliance to increase their use. Additionally, NIH may wish to provide guidance and build associated resources to assist investigators choose which of these repositories to use. If there are repositories that they must use (e.g. clinicaltrials.gov), or that NIH would prefer them to use, or that NIH has no preference (i.e., it would like the "market" to arrive at the best option), then NIH should make these degrees of requirement plain to investigators and make tools and infrastructure available to help them to decide.

Recommendation IX.c: Clarify Sharing Requirements for Data at Different Degrees of Processing and Curation.

Section 1 requires investigators to describe "the degree of data processing that has occurred (i.e., how raw or processed the data will be)." This raises the question as to whether the investigator can choose the level of processing and/or curation of the data to share, or if the investigator must share data at all levels of processing/curation. For purposes of reproducibility, we should encourage -- or require -- not only the sharing of data, but descriptions of data

processing at each level (per Section 2: Related Tools, Software and/or Code). This may, of course, increase the costs of DMS, so additional guidance would also be needed on what thresholds there may be and, the NIH should designate where the investigator has freedom to choose the levels of data shared and how the investigator should make tradeoffs.

Recommendation IX.d: Expand the Requirements and Guidance for Rationale.

Section 1 requires a rationale of which data to preserve or share based on the criteria of "scientific utility, validation of results, availability of suitable data repositories, privacy and confidentiality, cost, consistency with community practices, and data security." This rationale is limited to the choice of which data to share, while there are other important DMS decisions that warrant rationales. We recommend NIH require a rationale on where to share it and how long it will be available (Section 4), in what format it is shared (Section 3), and what other things might be shared, such as algorithms (Per Section 2). As with the choice of which data to preserve and share, NIH should offer criteria for decisions in each of these areas as well.

For choices regarding data preservation and sharing, as well as these other choices, if NIH has any preferences on how to weigh and balance criteria, we recommend it make those plain through additional guidance. Further, it should develop tools and infrastructure to help investigators to weigh and balance them, and conduct periodic audits/evaluations to understand how investigators across fields, over time, are making these judgements, if those judgements are in the best interest of the scientific community, and what additional incentives/requirements might be put in place.

Other Considerations Relevant to this DRAFT Policy Proposal:

Recommendation X.a: Detail how NIH will Monitor and Evaluate the Implementation of this Policy

A planning mechanism without an evaluation mechanism is only half complete. This policy should establish an adaptive system that improves DMS over time through feedback and learning. We recommend that this policy contain a new section that details how NIH will monitor and evaluate performance toward individual DMSPs during the funding period and after, to the extent that data are planned to be preserved after. Further, we recommend this new section also detail how NIH will monitor and evaluate implementation of this policy across all DMSPs, using evidence to illustrate how its purpose is or is not being achieved and what changes might be made to improve it. Policy-wide monitoring and evaluation information and reports should be made publicly available. Publicizing measures (e.g., usage rates and impact of

previously shared data) is also a way to promote a culture where investigators are incentivized to produce datasets that are valuable, reusable, and available.

Attachment:

Description:

Submission ID: 1353

Date: 1/10/2020

Name: Jeffrey Kidd

Name of Organization: University of Michigan

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

genomics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

I am concerned that this policy appears to be vague and represents a step backward in requiring open access to generated data sets.

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1354

Date: 1/10/2020

Name: Stuart Buck

Name of Organization: Arnold Ventures

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of the above

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Biothnicist/Social Science Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Biomedical research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

See attachment.

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Comments on NIH Data Sharing2019.docx

Description:

Response to Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Stuart Buck
Vice President of Research, Arnold Ventures

Susan M. Fitzpatrick
President, James S. McDonnell Foundation

Dawid Potgieter
Senior Program Officer and Head of Program Management
Templeton World Charity Foundation

January 10, 2020

As private philanthropic funders, we are dedicated to improving the reliability and validity of scientific evidence across fields that inform governmental policy, philanthropic endeavors, and individual decision-making. As part of our continued efforts to ensure that scientific research is fundamentally sound, these comments will address NIH's requirements for sharing data, code, and other research materials.

I. The Definition of Scientific Data

The NIH currently proposes to define “scientific data” as the “recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications.” The definition excludes, however, materials such as “laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects.”

We would suggest a few modifications that would further increase the value obtained from NIH-funded projects.

Possible Loopholes

As an initial matter, the focus on data that support “research findings” and/or “scholarly publications” could create at least two loopholes.

The first loophole is that due to how often journals and scholars alike prefer novel and positive results, some research projects may be seen as “failures” that do not lead to “scholarly publications.” Since the 1970s, this phenomenon has been known as the “file drawer effect.”

Yet the data collected in those research projects might often provide tremendous value to the scientific community. To take a hypothetical example, suppose an NIH-funded researcher explores genetic predictors of pancreatic cancer, but having found no significant genetic links, is unable to publish the results. The rest of the research community would benefit from knowing what happened in this line of research, so that they avoid further wasted effort and focus on other areas instead. Simultaneously, the NIH would have wasted its funding on that project if the results (however disappointing) remained buried. For another example, “failed” clinical trials can

be examined to look for genetic and other characteristics of so-called super-responders, that is, people in the treatment group who had surprising recoveries from a terminal disease.¹

As Francis Collins and Larry Tabak have said, “there is the problem of what is not published. There are few venues for researchers to publish negative data or papers that point out scientific flaws in previously published work. Further compounding the problem is the difficulty of accessing unpublished data — and the failure of funding agencies to establish or enforce policies that insist on data access.”²

The NIH’s general data sharing policy should address this problem by developing policies that would encourage scientists to share data and results *even from projects that went unpublished*.

For example, if applicants demonstrate that their lab has a track record of using preprint servers to post the results and accompanying data from studies that would otherwise have remained unpublished, study sections could rate that behavior under either Additional Review Criteria or Additional Review Considerations. More broadly, the NIH could reward other cases in which researchers take the trouble to share data even if otherwise unpublished.

The second loophole is over what it means to refer to “data used to support scholarly publications.”

In many cases, the scholarly community would benefit not just from the data that was literally “used to support” a publication, but from additional data that was *collected as part of the same project and could have been used*.

To take a hypothetical example, suppose that an NINDS-funded scientist runs a mouse experiment on a new stroke treatment, and collects data on 20 independent variables (such as experimenter’s gender, time of day, etc.). The scientist then publishes an article on the stroke treatment that uses *only* 5 of the 20 possible independent variables as part of the analysis.

If the scientist shares data on only those 5 variables, the relevant scholarly community would miss out on any insights to be gained from the other 15 independent variables. For us all to get the full value of the research data that NIH has funded, the other 15 variables should be shared as well.

The same is true for any research project that collects useful data but publishes an article only on a subset of that data.

Thus, rather than stating that “NIH does not expect researchers to share all scientific data generated in a study,” the rule should state, “NIH expects researchers to share all scientific data

¹ H. Ledford, “Cancer researchers revisit ‘failed’ clinical trials,” *Nature News & Comment* (18 April 2013).

² F. S. Collins and L. A. Tabak, “NIH plans to enhance reproducibility,” *Nature* 505 no. 7485 (27 Jan. 2014), available at <https://www.nature.com/news/policy-nih-plans-to-enhance-reproducibility-1.14586>.

generated in a study, with exceptions only when justified by researchers to a panel that includes subject matter experts and data experts.”

Some may contend that if researchers are required to share all data that *could have been* used to support a publication (rather than only the data elements that *were in fact* used), there will be a risk of scooping.

That risk, however, is minimal. For many experiments or clinical trials, the primary research article may take years before it is finally published. By that time, the original research team should have such a head start on analyzing the data that no one else could possibly beat them as to the secondary publications.

Moreover, the NIH funds scientific research to benefit humanity and the progress of science, not to advance the careers of particular researchers. If Researcher A makes available a dataset that leads to a scientific advance by Researcher B, we all benefit both from that scientific advance and from the greater efficiency and productivity of NIH’s funding, even if Researcher A wishes that he or she had gotten there first.

That said, the NIH should consider how to reform its practices of “crediting” the act of data sharing, so that research teams receive appropriate credit for creating a dataset that other researchers find useful. To take the most obvious example, NIH biosketches currently contain a section in which applicants can list their five “most significant contributions to science.” NIH could explicitly permit applicants to list occasions on which their sharing of data led to third-party publications and scientific advances.

A New Category: Raw Data

As a final note on the definition of “scientific data,” the proposed definition does not address an important point: does “data” mean the *final* data used for analysis purposes, or does it include data in a *raw* form before preprocessing and cleaning?

This ambiguity plays out in a specific example: the definition of “scientific data” excludes “case report forms,” but such forms are essentially the raw data about what happened in a clinical trial, and sharing that raw data can be immensely useful. One of the most well-known cases of clinical trial misreporting occurred as to the drug paroxetine (or Paxil); it was only when later investigators were able to get access to the case report forms that they found numerous cases of adverse events that had never been reported elsewhere.³

To clarify this point, the NIH should create a category of “Raw Data” that includes case report forms as well as the underlying raw data types for other studies. Then it should convene with ICs and experts in each substantive area of research to define what counts as “raw data” that would be useful to share.

³ J. Le Noury et al., “Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence,” *BMJ* 351 (Sept. 2015), available at <https://www.bmj.com/content/351/bmj.h4320>.

An applicant's promise to share Raw Data should be rated under either Additional Review Criteria or Additional Review Considerations.

II. The Substantive Requirements for Data Management and Sharing Plans

The NIH's current proposal is to require all NIH-supported research projects to include a plan for data management and sharing.

Unless the NIH adds *significant teeth* to the data management plan requirements, however, they will be circumvented or ignored as in the past.

The most important things the NIH could do as to these baseline requirements are:

- 1) Establish a firm default expectation that "scientific data" *must* be shared except where justified in advance to an expert panel, and the main purpose of a "plan" is to describe *how* (not whether) the sharing will occur;
- 2) NIH should develop a plan coordinated by Building One, in collaboration with each of the ICs, to create model data management plans in consultation with data science experts, and then it should add data management plans to the Scored Review Criteria as soon as the appropriate scoring standards can be developed for any given field.
- 3) Violations should be subject to clear sanctions, including findings of research misconduct with all the range of possible penalties available to the Office of Research Integrity.
- 4) NIH should make data management plans publicly available in machine-readable fashion, so that the availability of data can be publicly tracked.

There are two main reasons for the above requirements.

First, data sharing advances the NIH's mission of furthering scientific advancement. Sharing data enables other scientists to build upon previous work. As a *Science* editorial said, "Making data widely available is an essential element of scientific research."⁴ Sharing data has led to many scientific advances, particularly in genetics. Since 2007, NIH's Database of Genotypes and Phenotypes has allowed "2,221 investigators access to 304 studies, resulting in 924 publications and significant scientific advances."⁵ For example, a recent re-analysis of data made significant advances in our understanding of which genetic loci are associated with esophageal cancer.⁶

By contrast, failure to share data can halt scientific progress. For example, researchers tried to do a meta-analysis of techniques for treating newborn infants who have trouble

⁴ B. Hanson, A. Sugden, & B. Alberts, "Making Data Maximally Available," *Science* 331 (11 Feb. 2011): 649.

⁵ D. N. Paltoo et al., "Data use under the NIH GWAS Data Sharing Policy and future directions," *Nature Genetics* 46 (27 Aug. 2014): 934-38.

⁶ C. Wu et al., "Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations," *Nature Genetics* 46 (2014): 1001-06. Available at <http://www.nature.com/ng/journal/v46/n9/full/ng.3064.html>.

regulating their breathing, reflexes, etc., in the hopes of developing a prognostic tool for doctors to use. They found the meta-analysis impossible to carry out: over 60 percent of the data was unavailable because researchers either ignored the request or outright refused to share.⁷

Second, data sharing is essential to ensuring scientific reproducibility, which is of increasing concern. As Francis Collins and Larry Tabak have acknowledged, “the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring,” and “the recent evidence showing the irreproducibility of significant numbers of biomedical-research publications demands immediate and substantive action.”⁸

One of the best ways to improve reproducibility is to *require* the open sharing of data used to support scientific publications, while at least *rewarding* those who share the broader scientific workflow (as discussed above). When data are shared, other scientific investigators have the opportunity to double-check someone else’s analysis. Moreover, the original investigators will have a heightened incentive to analyze their data in a rigorous and defensible way if they foresee that the data could be re-examined by someone else.

The requirement to share data should be the default, and exceptions should be granted sparingly. While privacy and confidentiality are obviously of paramount importance for human subjects data, neither should those ideas be excessively used as an excuse for refusing to share data. For example, clinical trial data may be subject to the Health Insurance Portability and Accountability Act (HIPAA),⁹ but clinical trialists should be required to anonymize their data according to HIPAA standards, and then share the data under a confidentiality agreement just as multiple pharmaceutical companies have done.¹⁰

Some might object to a *broad* mandate for sharing data from NIH-funded research. After all, some sources or categories of data might be particularly cumbersome to share (particularly considering long-term preservation costs), and/or might be of little or no use to other investigators.

The best way to handle this objection is that even while NIH moves forward with a broad mandate, it should convene with ICs and with representatives of various scholarly communities to consider the exact nature and scope of what “data” has to be shared, and whether a limited exception to the data-sharing requirement is truly justifiable. Worth keeping in mind is that for

⁷ G. J. Jaspers & P. LJ Degraeuwe, “A failed attempt to conduct an individual patient data meta-analysis,” *Systematic Review* 3 (4 Sept. 2014): 97.

⁸ Collins and Tabak, footnote 3 above.

⁹ For a good discussion of the risks, see M. Mello et al., “Preparing for Responsible Sharing of Clinical Trial Data,” *New England Journal of Medicine* 369 (2013): 1651-1658, at <http://www.nejm.org/doi/full/10.1056/NEJMhle1309073>.

¹⁰ A. J. Vickers, “Whose data set is it anyway? Sharing raw data from randomized trials,” *Trials* 7 (2006): 15; M. A. Rodwin & J. D. Abramson, “Clinical Trial Data as a Public Good,” *JAMA* 308 (5 Sept. 2012): 871-72.

some types of data, no one may yet see the value of sharing *precisely* because no one has yet seen a systematically curated repository of all the data from that field.

Thus, rather than create a long string of exceptions in the initial rule, it would be better for the NIH to start with a universal mandate for broad data sharing to allow time for it to take root, and then grant exceptions in cases where a narrow category of data is provably of no scientific value. It's time to stop waffling on the open sharing of data created or collected with federal funds.

As a final note, the NIH says that it “encourages the use of established repositories for preserving and sharing scientific data.” This is a good step forward, but not far enough. If applicants are allowed, say, to “share” data merely by posting a PDF on Dropbox that could be deleted or lost in the near future, that would not be sufficient to protect the NIH's and the public's interest in seeing the greatest value come from publicly funded data. Long-term preservation is far more likely if the NIH specifies that sharing *must* occur via a trusted digital repository to the extent that one exists in a given field; if no such repository exists, the NIH could allow case-by-case exceptions where long-term preservation is guaranteed through some other mechanism, or it could consider creating and funding new repositories.

Deposited data should be locked, so that no one can later delete or modify it so as to frustrate the data sharing requirements. As well, there should be a default budgetary requirement that grants and contracts dedicate a portion of their resources to support the work of preparing data to be shared at a trusted digital repository, including a line item for the repository itself.

III. The Optimal Timing of New Requirements

The effective date for any new data sharing policy has yet to be determined. The NIH funds in so many areas (from cell biology, to clinical trials, to epidemiology) that in any discussion of timing, infrastructure, and standards, there is a risk of letting the general approach fall to the lowest common denominator.

Instead, the NIH should do the following: create a timeline and framework for implementation that looks roughly like this:

- For the types of research that already have a dedicated and trusted digital repository,¹¹ the new data sharing requirement goes into effect immediately.
- For all other types of research, the NIH will convene with the relevant ICs and representatives from individual scholarly communities to develop a plan for enforcing the data-sharing requirement as of one year of the rule's effective date.

With such an implementation framework in place, the NIH can move forward with a data-sharing requirement immediately for many areas of research, while still creating an impetus for developing standards and infrastructure in other areas as soon as possible.

¹¹This might include the NIH's official list of repositories at https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html, as well as others (perhaps including *Nature Scientific Data* at <https://www.nature.com/sdata/policies/repositories>).

IV. Code and Software

Code is an essential part of collecting, aggregating, cleaning, and analyzing data. As for code, the proposed rule merely says this: “An indication of whether specialized tools are needed to access or manipulate shared data to support replication or reuse, and name(s) of the needed tool(s) and software. Consider specifying how needed tools can be accessed, (i.e., open source and freely available, generally available for a fee in the marketplace, or available only from the research team or some other source).”

This is inadequate. We suggest that NIH could strengthen this requirement considerably as follows:

- When specialized software or code is *itself* developed with NIH funding, the NIH should require such software or code to be free and open source. For example, if someone uses NIH funding to create a new biostatistics package to handle high-throughput sequencing analyses, that package should be made freely available to the public. There is no reason for people to develop software with public funds but then keep it to themselves.
- When anyone uses software (including non-open software such as MATLAB or Stata) to clean and analyze data, the script(s) should be shared along with the data, so that anyone else can replicate the analysis. Even prestigious scholars have been tripped up by coding errors,¹² and making code available allows independent researchers the chance to exercise oversight over poorly written code.

With such requirements in place, the value of NIH-funded research will be increased.

¹² For one prominent example, see G. Miller et al, “A Scientist’s Nightmare: Software Problem Leads to Five Retractions,” *Science* 314 (22 Dec. 2006): 1856-57.

Submission ID: 1355

Date: 1/10/2020

Name: Janel Fedler

Name of Organization: University of Iowa

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Neurology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

To assist investigators in writing a Data Management and Sharing Plan, it would be helpful if the FOA or RFA specified which repository the data should be shared with so that the Plan meets the requirements of the repository and institute.

Section II: Definitions:

Metadata, an essential part of data sharing, can be provided in several different manners and levels to facilitate an outside user understanding the data. We ask for clarification of what would satisfy the metadata requirement. While we acknowledge standards have been developed for metadata such as CDISC, the expertise to formalize metadata according to such standards may be lacking by some investigators. Thus, support to meet such standards would need to be given. Adequate education of preparing documentation is critical so data are not mishandled.

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

The draft policy states, "Plans should also identify strategies or approaches to ensure data security and compliance with privacy protections are in place throughout the life of the scientific data." Once data is transferred to a designated repository, it is assumed the repository would take responsibility of data security. The applicant may have limited knowledge of the repositories data security policies and procedures.

While complete de-identification of data can never be achieved, repository guidance would mostly determine the extent of de-identification, i.e. repository requirements, open or limited access to data, etc.

Section VII: Compliance and Enforcement:

Beyond penalties, positive re-enforcement could be used to motivate data sharing. For example, better linkage in journals for publications using shared data, a website link to the data repository on ct.gov, encouraging collaboration with the original PI, etc.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The size of the study, complexity, amount of data provided by outside vendors, personnel expertise, degree of standards applied, and other factors would all play a part in estimating the costs to preparing data for sharing.

Often the data for a study is prepared and shared after the grant term ends. It is essential that grant terms be extended to account for this.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Ideally, the data sharing plan would be in alignment with the investigator's publication plan to share in a timely manner. Recognizing an investigator should not sequester data, however, he or she should have priority to analyze the data. Often the same resources would be split to accomplish these tasks.

Other Considerations Relevant to this DRAFT Policy Proposal:

We recommend the NIH clarify how the data sharing requirements would be communicated to and/or negotiated with industry partners

Consideration must be made to how de-identified scientific data would be linked to specimens.

We applaud the NIH for instigating this policy to encourage the sharing of scientific data collected; however, the many data repositories each with their own standards creates confusion among researchers and has been a barrier to sharing in the past. An effort to limit the number of repositories, or minimally standardize their requirements, would be in the best interest of the entire scientific community.

Attachment:

Description:

Submission ID: 1356

Date: 1/10/2020

Name: Ian Moss

Name of Organization: International Association of Scientific, Technical, and Medical Publishers (STM)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of the above

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other:

Domain of Research Most Important to You or Your Organization:

All areas

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

STM supports the intent of NIH to promote data sharing in order to improve reproducibility and transparency, as well as to improve analysis and enable further discovery. We and our members share NIH's commitment to good data management practices, rooted in community practices and widely-accepted standards. We look forward to working with NIH and the research community to help enable data management and sharing, particularly through our STM 2020 Research Data Year initiative.

Section II: Definitions:

As noted in our response to the "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research," STM believes that the definition of data needs to distinguish between data itself and the various interpretations and presentations of data. We do not believe that the definition of "scientific data" in the draft policy is explicit enough. While we appreciate that the definition has been modified to focus on validation and the replication of research findings, the exclusions are not clear enough, or extensive enough to exclude analyses or creative presentations of such data. We therefore recommend that the exclusions be expanded to read: "Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts or final

versions of scientific papers, plans for future research, peer reviews, communications with colleagues, visualizations, or physical objects, such as laboratory specimens."

In addition, we recommend that several other terms be defined to increase the potential implementation and effectiveness of the policy and to help researchers understand and implement steps within their Data Management and Sharing Plans. Specifically, we offer the following terms and potential definitions for consideration:

- **Data Availability Statement:** a statement, often published within an article, that indicates what data are available and how to access them
- **Data Citation:** a reference to the available and relevant data in the reference list, including, where available, a link to the resource
- **Persistent Identifiers (PIDs):** Persistent identifiers assigned to digital objects, such as data sets, in order to make them Findable, Accessible, Interoperable, and Re-usable (FAIR).
- **Data linking:** using PIDs to data sets to create links between articles and datasets, or between datasets and datasets and other related research outputs and artifacts.
- **Data publishing:** making data publicly available and linked to curated information related to the research; especially when an article reporting on research is published, making related data sets available alongside the article, via deposits in trustworthy repositories, making them FAIR by means of linking via persistent identifiers, the inclusion of a Data Availability Statement in the article and proper citation in the reference list.

Section III: Scope:

While STM supports better data management and sharing across the research ecosystem, we note that the implementation and extent of policy requirements may vary by the type of funding mechanism. Care must be taken to differentiate between requirements for contract work as opposed to researcher-led grant projects. This is in addition to – and distinct from – the already acknowledged differences between different research disciplines and communities.

Section IV: Effective Date(s):

STM appreciates that the implementation of the Policy will be dependent upon the willingness and ability of research communities to embrace FAIR principles and work towards greater data sharing. We are engaged in efforts to support both the infrastructure and cultural changes necessary to effect better data management and sharing and look forward to collaborating with NIH and others to accelerate the needed changes. We would welcome additional dialogue on how we can work together to achieve our shared goals.

In this section, and elsewhere in the document, there is a reference to "other funding agreements." In the context of the scope of the Policy, it would appear that this means "other funding agreements with NIH," and it would be helpful if that was so clarified here.

Section V: Requirements:

STM agrees that researchers should have a plan for managing and sharing data, as appropriate. We also appreciate the flexibility intrinsic in the Policy requirements, including the recognition that there may be restrictions or limitations on sharing.

Section VI: Data Management and Sharing Plans:

Good data management and sharing requires planning for long-term preservation and ensuring that data is FAIR, including through the creation of PIDs, proper data citation, and linking between articles and data sets. Although some of these issues are addressed in the "Supplemental DRAFT Guidance," they are central enough that it would be valuable to include them in the first paragraph of this section, just as data security has been mentioned. For example, where the Policy says "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public," it could add", and Plans should identify mechanisms for long-term preservation, where appropriate." The first paragraph should also note "Plans should explain how researchers will maximize the discoverability of shared data, through the creation of PIDs, citation, linking, and the like."

Section VII: Compliance and Enforcement:

As noted earlier, STM publishers are working to assist researchers in making data sharing a conscious part of their efforts to communicate the results of their research. As appropriate to the diversity of research communities they serve, journals have a variety of approaches to support data management and sharing. These include creating an explicit data policy, encouraging or requiring data availability statements, ensuring PIDs for shared data sets, providing proper data citation guidelines, and creating standard processes to link articles and data sets. All of these efforts would help to support implementation of and compliance with any commitments in a Plan.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The costs, both direct and indirect, of good data management and sharing practices can be considerable. STM applauds NIH's recognition that these costs may not be captured in current research practices and its explicit recognition that researchers will need to consider these additional costs in their budgets. We especially appreciate that the document acknowledges that data management and sharing may have ongoing costs and that long-term preservation costs need to be considered.

In addition to the costs listed in this guidance document, STM encourages NIH to explicitly note the costs of the assignment of a persistent identifier, whether directly incurred or as part of the assessment of an appropriate repository. This could be achieved by adding to section 1: "ensuring that the data is deposited at a selected repository that will assign a Persistent Identifier (PID) to each data set, which is endorsed by the research community (DOI's or Accession numbers, etc)."

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

STM's members publish in a wide variety of research areas, each of which has different practices with respect to data collection, use and sharing. The plan requirements must be flexible enough to support the diverse nature of the research that NIH funds, while also providing guidance to all researchers to encourage and enable sharing. With the diversity of data practices and differences in the intensity of data usage across different fields, it may not be appropriate to limit data management plans to two pages in all cases. NIH may want to consider providing the limit as a guideline, or adjusting it in the case of multi-institutional or more complex data plans.

We appreciate that the plan elements described in the guidance are generally flexible and open to interpretation by researchers to best suit their project, consistent with their research community standards. In particular, guidance on the data types and on related tools, software, and code provide appropriate openness to be adaptable to a variety of settings. At the same time, understanding the work involved in the development of good data management and sharing plans and practices, it might be helpful to provide additional guidance to researchers on repositories and practices, perhaps even with reference to a template for the creation of a plan or by providing examples.

We also appreciate the call for the use of standards that are community-endorsed, compatible and interoperable. In fact, this could even be strengthened to clarify that NIH is encouraging the use of standards that meet the two bulleted criteria.

To strengthen the guidance on data preservation and access we recommend several improvements:

- NIH may want to provide guidance to researchers on the criteria for an appropriate and trusted location for data, including plans for perpetual access and a commitment to the FAIR Data principles. Several initiatives offer certification for or recommendations of trusted data repositories, including CoreTrustSeal (<https://www.coretrustseal.org/>) and Repository Finder (<https://repositoryfinder.datacite.org/about>; <https://www.re3data.org/>).

- NIH may want to strengthen the recommendation on PIDs, which are a widely accepted best practice in data sharing and critical to implementing the FAIR Data principles. Rather than considering "whether a persistent unique identifier or other standard indexing tools will be used," researchers should consider "which persistent unique identifier or other standard indexing tools will be used.
- NIH may want to consider adding an additional consideration in support of data publishing. For example, "when and how the data will be made part of any article that reports on funded research; indicating whether journals will be considered that have data policies, entertain data availability statements and make data citations part of the reference lists, as well as linking from the published article to the Persistent Identifiers of the deposited data sets in trustworthy repositories."

Finally, the guidance does not directly address researcher rights to the commercialization of data, which is a key incentive in the research enterprise. NIH may want to consider the use of language that currently appears in the Cancer Moonshot and HEAL Initiative Public Access and Data Sharing policies which explicitly allow researchers to use "licenses that retain intellectual property for commercialization."

Other Considerations Relevant to this DRAFT Policy Proposal:

One of the most significant challenges to better data management and sharing is the current lack of understanding in the research communities we serve of the benefits and motivations for data sharing. In an environment where researchers are under increasing pressure and have limited resources, any additional cost or effort needs to be motivated and aligned with incentives. Here, it can be helpful to note that it is increasingly clear that those that share data have a greater impact. Research has shown that publications with links to shared data receive more citations (see, e.g., Colavizza, G. et al. "The citation advantage of linking publications to research data." ArXiv abs/1907.02565 (2019) <https://arxiv.org/pdf/1907.02565.pdf>).

Publishers stand ready to work with NIH and others to help researchers realize the increased impact for funded research and articles that report on that research. Further, publishers are in a unique position to help drive that change. Studies by a number of publishers that have introduced data policies and data availability statements have shown that such policies have significant impact. For example, when PLOS and BioMed Central introduced Data Availability Statement requirements, authors immediately responded with significant increases in sharing of datasets (see Fig. 2 of Colavizza, G. et al. "The citation advantage of linking publications to research data." ArXiv abs/1907.02565 (2019) <https://arxiv.org/pdf/1907.02565.pdf>). Similarly, Elsevier has reported that the percentage of articles that carry links to deposited data sets increased from about 7% to more than 20 % in the 3 years since it implemented data

availability statements. Publishers are eager to support funders in our common mission to share more data.

STM is already contributing significantly to the development of the standards, resources, policies, and infrastructure needed to enable robust data sharing across the research community, through its involvement in the Research Data Alliance, our own STM 2020 Research Data Year, and other initiatives. We welcome further discussion on how NIH, STM, and our member publishers can work together to build greater trust in science and promote the use of research data for the benefit of research and the public. Please feel free to contact me or David Weinreich, Director of Public Affairs in the Americas, for further information.

Attachment:

STM Response to Request for Information on NIH draft data management and sharing policy.pdf

Description:

Letter, including introduction and content of submission

10 January 2020

Response to Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

The International Association of Scientific, Technical and Medical Publishers (STM) is the leading global trade association for academic and professional publishers. It has more than 150 members in 21 countries who each year collectively publish more than 66% of all journal articles and tens of thousands of monographs and reference works. STM members include non-profit scientific and scholarly societies, commercial publishers, and university presses who work collectively to ensure broad access to and use of the latest scientific and scholarly information. The majority of our members are small businesses and not-for-profit organizations, who represent tens of thousands of publishing employees, editors and authors, and other professionals across the United States and world who regularly contribute to the advancement of science, learning, culture and innovation throughout the nation. They comprise the bulk of a \$25 billion publishing industry that contributes significantly to the U.S. economy and enhances the U.S. balance of trade.

Publishers sit at the interface between researchers, their research and the rest of the world through our work to improve the quality and availability of information related to research. STM shares our members' commitment to supporting researchers in the sharing, discoverability, and reuse of research data. Individual publishers are developing tools and services to support researchers to make their data FAIR (Findable, Accessible, Interoperable, and Re-usable), and have actively responded to community demand for citation principles for data. STM itself has been involved in numerous projects looking at data access, citation, and preservation, the most recent examples of which have been our recently announced 2020 STM Research Data Year and our ongoing support for the development of [SCHOLIX](#), an easy and universal linking mechanism between scholarly publications and research data.

We therefore welcome the opportunity to comment on the "DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance," published on November 6, 2019, and offer the following as response to the "Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance" (NOT-OD-20-013 / 84 FR 60398). We wish to reiterate our interest – consistent with our commitment to promote sustainable Open Science – in ongoing dialogue with NIH on how to best to promote openness and sharing in research communication. We hope that we can engage further with NIH's Office of Science Policy on these issues over the coming year. Our submission builds on responses that STM has submitted to previous NIH RFIs

on research data and digital repositories, as well as responses that STM has submitted to previous government-wide RFCs on the Federal Data Strategy.

DRAFT NIH Policy for Data Management and Sharing

Section I. Purpose

STM supports the intent of NIH to promote data sharing in order to improve reproducibility and transparency, as well as to improve analysis and enable further discovery. We and our members share NIH's commitment to good data management practices, rooted in community practices and widely-accepted standards. We look forward to working with NIH and the research community to help enable data management and sharing, particularly through our [STM 2020 Research Data Year](#) initiative.

Section II. Definitions

As noted in [our response](#) to the "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research," STM believes that the definition of data needs to distinguish between data itself and the various interpretations and presentations of data. We do not believe that the definition of "scientific data" in the draft policy is explicit enough. While we appreciate that the definition has been modified to focus on validation and the replication of research findings, the exclusions are not clear enough, or extensive enough to exclude analyses or creative presentations of such data. We therefore recommend that the exclusions be expanded to read: "Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts [or final versions](#) of scientific papers, plans for future research, peer reviews, communications with colleagues, [visualizations](#), or physical objects, such as laboratory specimens."

In addition, we recommend that several other terms be defined to increase the potential implementation and effectiveness of the policy and to help researchers understand and implement steps within their Data Management and Sharing Plans. Specifically, we offer the following terms and potential definitions for consideration:

- **Data Availability Statement:** a statement, often published within an article, that indicates what data are available and how to access them
- **Data Citation:** a reference to the available and relevant data in the reference list, including, where available, a link to the resource
- **Persistent Identifiers (PIDs):** Persistent identifiers assigned to digital objects, such as data sets, in order to make them Findable, Accessible, Interoperable, and Re-usable (FAIR).
- **Data linking:** using PIDs to data sets to create links between articles and datasets, or between datasets and datasets and other related research outputs and artifacts.
- **Data publishing:** making data publicly available and linked to curated information related to the research; especially when an article reporting on research is published, making related data sets available alongside the article, via deposits in trustworthy repositories, making them FAIR by means of linking via persistent identifiers, the inclusion of a Data Availability Statement in the article and proper citation in the reference list.

Section III: Scope

While STM supports better data management and sharing across the research ecosystem, we note that the implementation and extent of policy requirements may vary by the type of funding mechanism. Care must be taken to differentiate between requirements for contract work as opposed to researcher-led grant projects. This is in addition to – and distinct from – the already acknowledged differences between different research disciplines and communities.

Section IV: Effective Dates

STM appreciates that the implementation of the Policy will be dependent upon the willingness and ability of research communities to embrace FAIR principles and work towards greater data sharing. We are engaged in efforts to support both the infrastructure and cultural changes necessary to effect better data management and sharing and look forward to collaborating with NIH and others to accelerate the needed changes. We would welcome additional dialogue on how we can work together to achieve our shared goals.

In this section, and elsewhere in the document, there is a reference to “other funding agreements.” In the context of the scope of the Policy, it would appear that this means “other funding agreements with NIH,” and it would be helpful if that was so clarified here.

Section V: Requirements

STM agrees that researchers should have a plan for managing and sharing data, as appropriate. We also appreciate the flexibility intrinsic in the Policy requirements, including the recognition that there may be restrictions or limitations on sharing.

Section VI: Data Management and Sharing Plans

Good data management and sharing requires planning for long-term preservation and ensuring that data is FAIR, including through the creation of PIDs, proper data citation, and linking between articles and data sets. Although some of these issues are addressed in the “Supplemental DRAFT Guidance,” they are central enough that it would be valuable to include them in the first paragraph of this section, just as data security has been mentioned. For example, where the Policy says “NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public,” it could add “, and Plans should identify mechanisms for long-term preservation, where appropriate.” The first paragraph should also note “Plans should explain how researchers will maximize the discoverability of shared data, through the creation of PIDs, citation, linking, and the like.”

Section VII: Compliance and Enforcement

As noted earlier, STM publishers are working to assist researchers in making data sharing a conscious part of their efforts to communicate the results of their research. As appropriate to the diversity of research communities they serve, journals have a variety of approaches to support data management and sharing. These include creating an explicit data policy, encouraging or requiring data availability statements, ensuring PIDs for shared data sets, providing proper data citation guidelines, and creating

standard processes to link articles and data sets. All of these efforts would help to support implementation of and compliance with any commitments in a Plan.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

The costs, both direct and indirect, of good data management and sharing practices can be considerable. STM applauds NIH's recognition that these costs may not be captured in current research practices and its explicit recognition that researchers will need to consider these additional costs in their budgets. We especially appreciate that the document acknowledges that data management and sharing may have ongoing costs and that long-term preservation costs need to be considered.

In addition to the costs listed in this guidance document, STM encourages NIH to explicitly note the costs of the assignment of a persistent identifier, whether directly incurred or as part of the assessment of an appropriate repository. This could be achieved by adding to section 1: "ensuring that the data is deposited at a selected repository that will assign a Persistent Identifier (PID) to each data set, which is endorsed by the research community (DOI's or Accession numbers, etc)."

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)

STM's members publish in a wide variety of research areas, each of which has different practices with respect to data collection, use and sharing. The plan requirements must be flexible enough to support the diverse nature of the research that NIH funds, while also providing guidance to all researchers to encourage and enable sharing. With the diversity of data practices and differences in the intensity of data usage across different fields, it may not be appropriate to limit data management plans to two pages in all cases. NIH may want to consider providing the limit as a guideline, or adjusting it in the case of multi-institutional or more complex data plans.

We appreciate that the plan elements described in the guidance are generally flexible and open to interpretation by researchers to best suit their project, consistent with their research community standards. In particular, guidance on the data types and on related tools, software, and code provide appropriate openness to be adaptable to a variety of settings. At the same time, understanding the work involved in the development of good data management and sharing plans and practices, it might be helpful to provide additional guidance to researchers on repositories and practices, perhaps even with reference to a template for the creation of a plan or by providing examples.

We also appreciate the call for the use of standards that are community-endorsed, compatible and interoperable. In fact, this could even be strengthened to clarify that NIH is encouraging the use of standards that meet the two bulleted criteria.

To strengthen the guidance on data preservation and access we recommend several improvements:

- NIH may want to provide guidance to researchers on the criteria for an appropriate and trusted location for data, including plans for perpetual access and a commitment to the FAIR Data

principles. Several initiatives offer certification for or recommendations of trusted data repositories, including CoreTrustSeal (<https://www.coretrustseal.org/>) and Repository Finder (<https://repositoryfinder.datacite.org/about>; <https://www.re3data.org/>).

- NIH may want to strengthen the recommendation on PIDs, which are a widely accepted best practice in data sharing and critical to implementing the FAIR Data principles. Rather than considering “whether a persistent unique identifier or other standard indexing tools will be used,” researchers should consider “*which* persistent unique identifier or other standard indexing tools will be used.
- NIH may want to consider adding an additional consideration in support of data publishing. For example, “when and how the data will be made part of any article that reports on funded research; indicating whether journals will be considered that have data policies, entertain data availability statements and make data citations part of the reference lists, as well as linking from the published article to the Persistent Identifiers of the deposited data sets in trustworthy repositories.”

Finally, the guidance does not directly address researcher rights to the commercialization of data, which is a key incentive in the research enterprise. NIH may want to consider the use of language that currently appears in the Cancer Moonshot and HEAL Initiative Public Access and Data Sharing policies which explicitly allow researchers to use “licenses that retain intellectual property for commercialization.”

Other Considerations Relevant to this DRAFT Policy Proposal

One of the most significant challenges to better data management and sharing is the current lack of understanding in the research communities we serve of the benefits and motivations for data sharing. In an environment where researchers are under increasing pressure and have limited resources, any additional cost or effort needs to be motivated and aligned with incentives. Here, it can be helpful to note that it is increasingly clear that those that share data have a greater impact. Research has shown that publications with links to shared data receive more citations (see, e.g., Colavizza, G. et al. “The citation advantage of linking publications to research data.” *ArXiv* abs/1907.02565 (2019) <https://arxiv.org/pdf/1907.02565.pdf>).

Publishers stand ready to work with NIH and others to help researchers realize the increased impact for funded research and articles that report on that research. Further, publishers are in a unique position to help drive that change. Studies by a number of publishers that have introduced data policies and data availability statements have shown that such policies have significant impact. For example, when PLOS and BioMed Central introduced Data Availability Statement requirements, authors immediately responded with significant increases in sharing of datasets (see Fig. 2 of Colavizza, G. et al. “The citation advantage of linking publications to research data.” *ArXiv* abs/1907.02565 (2019) <https://arxiv.org/pdf/1907.02565.pdf>). Similarly, Elsevier has reported that the percentage of articles that carry links to deposited data sets increased from about 7% to more than 20 % in the 3 years since it

implemented data availability statements. Publishers are eager to support funders in our common mission to share more data.

STM is already contributing significantly to the development of the standards, resources, policies, and infrastructure needed to enable robust data sharing across the research community, through its involvement in the Research Data Alliance, our own STM [2020 Research Data Year](#), and other initiatives. We welcome further discussion on how NIH, STM, and our member publishers can work together to build greater trust in science and promote the use of research data for the benefit of research and the public. Please feel free to contact me or David Weinreich, Director of Public Affairs in the Americas, for further information.

Very truly yours,

Ian Moss
CEO

Submission ID: 1357

Date: 1/10/2020

Name: Jennifer Doty

Name of Organization: Emory University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All types

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Librarian

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

No comments on this section.

Section II: Definitions:

Scientific Data: "NIH expects that reasonable efforts will be made to digitize all scientific data."
(p. 2)

- We recommend clarification of what the NIH considers to be "reasonable efforts" and examples of the types of analog data that would benefit from digitization, given that laboratory notebooks and physical objects are not included in the draft policy's definition of Scientific Data.

Section III: Scope:

"This Policy applies to all research, funded or conducted in whole or in part by NIH..." (p. 2)

- We recommend including language to address how this policy affects projects funded by additional sponsors, whether federal agencies or private foundations. Will the NIH policy take precedence? Should the NIH ICO be consulted for guidance?

Section IV: Effective Date(s):

No comments on this section.

Section V: Requirements:

No comments on this section.

Section VI: Data Management and Sharing Plans:

"Researchers ... are required to submit a Plan to the funding NIH ICO as part of Just-in-Time for extramural awards..." (p. 3)

- If Plans are part of Just-in-Time submissions rather than the competitive review process, we are concerned that researchers' peers in the scientific community will not have an opportunity to provide input on the proposal's anticipated data management and sharing practices. This is different from how other funders (e.g. NSF, Gates Foundation) include data management plans in the merit review of proposals. We recommend that NIH reconsider this requirement and align their submission practice with other funders.

"NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." (p. 3)

- Who will determine the utility of scientific data in the long-term? We recommend that NIH include specific guidance on how data are to be deemed useful, or to consider removing this sentence.

"NIH may make Plans publicly available." (p. 3)

- If this is a reference to developing standards to make data management plans machine-actionable (<https://doi.org/10.1371/journal.pcbi.1006750>) and therefore more open for both human and machine consumption, we recommend that NIH make that explicit.

"NIH encourages the use of established repositories for preserving and sharing scientific data." (p. 3)

- We recommend more specificity about the criteria used to consider whether any repository is "established." This would also be a good place to reference the NLM-maintained list of NIH Data Sharing Repositories:
https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

"Plan Elements: Consider addressing specific elements outlined in Supplemental DRAFT Guidance..." (p. 3)

- We recommend removing the word "consider" so it is clear that all elements of the Plan should be addressed.

"Extramural Awards: Plans will undergo a programmatic assessment by NIH staff within the proposed funding NIH ICO." (p. 3)

- Like our first comment in this section, we are concerned that researchers' peers in the scientific community will not have an opportunity to provide input on the proposal's anticipated data management and sharing practices. We reiterate our recommendation that NIH reconsider this requirement and align their submission practice with other funders.

Section VII: Compliance and Enforcement:

"During the funding period, compliance with the Plan will be determined by the funding NIH ICO." (p. 4)

- It is good that this aligns somewhat with the NIH Public Access Policy, but is currently lacking specific information about when the NIH would expect data to be shared. In the case of articles, the full text must be shared at the time of publication. Any NIH requirement to share data within a certain timeframe could result in a high volume of investigators seeking help from an institution's IT services and/or library during reporting intervals to deposit data quickly and get their funds released. Depositing data in a repository is arguably more complicated and time-consuming than depositing articles with PMC.

"After the end of the funding period, non-compliance with the NIH ICO-approved Plan may be taken into account by the funding NIH ICO for future funding decisions for the recipient institution..." (p. 4)

- This is also different from the current NIH Public Access Policy, where non-compliance does not impact future funding decisions. This draft policy language could be read to mean that a given institution would not receive any NIH funds when just one investigator is non-compliant.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We are glad to see that NIH is explicitly stating that there are allowable costs to make data accessible in established repositories and to prepare and curate data for reproducible research.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We are glad to see that the two-page limit is consistent with other funders' plan requirements.

1. Data Type:

"Providing a rationale for decisions about which scientific data are to be preserved and made available for sharing, taking into consideration scientific utility, validation of results, availability of suitable data repositories, privacy and confidentiality, cost, consistency with community practices, and data security." (p. 1)

- We recommend moving this to section 4, Data Preservation, Access, and Associated Timelines.

4. Data Preservation, Access, and Associated Timelines:

"If scientific data will be archived in an existing data repository(ies), consider providing the name and URL web address of the repository(ies). If an existing data repository(ies) will not be used, consider indicating why not and how scientific data will be preserved and shared." (p. 2)

- We recommend more specificity about the criteria used to consider whether any repository is "established." This would also be a good place to reference the NLM-maintained list of NIH Data Sharing Repositories:

https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

- We also recommend removing both instances of the word "consider" from these sentences so it is clear that details should be provided.

"Anticipated timeframes for preserving scientific data..." (p. 3)

- We recommend including data sharing at the time of publications as one of the anticipated timeframes.

6. Oversight of Data Management

- We recommend moving this section earlier so that it's clear that responsibility for data stewardship is valued by the NIH.
- We recommend considering asking investigators to provide a list of key personnel and the data management training they have completed.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1358

Date: 1/10/2020

Name: Jerry Blancato

Name of Organization: EPA/ORD

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Research

Type of Organization: Government Agency

Type of Organization - Other:

Role: Government Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

1. Scope. What is the limit of the term, 'scientific data' (e.g. raw, summarized, model input/output)? We know what it isn't (see definition) but that leaves the remainder of things that qualify as scientific data extremely broad.

"Scientific Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. NIH expects that reasonable efforts will be made to digitize all scientific data." Page 1.

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

2. Data comes in various volumes and formats (e.g. some are created by specialized equipment in proprietary format) these proprietary datasets may require specialized

equipment or software to be used and interpreted and may not fit easily into the Findable, Accessible, Interoperable, and Reusable (FAIR) principle.

"This Policy applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data. This includes research funded or conducted by extramural grants, contracts, intramural research projects, or other funding agreements regardless of NIH funding level or funding mechanism." Page 2.

3. Does data ever expire or is it intended to be maintained publicly available into perpetuity? I think a term of guaranteed access to the data should be established in order to contain costs (e.g. 5 year, 10 years or 20 years). Otherwise, NIH is paying for maintenance of public facing data into perpetuity regardless of its significance, relevance, impact or use.

"Costs associated with data management and data sharing may be allowable under the budget for the proposed project (Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing)." Page 2.

4. Who deems scientific data useful (e.g. data owner, NIH or public)?

"NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." Page 3.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1359

Date: 1/10/2020

Name: Jeffery Smith

Name of Organization: AMIA

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: ALL DATA

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

All clinical, translational and biomedical research; health services research; and epidemiology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Recommendation I.a: Add Language to the Purpose Section

The draft DMSP should bolster the Purpose section by adding language similar to the introductory language, beginning, "NIH has a longstanding commitment to making the results and accomplishments of the research that it funds and conducts available to the public. Increasing access to scientific data resulting from NIH funding or support offers many benefits and reflects NIH's responsibility to maintain stewardship over taxpayer funds." AMIA recommends the draft DMSP adds to this with the following:

"Specifically, systematic management and sharing of scientific data and results enables researchers to more vigorously test the validity of research findings, strengthen analyses by combining data sets, access hard-to-generate data, and explore new frontiers. Data management and sharing also informs future research pathways, increases the return on investment of scientific research funding, and accelerates the translation of research results into knowledge, products, and procedures to improve health and prevent disease.

This Policy seeks to identify, adopt, and credit data management and sharing best practices, consistent with FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles, so that the United States remains the leader in biomedical and life sciences research. This Policy establishes the requirements and responsibilities of researchers generating scientific data resulting from NIH-funded or -supported research and it will govern development and implementation of other NIH Policies related to the management and sharing of scientific data, such as the NIH Genomic Data Sharing Policy, the NIH Policy on the Dissemination of NIH-funded Clinical Trial Information, and the Intramural Research Program Human Data Sharing (HDS) Policy."

Section II: Definitions:

Recommendation II.a: Amend the Definition of Data Management and Sharing Plan.

AMIA recommends the following amendments to the Plan's definitions to acknowledge differences in data management and data sharing. Further AMIA recommends the draft DMSP remove all references to "(e.g. researchers and the broader public)" when describing potential users of scientific data:

"A plan describing how scientific data will be generated, managed, described, analyzed, preserved, shared, and made accessible to others for supplemental uses, as appropriate. This plan should include two distinct sections describing how scientific data will be managed across the life-cycle of the project and how scientific data will be shared at the project close, or at another appropriate interval(s)."

Recommendation II.b: Replace the Definition of Data Management.

As discussed above, the DMSP should explicitly describe what is necessary to manage data, not just share data, given that data management and data sharing are distinct. Data management is prerequisite for data sharing, ensuring that the data are accurate, complete, and maintained in a standardized manner. Without effective data management, you cannot have effective data sharing, thus we recommend the DMSP consider additional Plan Elements as described in that section of our comments. Given this view, we recommend the draft DMSP include a new definition for data management as follows:

"The upstream management of scientific data that documents actions taken in making research observations, collecting research data, describing data (including relationships between datasets), processing data into intermediate forms as necessary for analysis, integrating distinct datasets, and creating metadata descriptions. Specifically, those actions that would likely have impact on the quality of data analyzed, published, or shared."

Recommendation II.c: Amend the Definition of Data Sharing

Amend the definition of data sharing to the following:

"Making scientific data accessible for use by others in a manner that is consistent with the FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles."

Recommendation II.d: Refine the definition of Metadata

We found the definition for metadata in need of refinement. Specifically, the phrase "additional information to make data more usable" implies that a data set could be usable at all without metadata, which is simply not the case. There is no data that can be correctly understood, much less re-used, without at least a data dictionary with field definitions and data types. Further, we view "Outcome measures" as actual data, not metadata. There may be metadata that defines how an outcome measure was derived, but the outcome data itself is not metadata. Given this view, we recommend the draft DMSP amend the definition of metadata as follows:

"Metadata is descriptive information about data, including variable/document definition/description, data type, and other characteristics. Areas discussed in metadata include, but are not limited to, instruments used to collect data; parameters or settings for such instruments; descriptors of physical samples from which data were collected; dates and times of data collection; any transformations applied to the data; relationships between datasets; provenance linking derived or modified datasets to original sources; phenotypic descriptors of data sources; and institutional/personal identifying information associated with the group or person(s) responsible for the data. Metadata also help establish (confidence in) the credibility of the data."

Recommendation II.e: Amend the Definition of Scientific Data

We support the concept of "Scientific Data," but do not support a definition of this concept through negation. The listing of what Scientific Data is not may serve better as part of ancillary materials published by the NIH, such as Frequently Asked Questions, rather than be included in a definition. Further, it is odd to place a command, "NIH expects..." into a definition. Given this view, we recommend the draft DMSP include a new definition for Scientific Data as follows:

"Information that is gathered, derived, or generated in the course of conducting research. It is the basis for reaching conclusions and inferences based on scientific principles and methodologies. Scientific data can be used to test existing hypotheses, to generate new hypotheses for future research, to validate or replicate prior research as well as for more exploratory purposes. Scientific data represent the foundation for both scientific theories and publications."

Recommendation II.f: Add a Definition of "Scientific Software Artifacts"

AMIA recommends the draft DMSP includes a definition for "Scientific Software Artifacts," so that grantees clearly understand that both data and software tools created with NIH funds should be included as part of their data management and sharing plan. This definition would be limited to artifacts created with NIH funds, and omit proprietary software tools used to conduct research, such as a stat package. We recommend a definition such as:

"Software, code, analytic programs, and other knowledge artifacts developed to conduct research or resulting from the conduct of research."

Recommendation II.g: Add a Definition of "Covered Data"

AMIA recommends the draft DMSP includes a definition for "Covered Data," so that grantees clearly understand which data must be included as part of their data management and sharing plan. We recommend a definition such as:

"Those newly generated or derived Scientific Data used to conduct NIH-funded or -supported research and subject to this Policy. Such data may or may not be proprietary or subject to various access controls."

Recommendation II.h: Add a Definition of "Covered Period"

We recommend the NIH address these and other questions by incorporating a concept of "Covered Period." This term would facilitate greater understanding of the obligations of grantees

"The period of time for which the Scientific Data is expected to be maintained by the grantee and for which it is to be made available to others."

Section III: Scope:

Recommendation III.a: Include Scientific Software Artifacts as an Asset to Managed/Shared.

We urge the NIH to proceed with the proposed DMSP scope, ensuring that the policy requirements are constructed in a way that both small and large awardees can comply. While we agree that it is important for all NIH research to be subject to this policy, regardless of funding or mechanism, the policy must maintain flexibility to accommodate individual ICs and individual project characteristics.

AMIA recommends the NIH draft this section as "III. Scope" and position the aspects of the current provisions related to "requirements" in the next section, "IV Requirements for Data Management and Sharing Plans." The draft DMSP could expand on the rationale for its scope, similar to the Purpose section. We discuss issues related to IC-specific requirements and "reasonable costs," for data management and sharing below.

Section IV: Effective Date(s):

Recommendation IV.a: Establish a phased implementation timeline, beginning with grant awards above \$500,000 per year, six months after finalization of the NIH DMSP.

Recognizing the need to have all NIH-funded research comply with this DMSP, and with appreciation of what AMIA sees as necessary components of a data management and sharing

plan, we recommend a phased compliance timeline based on funding levels. This phased implementation would only apply to new research funded after the DMSP is final. First, new research funded above \$500,000 per year and subject to the existing data sharing policy should comply with the final DMSP within one year of its adoption. Second, new research funded above \$250,000 per year should comply with the provisions of the DMSP within 2 years of its adoption, and finally, all grants funded below \$250,000 per year should comply with the DMSP within 3 years of adoption. This compliance approach would focus efforts on those grants that already must comply with the existing policy and likely have the richest cache of scientific data, while giving smaller projects more time to become familiar with the DMSP.

This implementation strategy and timeline should guide all ICO-specific requirements and apply equally to intramural, extramural, and other funding agreements.

Section V: Requirements:

Recommendation V.a: Incorporate Supplemental Guidance documents on Plan Elements and Allowable Costs (as amended by our recommendations) into this section on Requirements.

Perhaps the most disappointing aspect of this proposal is the idea that the NIH could adequately coordinate data management and sharing of scientific data across its 27 ICOs with two simple requirements: (1) submit a Plan and (2) comply with ICO-specific requirements (if any). This strategy will lead to wide variation in requirements, implementation expectations, and researcher experiences in managing and sharing scientific data because this section is so sparse.

AMIA recommends this section include Supplemental Guidance documents on Plan Elements and Allowable Costs (as amended by our recommendations) so that ICOs have more direction and so that improved and consistent data management and sharing occurs across NIH-funded projects.

Recommendation V.b: Subject ICO-specific Policies to approval by the NIH Office of Data Science Strategy and the Office of Science Policy.

The NIH must establish a process to ensure alignment across ICO-specific DMSPs. While we do not dispute the need for variation based on domain and other circumstances, the NIH must coordinate disparate ICO DMSPs.

Recommendation V.b.1: Require ICOs to factor the quality of grantees' Plans into the overall impact score through a peer-review process for those grants that are supported at high levels or focused on programmatic priorities.

This is critical. Without accounting for the quality of data management and sharing as part of the grant selection process, this policy is feckless. As stated previously, making data sharing and management plans scorable elements of grants – not "just-in-time" requirements – is the best way to incentivize FAIR data principles.

Recommendation V.b.2: Require ICOs to identify and incentivize deposition of scientific data in endorsed depositories and knowledgebases.

The NIH has done a lot of work to determine how to differentiate between good and poor depositories and knowledgebases. The likely outcome of this policy will be a lot more scientific data and the NIH must play an active role in helping steer researchers towards quality depositories. We are happy to describe our thoughts further, but please see our response to a 2016 NIH RFI on Metrics to Assess Value of Biomedical Digital Repositories: <https://www.amia.org/sites/default/files/AMIA-Response-to-NIH-RFI-on-Metrics-to-Assess-Value-of-Biomedical-Digital-Repositories.pdf>

Section VI: Data Management and Sharing Plans:

Recommendation VI.a: Establish Parity Between the Rigor of Plan Review/Evaluation and Amount of NIH Funding Support.

We recommend the DMSP establish parity between the rigor of Plan review/evaluation and amount of NIH funding support. We strongly recommend that the draft DMSP encourage ICs to factor the quality of the Plan into the overall impact score through the peer review process for those grants that are supported at high levels or support programmatic priorities. While we support negotiation, making Plans scorable will improve the use of best practices and the general management and sharing posture of applicants far more efficiently than an "acceptable or unacceptable," evaluation schema. Rather than discouraging ICs from factoring Plan reviews/evaluations into the overall impact score, AMIA recommends ICs view quality Plans as essential to important research and design evaluation schemas to reflect this view.

Alternatively, the ICs could incentivize quality Plans by funding data management and sharing activities in an amount corresponding to the completeness of the Plan. For example, specific support of data managing and sharing activities might reflect the completeness of the plan, scored as "unsatisfactory" (0% of requested funds), "minimal" (25%), "adequate" (75%), "excellent" (100%). (Percentages for illustration only).

Recommendation VI.b: Require ICOs to make Plans Publicly Available

As with other aspects of this draft policy, the language is suggestive, but not explicit. This section states that, "NIH may make Plans publicly available," [emphasis on may]. The NIH should establish a policy that requires Plans publicly available, unless there are compelling reasons not to do so. A key goal of this DMSP should be to improve data management and sharing activities over time and making Plans publicly available will assist in this objective. We also contend that such transparency will improve accountability for funded projects to actually adhere to their Plans. AMIA recommends that this section state that "NIH will make Plans publicly available," [emphasis on will].

Section VII: Compliance and Enforcement:

Recommendation VII.a: Develop a Formal Endorsement Process of Preferred Databases and Knowledgebases.

AMIA generally supports the compliance section "During the Funding or Support Period," and "Post-Funding or Support Period." However, we note that data management is an ongoing process and that a management plan is updated, modified, and versioned. We anticipate that this part of the Plan could be part of the progress report statement. As for data sharing, we reiterate our recommendation that NIH develop a formal endorsement process of preferred databases and knowledgebases. These endorsed repositories would facilitate DMSP compliance and enforcement by having transparent terms and conditions and abide community consensus best practices. Researchers who use these NIH endorsed repositories would have a streamlined compliance process.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Recommendation SG Costs 1: Include this guidance as part of the NIH DMSP, not a supplemental guidance.

AMIA views supplemental guidance as too weak a designation for how to fund data management and sharing activities. Under the NIH strategy to empower ICOs to develop their own policies, we see this as yet another instance where the NIH should dictate with greater clarity its expectations – not leave it to ICOs to use (or not) the supplemental guidance.

Recommendation SG Costs 2: Establish a funding policy for data management and sharing activities that earmarks a percentage (at least 5 percent) of a grant award for such activities, rather than merely allow for such activities to be included in NIH budget requests.

We note that an advisory group to the European Commission has recommend that "well budgeted data stewardship plans should be made mandatory and we expect that on average about 5% of research expenditure should be spent on properly managing and stewarding data." AMIA believes that a similar expectation be set so as to help guide ICO-level policies. The citation for the above quote is: Commission High Level Expert Group on the European Open Science Cloud."Realising the European Open Science Cloud." 2016. ISBN 978-92-79-61762-1 doi:10.2777/940154.
https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Recommendation SG Elements 1: Include this guidance as part of the NIH DMSP, not a supplemental guidance.

Additional SG Element Recommendations: Below we offer comment and recommendation for each of the listed Elements.

i. Data Type

We recommend listing the find the term "rationale" in this section confusing. Given that the DMSP clearly articulates a rationale for scientific data preservation and sharing, we recommend this section simply state:

1. Data Type: Indicate the types and estimated amount of scientific data that will result from NIH-funded or -supported research and indicate how scientific data will be preserved and shared.

1.1. Amendments: We recommend inserting "expected" following "scientific data" in 1.1 to reflect that the data actually collected may change slightly over time. The expectation should be that the Plan will be directionally correct and complete, but that it could be subject to amendment. Further, we recommend rewording the second sentence of 1.1 as follows:

1.2. Amendments: We recommend adding the word "metadata" to 1.2, and we encourage the NIH to reference this defined term as appropriate throughout the document.

ii. Related Tools, Software and/or Code

We recommend the following changes to reflect these recommendations:

ii. Related Tools, Software and/or Code: Indicate what tools, software and/or code will be used to process or analyze the scientific data, why the software/code was chosen, and whether it is free and open source. Also indicate whether tools, software and/or code were developed to conduct NIH-supported research resulting in scientific data and if such artifacts are expected to be shared. The inclusion of scripts and the use of data and workflow diagrams, which graphically depicts at a high level the data sources, operations performed on the data, and the path taken by the data through information systems and operations may be useful.

iii. Standards

We recommend the following changes to reflect these recommendations:

iii. Standards: Indicate what standards, if any, apply to the scientific data to be collected, including data formats, data identifiers, data models, definitions, metadata and other data documentation, including terms of use. NIH encourages the use of existing data standards, such as standards for collecting and representing scientific data and information describing the

scientific data. NIH encourages the use of common data elements (CDEs) to facilitate broader and more effective use of scientific data and to advance research across studies. For assistance in identifying NIH-supported CDEs, the NIH has established a Common Data Element Resource Portal. For a list of established clinical data standards, please see the most recent Office of the National Coordinator for Health Information Technology Standards Advisory. Where commonly accepted standards don't exist, the Plan should include description of these standards in this section.

iv. Data Preservation and Access

4.1 Amendments: Data Deposition and Archiving: Indicate where scientific data will be archived to ensure its long-term preservation. If scientific data will be stored in an existing repository, provide the name and URL web address of the repository. If an existing repository will not be used, indicate why not and how scientific data preservation will be assured (e.g., in a newly created repository or by the investigator's organization).

4.2 Amendments Discoverability: Indicate how the scientific data will be made discoverable and whether a persistent unique identifier or other standard indexing tools will be used.

4.3 Amendments Security: Describe any provisions for maintaining the security and integrity of the scientific data (e.g., encryption and backups).

4.4 Amendments Plan Alternatives: Describe alternative plans for maintaining, preserving, and providing access to scientific data should the original Plan not be achieved.

4.5 Amendments Barriers: If perceived barriers to preserving and making accessible scientific data exist include an explanation of the perceived barriers.

4.6 Amendments Other Considerations: Indicate whether additional considerations are needed to preserve and make accessible the scientific data.

4.7 Amendments Biospecimens: Indicate whether scientific data generated from humans or human biospecimens will be available through unrestricted (made publicly available to anyone) or restricted access (made available after the requestor has received approval to use the requested scientific data for a particular project or projects). If the scientific data will be shared through a restricted access mechanism, describe the terms of access for the data.

4.8 Timeline Provide information on the anticipated timeframes for scientific data storage and accessibility, and criteria for how decisions affecting scientific data storage and accessibility will be made throughout the course of the study.

4.9 Amendments Secondary Use Timeline: Describe when the scientific data will be made available to secondary data users. This should be expressed in relation to some critical event, such as the publication of the major study findings, the end of data collection, or other similar activity.

v. Data Preservation and Access Timeline

AMIA recommends the DMSP merge Element 5 as subordinate points of Element 4 (see above Elements 4.8 and 4.9). We recommend that Element 5.2 be removed from the DMSP.

vi. Data Sharing Agreements, Licensing, and Intellectual Property

6.1 Amendments Data Sharing Agreements: Describe any existing data sharing agreement(s), outlining the responsibilities of each party, as well as how scientific data can and cannot be used.

6.2 Amendments Licensing: Describe any existing licensing terms, and any limitations on the scientific data use and reuse based on these terms. Describe whether the licensing is imposed by the applicant institution or whether it comes from any existing agreement(s).

6.3 Amendments Intellectual Property: If applicable, indicate how intellectual property, including invention or other proprietary rights, will be managed in a way to maximize sharing of

scientific data. Include any information relevant to the intellectual property rights associated with the scientific data, such as whether the intellectual property stems from an existing agreement or is anticipated to arise from the proposed research project itself.

vii. Oversight of Data Management

AMIA recommends removal of this section, given that grantees already provide personnel information in other parts of the grant. If it remains in the draft DMSP, we recommend a focus on the role rather than the individual to describe data management oversight and execution of the Plan.

Other Considerations Relevant to this DRAFT Policy Proposal:

AMIA additionally recommends that NIH take a proportionate approach to govern the sharing of health-related data involving populations with consent-related vulnerabilities, including but not limited to children and incompetent adults. Data from such populations should be collected, accessed, and exchanged for the purposes of advancing clinical understanding, and warrant special protections consistent with existing human subject regulations, international conventions, and jurisdictional data protection laws.

Attachment:

AMIA Response to 2019 NIH RFC on Data Management and Sharing Policy.pdf

Description:

Transmittal letter of AMIA Comments (please read in addition to template comments)



January 10, 2020

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
NIH Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Dr. Wolinetz:

Health Informatics is the science of how to use data, information, and knowledge to improve human health, the delivery of health care services, and the execution of scientific research. AMIA is the professional home for more than 5,500 informatics professionals, representing frontline clinicians, biomedical researchers, public health experts, and educators who bring meaning to data, manage information, and generate new knowledge across the healthcare system and research enterprise. AMIA members advance health and wellness by implementing and evaluating informatics interventions, innovations, and public policy across settings and patient populations, adding to our collective understanding of health in the 21st century through peer-reviewed journals and scientific meetings.

In 2018, AMIA responded to the “Proposed Provisions for a Draft NIH Data Management and Sharing Policy,” with enthusiastic support for a pan-NIH strategy and we commended the NIH for initiating a process to update its policy for the first time since 2003. If executed effectively, AMIA believes this policy could be transformative in how NIH-funded scientific data is accessed, exchanged, and used for secondary analysis and data-driven discovery. But such transformation will take leadership and coordination from Building 1, especially if the strategy is to empower individual Institutes, Centers, and Offices (ICOs) to establish their own, domain-specific requirements.

The opportunity inherent in this policy is to organize, categorize, and manage scientific data for retrospective and observational research, and to make publicly funded scientific data appropriately findable, accessible, interoperable, and reproducible, or FAIR. However, this proposed policy seems to perpetuate a check-the-box approach that subjugates the systematic collection, management, and deposition of data to a custodial exercise that will increase compliance burden without commensurate, downstream benefits or utility. Further, the proposed policy provides insufficient direction to ICOs through a weak “guidance” mechanism that is unlikely to result in coordinated, consistent, and harmonized data management and sharing activities across NIH ICOs. This is especially confusing and problematic given that several NIH efforts, spanning billions of

January 10, 2020

dollars in annual funding, primarily focus on supplemental use and secondary analysis of data, including the All of Us Research Program,¹ the National Center for Data to Health (CD2H),² the Accrual to Clinical Trials (ACT) network,³ Informatics for Integrating Biology and the Bedside (i2b2)⁴ and its own Data Science Strategy.⁵ Additionally, the Administration has charged Executive Branch agencies and offices to consider how they will leverage data as a strategic asset, through the Federal Data Strategy,⁶ elevating data management and sharing to one of the highest priorities of the White House Office of Management and Budget and Office of Science and Technology Policy.

Unfortunately, much of what we recommended more than a year ago still pertains to this second iteration Request for Public Comment. We have reproduced AMIA’s comments to the 2018 document in full at [Appendix A](#). As proposed, the most recent Data Management and Sharing Policy (DMSP) represents a missed opportunity to modernize the 2003 policy and to reorient data management and sharing activities at the NIH for data-driven discovery. Specifically, AMIA strongly objects to:

- The DMSP’s requirement for “just-in-time” development of data management and sharing plans (Plans) rather than a requirement to develop a Plan as part of the grant proposal;
- The use of Supplemental Guidance to discuss potential Plan Elements and Allowable Costs, rather than including these as policies in the proposed DMSP;
- The strategy to subject Plans to programmatic assessment by NIH staff rather than experts who can differentiate between high- and low-quality data management and sharing activities;
- The contention described in the Supplemental Guidance on Elements of a Plan that a data management and sharing plan could be adequately described in “two pages or less”; and
- Continuing definitional ambiguities and omissions in the DMSP and Plan Elements.

The net result of the proposed DMSP will be wasted time, effort, and money on behalf of researchers, widely divergent policies across ICOs, and a rapidly growing corpus of scientific data with limited utility for observational research and secondary analysis. **Thus, AMIA strongly recommends the NIH re-consider our 2018 comments and dramatically amend the current proposed policy to achieve three core goals: (1) Optimize scientific data once generated; (2) Incentivize improvements in data management and sharing practices; and (3) Coordinate disparate ICO data management and sharing policies.**

To achieve these goals, AMIA recommends that the NIH:

1. Finalize a pan-NIH DMSP that positions ICOs to develop their own requirements, subject to approval by the NIH Office of Data Science Strategy and the Office of Science Policy;
2. Take a stronger leadership position in establishing guardrails for ICOs by
 - a. Requiring ICOs to factor the quality of grantees’ Plans into the overall impact score of applications through a peer-review process for those grants that are supported at high levels or focused on programmatic priorities;

¹ <https://allofus.nih.gov/>

² <https://ctsa.ncats.nih.gov/cd2h/>

³ <https://www.actnetwork.us/National>

⁴ <https://www.i2b2.org/>

⁵ <https://datascience.nih.gov/>

⁶ <https://strategy.data.gov/>

January 10, 2020

- b. Requiring ICOs to identify and incentivize deposition of scientific data in endorsed depositories and knowledgebases;
 - c. Enable ICOs to establish graduated Plan requirements based on funding levels, subject to the aforementioned NIH review
3. Implement the NIH DMSP over the span of three years, requiring grant proposals subject to the existing policy (i.e., grants above \$500,000 per year) to comply initially, giving grants of lesser amounts additional time to comply;
4. Establish a funding policy for data management and sharing activities that earmarks a percentage (at least 5 percent)^{7, 8} of a grant award for such activities, rather than merely allow for such activities to be included in NIH budget requests;
5. Include the Supplemental Guidance on Allowable Costs for Data Management and Sharing as part of the DMSP, not as Supplemental Guidance, to ensure consistency across ICOs;
6. Include the Supplemental Guidance on Elements of a NIH Data Management and Sharing Plan as part of the DMSP, not as Supplemental Guidance, and adopt more directive language to establish required elements for ICOs;

It is imperative that the NIH view scientific data – not the presentation at the conference or the journal publication that purports to describe the data – as the principal result of scientific research. Everything we value from scientific research follows from the right analysis of data, so the NIH must take the position that good data stewardship is an essential component of the scientific enterprise, rather than a “just-in-time” afterthought or byproduct of the “real” research activities.

Furthermore, there are technologies and toolkits available to make data management and sharing more efficient. Many AMIA members are actively engaged in building and evaluating tools such as the CEDAR Workbench, which makes it easy to create comprehensive metadata for experimental datasets, and to upload the data and metadata to public repositories.⁹ And global consortia like the Global Alliance for Genomics & Health have developed toolkits for Genomic Data deposition, regulatory and ethics compliance, and data security.¹⁰

A robust DMSP is necessary to optimize these investments so that new discoveries can be identified across these programs and so that all NIH-funded research can add to our national strategic asset of life sciences and biomedical data. As mentioned in our previous comments, AMIA has numerous experts that can be made available to NIH policymakers and our offer to assist still stands.

⁷ An advisory group to the European Commission has recommend that “well budgeted data stewardship plans should be made mandatory and we expect that on average about 5% of research expenditure should be spent on properly managing and stewarding data.” Commission High Level Expert Group on the European Open Science Cloud. “Realising the European Open Science Cloud.” 2016. ISBN 978-92-79-61762-1 doi:10.2777/940154.
https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

⁸ Similar to the High Level Expert Group, the European Research Council Scientific Council has recognized that “data annotation and deposition are time-consuming activities. ERC grant money can be specifically earmarked for this purpose, for example to contribute to the salary of a research assistant or to the costs of a commercial provider” via the report “Open Research Data and Data Management Plans, Information for ERC grantees.” Version 3.1. 3 July 2019.
https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf

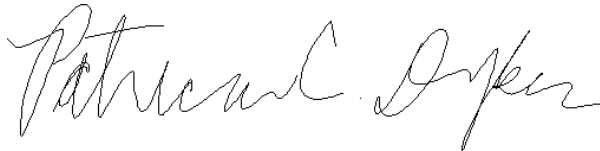
⁹ See: <https://metadacenter.org/>. Also: Musen, M.A., Bean, C.A., Cheung, K.-H., et al. The Center for Expanded Data Annotation and Retrieval. JAMIA 22(6):1148–1152, 2015

¹⁰ See: <https://www.ga4gh.org/> for more.

January 10, 2020

We hope our comments are helpful as you undertake this important work. Should you have questions about these comments or require additional information, please contact Jeffery Smith, Vice President of Public Policy at jsmith@amia.org or (301) 657-1291. We look forward to continued partnership and dialogue.

Sincerely,



Patricia C. Dykes, PhD, RN, FAAN, FACMI
Chair, AMIA Board of Directors
Program Director Research
Center for Patient Safety, Research, and Practice
Brigham and Women's Hospital

January 10, 2020

Appendix A: AMIA Response to Proposed Provisions for a Draft NIH Data Management and Sharing Policy



December 10, 2018

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
NIH Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: Proposed Provisions for a Draft NIH Data Management and Sharing Policy

Dr. Wolinetz:

AMIA's membership is comprised of informaticians across the spectrum of biomedical research, clinical care, public health, and consumer health, with backgrounds in medicine, biomedical sciences, and informatics. Our comments are rooted in this expertise and are representative of diverse and multidisciplinary stakeholders who are deeply experienced in the systematic collection, analysis, application and responsible sharing of data for health.

AMIA enthusiastically supports development of a pan-NIH Data Management and Sharing Policy (DMSP) and we commend the NIH for initiating this effort. We are pleased to see several elements of AMIA's Data Sharing Principles & Positions incorporated in the Proposed Provisions, including a reliance on FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles, and acknowledgment that the DMSP should support underlying infrastructure and curation activities with funding.

We are especially pleased that the NIH envisions a DMSP that applies to "all intramural and extramural research, funded or supported in whole or in part by NIH, that results in scientific data, regardless of NIH funding level or mechanism." While this scope is ambitious, quality data management and sharing plans (Plans) are prerequisite to achieve the vision of FAIR data principles and such a scope should be the long-term goal of the NIH DMSP.

January 10, 2020

Recognizing the need to have all NIH-funded research comply with this DMSP, and with appreciation of what AMIA sees as necessary components of a data management and sharing plan, we recommend a phased compliance timeline based on funding levels. This phased implementation would only apply to new research funded after the DMSP is final. First, new research funded above \$500,000 per year and subject to the existing data sharing policy should comply with the final DMSP within one year of its adoption. Second, new research funded above \$250,000 per year should comply with the provisions of the DMSP within 2 years of its adoption, and finally, all grants funded below \$250,000 per year should comply with the DMSP within 3 years of adoption. This compliance approach would focus efforts on those grants that already must comply with the existing policy and likely have the richest cache of scientific data, while giving smaller projects more time to become familiar with the DMSP.

Alongside this phased adoption timeline, the NIH should consider a graduated DMSP that appropriately calibrates requirements based on funding level and whether scientific data are deposited in an NIH-endorsed depository or knowledgebase. We strongly recommend that the draft DMSP encourage Institutes and Centers (ICs) to factor the quality of the Plan into the overall impact score through the peer review process for those grants that are supported at high levels or support programmatic priorities. We also recommend that NIH incentivize deposition of scientific data in NIH-endorsed databases and knowledgebases by allowing such Plans to comply with a streamlined DMSP.

We note several high-level observations and recommendations for which we provide additional detail and rationale in the enclosure of this comment letter:

- 1. The draft DMSP should improve data management and sharing of scientific data to facilitate learning health systems and continuous discovery.**
 - a. While we are supportive of a pan-NIH DMSP, subject to ICs specific grant-types and awardees, AMIA recommends the DMSP encourage ICs to make Plans scorable elements of specific grants. This will improve Plans' quality and better ensure supplemental use of scientific data.
 - b. AMIA also recommends the DMSP seeks to improve the interoperability and supplemental uses of research data writ large by encouraging the use of established biomedical data standards and adherence to data management and data sharing best practices. Over time, better use of and refinement of data standards, buttressed by systematic scoring of plans, will optimize scientific data for continuous learning and discovery.
 - c. AMIA recommends the DMSP incentivize the deposition of scientific data and tools, software and/or code developed as part of NIH-supported projects into NIH-approved data repositories and knowledgebases. This will enable both large and small grantees to more easily comply with the DMSP.
- 2. The draft DMSP should improve institutional support and professional advancement for experts managing and sharing scientific data.**
 - a. We applaud NIH for suggesting that reasonable costs associated with data management and sharing could be requested under the budget for the proposed

January 10, 2020

project. AMIA recommends that the DMSP establish a standard way to account for data management and sharing costs as both Direct costs and F&A costs.

- b. The DMSP should facilitate implementation of the NIH Data Science Strategic Plan, especially the relevant aspects of the Strategic Plan that seek to credit experts who manage and share valuable data sets / software for their work. If data is seen as valuable, experts who enable FAIR data should also be valued. The NIH should support certifications for experts that manage and share scientific data. We also see a need for R&D on data management tools to facilitate compliance with the DMSP.
- 3. To operationalize the DMSP more specificity and clarity around concepts is needed.**
- a. Data management is distinct from data sharing. The processes and activities that support data management and sharing are also different. AMIA recommends the NIH develop a DMSP that specifies these distinctions through additional Plan Elements as described below.
 - b. AMIA recommends that the DMSP expand the current list of definitions to include concepts for “Data Management,” “Covered Data,” “Covered Timeframe,” and refine definitions for “Metadata” and “Scientific Data.”
 - c. While we support the scope of a pan-NIH DMSP that covers all grants, contracts, and/or other funding agreements, AMIA recommends the NIH convene stakeholders with individual ICs to operationalize the DMSP.

Finally, we offer AMIA and its members as resources during subsequent work on the DMSP. We strongly recommend the NIH develop a subsequent draft DMSP based on stakeholder feedback to the concepts in this RFI. Another comment period will provide NIH with valuable insights before issuing a final DMSP.

The enclosure includes detailed AMIA comments regarding the Proposed Provisions for a Draft NIH Data Management and Sharing Policy. Where possible, we provide both in-line edits and rationale for suggested edits.

- I. [Definitions](#)
 - a. [Data Management and Sharing Plan](#)
 - b. [Data Management](#)
 - c. [Data Sharing](#)
 - d. [Metadata](#)
 - e. [Scientific Data](#)
 - f. [AMIA Recommended New Definitions](#)
- II. [Purpose](#)
- III. [Scope and Requirements](#)
- IV. [Requirements for Data Management and Sharing Plans](#)
 - a. [Plan Review and Evaluation](#)
 - b. [Plan Elements](#)
 - i. [Data Type](#)
 - ii. [Related Tools, Software and/or Code](#)
 - iii. [Standards](#)
 - iv. [Data Preservation and Access](#)

January 10, 2020

- v. [Data Preservation and Access Timeline](#)
 - vi. [Data Sharing Agreements, Licensing, and Intellectual Property](#)
 - vii. [Oversight of Data Management](#)
 - viii. [Other Considerations](#)
- V. [Compliance and Enforcement](#)

We hope our comments are helpful as you undertake this important work. Should you have questions about these comments or require additional information, please contact Jeffery Smith, Vice President of Public Policy at jsmith@amia.org or (301) 657-1291. We look forward to continued partnership and dialogue.

Sincerely,



Douglas B. Fridsma, MD, PhD, FACP, FACMI
President and CEO
AMIA

Enclosed: Detailed AMIA comments regarding the Proposed Provisions for a Draft NIH Data Management and Sharing Policy.

January 10, 2020

I. Definitions

a. Data Management and Sharing Plan

AMIA Comments: The draft Data Management and Sharing Policy (DMSP) should differentiate between “data management” and “data sharing” as two distinct concepts and sets of activities with different, if overlapping, considerations and timeframes. For clarity, we refer to the Data Management and Sharing Plan as “Plan” and DMSP refers to the policy. While we are supportive of the focus on data sharing as part of data management, it is critical to acknowledge that upstream data collection and handling processes largely determine data quality necessary for research replicability, reproducibility, and traceability.¹¹

AMIA Comments: The draft DMSP should not distinguish between potential “others” who may use scientific data. The number and heterogeneity of “others,” even when confined to “researchers,” and “the broader public,” would needlessly complicate compliance with the DMSP. AMIA members note a major discrepancy between making scientific data available to another scientist in the same discipline and making it available to the general public. Further, we note that even within the scientific community, there will be wide gaps in knowledge across disciplines that requires extensive annotation and training to be understood.

AMIA Recommendation: AMIA recommends the following amendments to the Plan’s definitions to acknowledge differences in data management and data sharing. Further AMIA recommends the draft DMSP remove all references to “(e.g. researchers and the broader public)” when describing potential users of scientific data:

Data Management and Sharing Plan: A plan describing how scientific data will be generated, managed, described, analyzed, preserved, shared, and made accessible to others for supplemental uses, (e.g., other researchers and the broader public), as appropriate. This plan should include two distinct sections describing how scientific data will be managed across the life-cycle of the project and how scientific data will be shared at the project close, or at another appropriate interval(s).

b. Data Management

AMIA Comments: As discussed above, the DMSP should explicitly describe what is necessary to manage data, not just share data, given that data management and data sharing are distinct. Data management is prerequisite for data sharing, ensuring that the data are accurate, complete, and maintained in a standardized manner. Without effective data management, you cannot have effective data sharing, thus we recommend the DMSP consider additional Plan Elements as described in that section of our comments.

AMIA Recommendation: Given this view, we recommend the draft DMSP include a new definition for data management as follows:

¹¹ Traceability of research data is the ability to reproduce the raw data from the analysis datasets and vice versa.

January 10, 2020

Data Management: The upstream management of scientific data that documents actions taken in making research observations, collecting research data, describing data (including relationships between datasets), processing data into intermediate forms as necessary for analysis, integrating distinct datasets, and creating metadata descriptions. Specifically, those actions that would likely have impact on the quality of data analyzed, published, or shared.¹²

c. Data Sharing

Data Sharing: ~~To make~~ **Making** scientific data accessible for use by others (e.g., other researchers and the broader public) in a manner that is consistent with the FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles.

d. Metadata

AMIA Comments: We found the definition for metadata in need of refinement. Specifically, the phrase “additional information to make data more usable” implies that a data set could be usable at all without metadata, which is simply not the case. There is no data that can be correctly understood, much less re-used, without at least a data dictionary with field definitions and data types. Further, we view “Outcome measures” as actual data, not metadata. There may be metadata that defines how an outcome measure was derived, but the outcome data itself is not metadata.

AMIA Recommendation: Given this view, we recommend the draft DMSP amend the definition of metadata as follows:

Metadata: ~~Data that provide additional information to make data more usable (e.g., independent sample and variable description, outcome measures, and any intermediate, descriptive, or phenotypic observational variables).~~ Metadata is descriptive information about data, including variable/document definition/description, data type, and other characteristics. Areas discussed in metadata include, but are not limited to, instruments used to collect data; parameters or settings for such instruments; descriptors of physical samples from which data were collected; dates and times of data collection; any transformations applied to the data; relationships between datasets; provenance linking derived or modified datasets to original sources; phenotypic descriptors of data sources; and institutional/personal identifying information associated with the group or person(s) responsible for the data. Metadata also help establish (confidence in) the credibility of the data.

In survey data, “paradata” is used to describe confidence in the credibility of data. This may be an evaluation of the sincerity or seriousness of the respondent by the questioner (e.g. “Open/Frank” to “Uncomfortable/Evasive”, “Earnest” to “Flippant”, etc.), or less subjectively, in an online survey, the time the respondent spent to complete the survey.

¹² Adapted from Williams, Bagwell and Zozus “Data management plans, the missing perspective,” Journal of Biomedical Informatics 71 (2017) 130–142

January 10, 2020

e. Scientific Data

AMIA Comments: We support the concept of “Scientific Data,” but do not support a definition of this concept through negation. The listing of what Scientific Data is not may serve better as part of ancillary materials published by the NIH, such as Frequently Asked Questions, rather than be included in a definition. Further, it is odd to place a command, “NIH expects…” into a definition.

AMIA Recommendation: Given this view, we recommend the draft DMSP include a new definition for Scientific Data as follows:

Scientific Data: ~~The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, scientific data may include certain individual level and summary or aggregate data, as well as metadata. NIH expects that reasonable efforts should be made to digitize all scientific data.~~ 1. Information that is gathered, derived or generated in the course of conducting research. It is the basis for reaching conclusions and inferences based on scientific principles and methodologies. Scientific data can be used to test existing hypotheses, to generate new hypotheses for future research, to validate or replicate prior research as well as for more exploratory purposes. Scientific data represent the foundation for both scientific theories and publications.

f. AMIA Recommended New Definitions

1. Scientific Software Artifacts

AMIA Comment: Increasingly, the NIH funds research that results in software tools, code, and analytic programs. These “software artifacts” are both explicitly funded as part of extramural research and developed as a means to conduct NIH-funded research. These software artifacts can be deposited in knowledgebases analogous to data into databases.

AMIA Recommendation: AMIA recommends the draft DMSP includes a definition for “Scientific Software Artifacts,” so that grantees clearly understand that both data and software tools created with NIH funds should be included as part of their data management and sharing plan. This definition would be limited to artifacts created with NIH funds, and omit proprietary software tools used to conduct research, such as a stat package. We recommend a definition such as:

Scientific Software Artifacts: Software, code, analytic programs, and other knowledge artifacts developed to conduct research or resulting from the conduct of research.

2. Covered Data

AMIA Comment: We see the need to define two additional terms so that the DMSP can address a number of questions that arise throughout our deliberations. Specifically, grantees need to have a

January 10, 2020

clear understanding of which Scientific Data are covered by the Policy and for what period of time those data are covered. These definitions do not need to establish a policy for these questions; rather, these concepts should facilitate conversations to answer those questions.

There is a distinction between data generated by and for research, and data that is used in research. We see a need to define what scientific data is covered under the DMSP and what data is not. For example, clinical trials routinely rely on data that has been generated during the course of clinical care and collected as part of research participants' electronic health record (EHRs). This data may be included in the study data set and used as part of an analysis. Such data was not specifically generated for the trial and the tests or other work involved in generating them were not paid for by the trial. Is such data covered by the policy or not?

As another example, we note that a number of large databases are currently used and made available for epidemiological research or data mining projects based entirely on real-world evidence. These data are generated and paid for in the course of routine clinical care and are maintained under private funding. If an NIH funded analytic project is based on the use of such data, can that data now be required to be made available more generally to the public? If so, this could represent a disincentive to the private organization to make such data available for research and might have the paradoxical effect of making less data available for research or making whole classes of data unavailable for research.

AMIA Recommendation: AMIA recommends the draft DMSP includes a definition for “Covered Data,” so that grantees clearly understand which data must be included as part of their data management and sharing plan. We recommend a definition such as:

Covered Data: Those newly generated or derived Scientific Data used to conduct NIH-funded or -supported research and subject to this Policy. Such data may or may not be proprietary and subject to various access controls.

3. Covered Period

AMIA Comment: We also note a need to define the expected timeframe for which grantees must steward Scientific Data. While we have numerous questions, such as, does the transfer of data to an NIH-supported or endorsed repository complete the obligation of the grantee? Will there be funding available to grantees who steward their own Scientific Data associated with tracking and satisfying data use requests? Would there be some appeal process if the volume/complexity of requests exceeds what was anticipated or funded? Or could the grantee charge some reasonable administrative fee if total costs incurred exceed some threshold?

AMIA Recommendation: We recommend the NIH address these and other questions by incorporating a concept of “Covered Period.” This term would facilitate greater understanding of the obligations of grantees

Covered Period: The period of time for which the Scientific Data is expected to be maintained by the grantee and for which it is to be made available to others.

January 10, 2020

II. Purpose

AMIA Comment: This section describes what the DMSP is, but only hints at why the NIH is proposing one and how it will interact with other NIH policies. This section should describe why a DMSP is necessary and what a DMSP will achieve.

AMIA Recommendation: The draft DMSP should bolster the Purpose section by adding language similar to the introductory language, beginning, “**NIH has a longstanding commitment to making the results and accomplishments of the research that it funds and conducts available to the public. Increasing access to scientific data resulting from NIH funding or support offers many benefits and reflects NIH’s responsibility to maintain stewardship over taxpayer funds.**” AMIA recommends the draft DMSP adds to this with the following:

Specifically, systematic management and sharing of scientific data and results enables researchers to more vigorously test the validity of research findings, strengthen analyses by combining data sets, access hard-to-generate data, and explore new frontiers. Data management and sharing also informs future research pathways, increases the return on investment of scientific research funding, and accelerates the translation of research results into knowledge, products, and procedures to improve health and prevent disease.

This Policy seeks to identify, adopt, and credit data management and sharing best practices, consistent with FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles, so that the United States remains the leader in biomedical and life sciences research. This Policy establishes the requirements and responsibilities of researchers generating scientific data resulting from NIH-funded or -supported research and it will govern development and implementation of other NIH Policies related to the management and sharing of scientific data, such as the NIH Genomic Data Sharing Policy, the NIH Policy on the Dissemination of NIH-funded Clinical Trial Information, and the Intramural Research Program Human Data Sharing (HDS) Policy.

III. Scope and Requirements

AMIA Comment: We applaud the NIH for considering a comprehensive DMSP that would “apply to all intramural and extramural research, funding or supported in whole or in part by NIH, that results in scientific data, regardless of NIH funding or mechanism.” A pan-NIH DMSP will improve our national culture of data sharing, as well as facilitate the FAIR data principles.

We also support submission of a Data Management and Sharing Plan (Plan) as part of the funding/support application process and articulating if there are perceived barriers to sharing scientific data in this Plan. Finally, we greatly appreciate that these draft policy provisions state that “Reasonable costs associated with data management and sharing could be requested under the budget for the proposed project.”

January 10, 2020

AMIA Recommendation: We urge the NIH to proceed with the proposed DMSP scope, ensuring that the policy requirements are constructed in a way that both small and large awardees can comply. While we agree that it is important for all NIH research to be subject to this policy, regardless of funding or mechanism, the policy must maintain flexibility to accommodate individual ICs and individual project characteristics.

AMIA recommends the NIH draft this section as “**III. Scope**” and position the aspects of the current provisions related to “requirements” in the next section, “IV Requirements for Data Management and Sharing Plans.” The draft DMSP could expand on the rationale for its scope, similar to the Purpose section. We discuss issues related to IC-specific requirements and “reasonable costs,” for data management and sharing below.

IV. Requirements for Data Management and Sharing Plans

a. Plan Review and Evaluation

AMIA Comment: Establishing a flexible, yet consistent, and fair review and evaluation strategy will greatly improve the likelihood that this DMSP is successful. We note that the proposed policy envisions that review and evaluation would be the primary responsibility of the funding or supporting NIH IC, “which could be implemented in a variety of ways...” and that this section delineates how various funding mechanisms might differently approach the task of review and evaluation. We are generally supportive of this strategy as long as the DMSP provides direction for ICs to rationalize and harmonize their specific requirements.

As it relates to Extramural Grants, we are concerned that scoring in a binary way has contributed to our current shortcomings in quality data management and sharing. As stated previously, we view rigorous review and evaluation of Plans as a means to improve the FAIR-ness of data and encourage the NIH to treat these Plans as scorable elements of certain grant applications.

AMIA Recommendation: We recommend the DMSP establish parity between the rigor of Plan review/evaluation and amount of NIH funding support. We strongly recommend that the draft DMSP encourage ICs to factor the quality of the Plan into the overall impact score through the peer review process for those grants that are supported at high levels or support programmatic priorities. While we support negotiation, making Plans scorable will improve the use of best practices and the general management and sharing posture of applicants far more efficiently than an “acceptable or unacceptable,” evaluation schema. Rather than discouraging ICs from factoring Plan reviews/evaluations into the overall impact score, AMIA recommends ICs view quality Plans as essential to important research and design evaluation schemas to reflect this view.

Alternatively, the ICs could incentivize quality Plans by funding data management and sharing activities in an amount corresponding to the completeness of the Plan. For example, specific support of data managing and sharing activities might reflect the completeness of the plan, scored as “unsatisfactory” (0% of requested funds), “minimal” (25%), “adequate” (75%), “excellent” (100%). (Percentages for illustration only).

January 10, 2020

This recommendation notwithstanding, we do see value in considering a binary evaluation in limited circumstances, such as small grants to new investigators, or in cases where scientific data cannot be de-identified and shared.

b. Plan Elements

AMIA Comment and Recommendation: Given the extent of information expected as part of a Plan, we do not envision a 2-page limit will be sufficient in most circumstances. Rather than setting arbitrary page limits through the DMSP, we recommend the NIH leave length and depth of Plans to peer review and IC guidance.

We are generally supportive of the Plan Elements listed. However, we believe there is a need to include additional Elements so that applicants can describe their Data Management activities. We also recommend “Data Preservation and Access Timeline” be included as a sub-point of “Data Preservation and Access,” rather than a standalone Element. Below we offer comment and recommendation for each of the listed Elements.

i. Data Type

AMIA Comment and Recommendation: We recommend listing the find the term “rationale” in this section confusing. Given that the DMSP clearly articulates a rationale for scientific data preservation and sharing, we recommend this section simply state:

1. Data Type: Indicate the types and estimated amount of scientific data that will result from NIH-funded or -supported research and indicate **how** the rationale for which scientific data will be preserved and shared.
- 1.1. Amendments:** We recommend inserting “**expected**” following “scientific data” in 1.1 to reflect that the data actually collected may change slightly over time. The expectation should be that the Plan will be directionally correct and complete, but that it could be subject to amendment. Further, we recommend rewording the second sentence of 1.1 as follows:

Describe the data modality (e.g., imaging, genomic, mobile, **patient-reported**, and survey) and whether the scientific data will be individual, aggregated, or summarized, and **whether the data will be** ~~how raw or processed the data will be.~~

- 1.2. Amendments:** We recommend adding the word “**metadata**” to 1.2, and we encourage the NIH to reference this defined term as appropriate throughout the document.

Describe any other information that is anticipated to be shared along with the scientific data, such as relevant associated data, and any other information necessary to interpret the data (e.g., study protocols ~~and~~ data collection instruments, **and other metadata**).

ii. Related Tools, Software and/or Code

January 10, 2020

AMIA Comment and Recommendation: AMIA supports efforts to make tools, software and/or code available for use, if such artifacts were developed as the result of NIH funding. However, there is a fundamental difference between sharing data and sharing code or software, particularly if the code is considered proprietary, such as a purchased stat package. The intent of this policy should be twofold: (1) To improve replicability by ensuring transparency in how data were transformed and (2) encourage the sharing of related tools, software and/or code generated through NIH funding. The intent should not be to make researchers provide an analytic environment, open source or otherwise. The use of data and workflow diagrams, which graphically depicts at a high level the data sources, operations performed on the data, and the path taken by the data through information systems and operations may be useful.

While we support the use of alternative free or open source code, we do not view the DMSP as an appropriate vehicle to encourage such solutions. The effort to identify such tools could be significant and may require skills well beyond those of the investigator and requiring assistance from staff not included in any of the grant funding. We recommend the following changes to reflect these recommendations:

ii. Related Tools, Software and/or Code: Indicate what **tools**, software **and/or computer** code will be used to process or analyze the scientific data (~~the inclusion of scripts may be helpful~~), why the software/code was chosen, and whether it is free and open source. **Also indicate whether tools, software and/or code were developed to conduct NIH-supported research resulting in scientific data and if such artifacts are expected to be shared.** ~~If software/code that is not free and open source is needed to access or further analyze the scientific data, briefly describe why this particular software/code is needed. Describe whether there is an alternative free and open source software/code that may be used to further analyze the scientific data.~~ **The inclusion of scripts and the use of data and workflow diagrams, which graphically depicts at a high level the data sources, operations performed on the data, and the path taken by the data through information systems and operations may be useful.**

iii. Standards

AMIA Comment and Recommendation: AMIA appreciates the NIH pointing towards and encouraging use of established data standards, common data elements, and other publicly funded initiatives. We support leveraging this DMSP to encourage the use of existing data standards and common data elements (CDEs) to “facilitate broader and more effective use of scientific data and to advance research across studies.” We hope that, over time, researchers will coalesce around common standards when appropriate and that when common standards can be used they are used. This will only happen if Plans are critically peer reviewed by experts trained in the systematic collection, analysis, and application of data. We recommend the following changes to reflect these recommendations:

iii. Standards: Indicate what standards, if any, apply to the scientific data to be collected, including data formats, data identifiers, **data models**, definitions, **metadata** and other data documentation, including terms of use. NIH encourages the use of existing data standards, such as standards for

January 10, 2020

collecting and representing scientific data and information describing the scientific data. NIH encourages the use of common data elements (CDEs) to facilitate broader and more effective use of scientific data and to advance research across studies. For assistance in identifying NIH-supported CDEs, the NIH has established a Common Data Element Resource Portal. **For a list of established clinical data standards, please see the most recent Office of the National Coordinator for Health Information Technology Standards Advisory.**¹³ Where commonly accepted standards don't exist, the Plan should include description of these standards in this section.

iv. Data Preservation and Access

AMIA Comment and Recommendation: AMIA encourages the NIH to be more prescriptive in its expectations that Plans leverage NIH-supported data repositories.¹⁴ AMIA recommends the NIH incentivize the deposition of scientific data into NIH-supported data repositories by scoring or funding such Plans higher than Plans that do not use an NIH-supported or NIH-approved data repository (unless sufficient justification can be made) and by allowing Plans that leverage such repositories to forego this section of the DMSP. This may require the NIH to better understand the relative strengths and weaknesses of repositories currently/potentially supported by the NIH, but it will improve the likelihood of long-term data FAIRness. AMIA recommends the NIH develop a formal endorsement process to approve and list preferred repositories for scientific data and scientific software artifacts.

In addition, the NIH should use information gathered during the 2016 RFI on “Metrics to Assess Value of Biomedical Digital Repositories,” to inform policy development in this area. While AMIA acknowledged there “will be no ‘one-size fits all’ scorecard” in comments to this RFI, we provided several recommendations for the NIH to develop a rating schema for deposition repositories and knowledgebases.¹⁵

If Plans wish to rely on data repositories other than those supported or endorsed by NIH, we recommend the following aspects be articulated (we reference existing sub-element numbers below):

4.1 Amendments Data Deposition and Archiving: Indicate where scientific data will be archived to ensure its long-term preservation. If scientific data will be stored in an existing repository, provide the name and URL web address of the repository. If an existing repository will not be used, indicate why not and how scientific data preservation will be assured (e.g., in a newly created repository or by the investigator's organization).

4.2 Amendments Discoverability: Indicate how the scientific data will be made discoverable and whether a persistent unique identifier or other standard indexing tools will be used.

4.3 Amendments Security: Describe any provisions for maintaining the security and integrity of the scientific data (e.g., encryption and backups).

¹³ <https://www.healthit.gov/isa/>

¹⁴ https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

¹⁵ AMIA Comments available at: <https://www.amia.org/sites/default/files/AMIA-Response-to-NIH-RFI-on-Metrics-to-Assess-Value-of-Biomedical-Digital-Repositories.pdf>

January 10, 2020

4.4 Amendments Plan Alternatives: Describe alternative plans for maintaining, preserving, and providing access to scientific data should the original Plan not be achieved.

4.5 Amendments Barriers: If perceived barriers to ~~sharing~~ **preserving and making accessible** scientific data exist (e.g., ~~sharing includes specific restrictions or sharing is not possible~~), outline how scientific data will be managed and preserved and include an explanation of the perceived barriers.

4.6 Amendments Other Considerations: Indicate whether additional considerations are needed to preserve and make accessible ~~implement~~ **the scientific data. Plan** (e.g., ~~prior permission to use a specific repository~~).

4.7 Amendments Biospecimens: Indicate whether scientific data generated from humans or human biospecimens will be available through unrestricted (made publicly available to anyone) or restricted access (made available after the requestor has received approval to use the requested scientific data for a particular project or projects). If the scientific data will be shared through a restricted access mechanism, describe the terms of **access** for the data.

New 4.8 Timeline: Provide information on the anticipated timeframes for scientific data storage and accessibility, and criteria for how decisions affecting scientific data storage and accessibility will be made throughout the course of the study.

New 4.9 Amendments: Secondary Use Timeline: Describe when the scientific data will be made available to secondary data users. This should be expressed in relation to some critical event, such as the publication of the major study findings, the end of data collection, or other similar activity.

v. Data Preservation and Access Timeline

AMIA Comment and Recommendation: AMIA recommends the DMSP merge Element 5 as subordinate points of Element 4 (see above Elements 4.8 and 4.9). We recommend that Element 5.2 be removed from the DMSP.

vi. Data Sharing Agreements, Licensing, and Intellectual Property

AMIA Comment and Recommendation: AMIA supports the expectation that scientific data will be broadly available, consistent with privacy, security, informed consent, and proprietary issues. We note that this information may be duplicative with information provided in prior Elements, such as barriers to preservation / access, and we encourage NIH to reduce sections that overlap in intent or required content.

6.1 Amendments Data Sharing Agreements: Describe any **existing** data sharing agreement(s), outlining the responsibilities of each party, as well as how scientific data can and cannot be used.

January 10, 2020

6.2 Amendments Licensing: Describe any ~~existing general~~ licensing terms, and any limitations on the scientific data use and reuse based on these terms. Describe whether the licensing is imposed by the applicant institution or whether it comes from any existing agreement(s).

6.3 Amendments Intellectual Property: If applicable, indicate how intellectual property, including invention or other proprietary rights, will be managed in a way to maximize sharing of scientific data. Include any information relevant to the intellectual property rights associated with the scientific data, such as whether the intellectual property stems from an existing agreement or is anticipated to arise from the proposed research project itself.

vii. Oversight of Data Management

AMIA Comment and Recommendation: AMIA recommends removal of this section, given that grantees already provide personnel information in other parts of the grant. If it remains in the draft DMSP, we recommend a focus on the role rather than the individual to describe data management oversight and execution of the Plan.

viii. Other Considerations

AMIA Comment and Recommendation: AMIA views the additional considerations as important context that could be used to

V. Compliance and Enforcement

AMIA Comment and Recommendation: AMIA generally supports the compliance section “During the Funding or Support Period,” and “Post-Funding or Support Period.” However, we note that data management is an ongoing process and that a management plan is updated, modified, and versioned. We anticipate that this part of the Plan could be part of the progress report statement. As for data sharing, we reiterate our recommendation that NIH develop a formal endorsement process of preferred databases and knowledgebases. These endorsed repositories would facilitate DMSP compliance and enforcement by having transparent terms and conditions and abide community consensus best practices. Researchers who use these NIH endorsed repositories would have a streamlined compliance process.

Submission ID: 1360

Date: 1/10/2020

Name: M. Saiful Huq, PhD, President ,AAPM

Name of Organization: American Association of Physicists in Medicine (AAPM)

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: Medical Physicist

Domain of Research Most Important to You or Your Organization:

radiation medicine

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

See attached comment letter

Section II: Definitions:

See attached comment letter

Section III: Scope:

See attached comment letter

Section IV: Effective Date(s):

See attached comment letter

Section V: Requirements:

See attached comment letter

Section VI: Data Management and Sharing Plans:

See attached comment letter

Section VII: Compliance and Enforcement:

See attached comment letter

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

See attached comment letter

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

See attached comment letter

Other Considerations Relevant to this DRAFT Policy Proposal:

See attached comment letter

Attachment:

AAPM Comment NIH Data Policy Final .pdf

Description:

Comment Letter of American Association of Physicists in Medicine



M. Saiful Huq, PhD, FAAPM, FInstP
Office of the President
UPMC Hillman Cancer Center and
University of Pittsburgh School of Medicine
5150 Centre Ave, Fifth Floor, Suite 542
Pittsburgh, PA 15232-1309
huqs@upmc.edu
412.647.1813

January 10, 2020

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

VIA Email: SciencePolicy@mail.nih.gov

RE: Request for Comment Draft NIH Policy for Data Management and Sharing and Supplemental Draft Guidance

Dear Sir or Madam:

The American Association of Physicists in Medicine (AAPM)¹ is pleased to submit comments to the National Institutes of Health (NIH) regarding its Draft NIH Policy for Data Management and Sharing and Supplemental

¹ The AAPM is the premier organization in medical physics, both in the U.S. and abroad. Medical physics is a scientific and professional discipline that uses physics principles to address a wide range of biological and medical needs. The mission of the AAPM is to advance medicine through excellence in the science, education and professional practice of medical physics. Currently, the AAPM represents over 9,000 medical physicists.

Medical physicists contribute to the effectiveness of medical imaging by ensuring the safe and effective use of radiant energy (e.g., optical, ionizing, ultrasonic, or radiofrequency) to obtain detailed information about the form and function of the human body. Medical physicists continue to play a leading role in the development of novel imaging technologies, as well as in guiding the optimization of existing imaging modalities. In addition, medical physicists contribute to development of new therapeutic technologies in radiation oncology, as well as in other disciplines, such as in thermal ablation or high intensity focused ultrasound. Clinically, medical physicists work side by side with radiation oncologists to design treatment plans and monitor equipment and procedures to ensure that cancer patients receive the prescribed dose of radiation at the correct location.

Draft Guidance. The AAPM commends the NIH on its work in advancing effective and efficient data management and sharing to maximize benefits from research efforts funded by the NIH.

General Comments

The AAPM believes this is an important NIH initiative. This policy review and update come at a critical time for data management and sharing where there is exponential growth in the amount of scientific data and an increasing need to leverage large data sets to advance research. The AAPM urges the NIH to maximize the value of data by taking a science-based approach to data sharing.

Draft Data Management and Sharing

The AAPM agrees that data produced by research that is publicly funded should be broadly and routinely shared. Data sharing improves the scientific process through independent verification, advances knowledge through efficient use of existing data, and promotes reproducible science. The AAPM, however, expresses its concern about the complexity of these plan requirements and the ability of principal investigators to successfully comply with these requirements.

We voice concern with requesting the data sharing plan as part of the just-in-time (JIT) information submission. Currently, resource sharing plans are evaluated during study section review and comments provided to the investigators to assist them in addressing potential weaknesses. Evaluation of the resource sharing plan, which contains a data sharing plan as one element, is not considered in the scoring of the proposal, but provides highly valuable information to the principal investigator. Proposals with potentially fundable scores that have weaknesses identified in the data sharing plan by the reviewers would benefit from an early request to the principal investigator to address such weaknesses. JIT information can be requested with relatively short response times, which may be inadequate for resolution of identified concerns. The JIT process represents the final step of the grant award process. Should a data sharing plan submitted at this late stage be found inadequate, the grant award process will be delayed, which can pose difficulties for not only the investigator and institution, but also the funding agency.

In addition, the AAPM is concerned with the statement: "The [data sharing] Plan will become a Term and Condition of the Notice of Award. Failure to comply with the Terms and Conditions may result in an enforcement action...". (See "Compliance and Enforcement"). The grant awards are in support of research, and plans will change during the execution of the research. Accordingly, we believe there should be a mechanism to support flexible alterations in the plan over time that is not punitive. For example, what will happen if the embargo period changes during the grant period? The AAPM asks whether the enforcement period could be phased in after several cycles, or after a pilot period.

The AAPM offers the following recommendations:

- Identify where NIH will keep any data that it gets and specify how the public will access these data. Alternatively, if NIH will require principal investigators to make the data available, specify how that will be accomplished and how the public will be able to search for and gain access to available data.
- Conduct a pilot program to test plan processes.
- Provide sample data sharing plans, like the sample NIH biosketches, to assist investigators in crafting data sharing plans.
- Consider including the data sharing plan as a formally peer-reviewed item to give the most highly-impacted stakeholders (i.e., the scientific community) greater input as to the acceptability of a plan².
- Ensure that adequate time is provided to the principal investigator to address weaknesses in the data sharing plan identified during the review process.
- Provide lists or examples of acceptable tools/resources for data de-identification and sharing, and specify whether generated repositories can be local, independent databases, or cloud-based.
- Devise a form to guide the principal investigator through the process to ensure that the principal investigator includes sufficient information. The form could include checkboxes or selection tools for major classifiers such as: Imaging, Modality, Approximate Number of Subjects, and Sequences.
- Include an embargo or delay on sharing data that enables researchers to publish their findings before handing data off to competitors. Differing data sets may require differing embargo periods and a mechanism is needed to allow flexibility in this regard.
- Implement a phased-in adoption. We believe the phase-in period would require the preparation and dissemination of educational materials, tools to de-identify data containing protected health information, standardized patient information and consent forms, tools to guide researchers on how

² This view is supported by Report of the Board of Scientific Advisors Ad Hoc Working Group entitled, "An Assessment of the Impact of the NCI Cancer Biomedical Informatics Grid"(caBIG) published in March 2011, but never implemented, Recommendation 10:

"Promote interoperability and data sharing by making them key review criteria for grant and cooperative agreement applications and R&D contracts and including them as requirements for award..."

(See <https://deainfo.nci.nih.gov/advisory/bsa/archive/bsa0311/caBIGfinalReport.pdf>.)

to complete relevant forms, and tools to ensure that all required information is provided in a data management and sharing plan.

- Develop guidance documents for institutional officials who authorize the submission of an application. Institutional grants management officials will need to be educated on the requirements of data sharing and management policies such that they can internally assess the adequacy of the plan contained in the proposal and ensure that the plan is consistent with institutional policies and procedures.

Supplemental Draft Guidance: Elements of NIH Data Management

We believe limiting the length of the data sharing plan to no more than two pages is unrealistic, and we urge you to consider allowing more than two pages so that principal investigators may properly present all of the information that is requested. We believe this will be particularly important if NIH makes the data sharing plan an essential part of the awarding criteria.

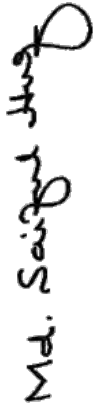
The AAPM recommends that NIH add greater clarity to this guidance document. For example, the words “broad” and “broadly” appear multiple times throughout the document, indicating a high degree of ambiguity in interpretation at different time points in a grant life cycle. We believe that without greater clarification, a number of provisions will likely cause problems in the review of the plan as well as in the implementation and interpretation of compliance.

The AAPM further recommends that more thought be given to the elements of data management. We believe the data sharing plan should be more structured, i.e., with sections, subsections, and sub-subsections with required information specifically laid out. As currently written, it is suggestive of the information that is sought, but does not clearly delineate what is required. For example, the text “Any other considerations that may result in limitations on the ability to broadly share scientific data” can be hard to interpret and does not help in identifying the appropriate materials for a response. (See paragraph entitled, “Data Sharing Agreements, Licenses, and Other Use Limitations”).

In summary, the AAPM supports NIH’s efforts to increase access to scientific data resulting from agency-funded research by crafting a viable, pragmatic policy for data management and sharing and urges NIH to assist investigators in navigating such a process. The AAPM hopes that the NIH will carefully consider the AAPM’s comments and adopt the AAPM’s recommendations when crafting the final policy.

Thank you for the opportunity to comment. If you have any questions or require additional information, please contact Richard J. Martin, JD, Government Relations Project Manager, at 571-298-1227 or Richard@aapm.org

Sincerely,



**M. Saiful Huq, PhD, FAAPM, FinstP
President, AAPM**

Professor of Radiation Oncology
Professor of Clinical and Translational Science Institute
Director, Division of Medical Physics

Department of Radiation Oncology
UPMC Hillman Cancer Center
University of Pittsburgh School of Medicine

Submission ID: 1361

Date: 1/10/2020

Name: Jaclyn Lucas

Name of Organization: Beckman Research Institution of the City of Hope

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role:

Role - Other:

Domain of Research Most Important to You or Your Organization:

biomedicine

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We thank the NIH for the opportunity to provide input on the Draft NIH Policy for Data Management and Sharing. We value NIH's commitment to promoting effective data management and sharing and to making results of NIH-funded research available to the public and scientific community. While recognizing the challenges inherent in developing a policy that fits the diversity of data generated by the biomedical research enterprise, and appreciating that the draft policy provides flexibility for investigators to design data sharing plans appropriate for their projects, there are several areas where we feel further guidance or clarification would be beneficial, as indicated below.

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Although different projects will have different timeframes in which it is reasonable to share data, nonetheless a timeframe by which data are required to be deposited (e.g., within one year of the end of the funding period) should be provided to ensure sharing occurs.

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:**Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:**

In some situations, it may not be appropriate to share scientific data or deposit it into a repository until after the end of the funding period or there may be recurring fees or costs (such as if data is locally managed) that would extend beyond the funding period. Guidance is needed on whether funds could be requested to continue to preserve and manage data or deposit it after the funding period, or if there would be mechanisms available to apply for funds to support these activities after the end of the funding period. Although the Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan indicates researchers should provide "[a]nticipated times frames for preserving scientific data..." without continuing financial support, there is risk that data will not be able to be preserved and shared for sufficiently long timeframes after the end of funding.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The draft guidance notes that "NIH does not expect researchers to share all scientific data generated in a study." However, clarification on the scope of data that would be expected to be shared would be beneficial. While the definition of "Scientific Data" in the draft policy suggests that data needed "to validate and replicate research findings" should be shared, there may be disagreement in research communities and fields on what constitutes this data. Perhaps examples such as currently provided under "Data Type" and "Standards" in the draft guidance could be included or further guidance given from the NIH Institutes, Centers, and Offices upon the publication of the final policy.

Other Considerations Relevant to this DRAFT Policy Proposal:

We would be curious if NIH is considering supporting new/additional data repositories to support data preservation, management and sharing. We recognize that NIH currently supports many data sharing repositories, but some of these repositories restrict data submission and given the diversity of data produced by NIH-funded research, additional repositories will likely be needed. Because development and maintenance of repositories to ensure long-term availability of data is costly, it seems appropriate that NIH facilitate sustained support of any additional repositories needed to ensure compliance with the final policy.

Attachment:**Description:**

Submission ID: 1362

Date: 1/10/2020

Name: Amazon Web Services

Name of Organization: Amazon Web Services

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other:

Type of Organization: Other

Type of Organization - Other: Cloud Services

Role: Member of the Public

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

The current language in the draft policy encourages the use of established repositories of scientific data. There are several concerns with this recommendation, as the current repositories do not always meet the data accessibility and protection requirements outlined in the draft policy. We recommend requiring new repositories and a data migration plan to move current datasets into new repositories. While these efforts will incur additional expenses, they are in line with related efforts to modernize the existing data infrastructure ecosystem.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Costs associated with effective data sharing can be substantial and include personnel to curate and convert data into appropriate archival formats, storage or storage services, and training researchers in the use of data sharing technologies. However, the data sharing plan is to be submitted just in time for extramural proposals, separate from budget line items. This may

result in questions during grant review about whether the budget is appropriate, especially if cloud services are included without explanation.

Further, data sharing is an increasingly important factor in evaluating a grant proposal's impact. Data shared as part of a large, harmonized repository that facilitates advanced analytics amplifies any impact for the project collecting it. This practice should be encouraged despite additional costs. Submission of a plan JIT for extramural proposals means that neither the data sharing plan nor its budget can be considered as part of the proposal.

We recommend that the budget for additional data management and sharing costs be considered in conjunction with the data sharing plan as an additional and separate component and that additional funding be allocated to support data sharing.

AWS recommends preserving and sharing data through established repositories. Unfortunately, many popular repositories lack programmatic access through application programming interfaces, do not enforce adequate metadata or documentation, require download of data to perform simple analytics on the content, and in many other ways do not represent the state-of-the-art in data warehousing and storage. Moreover, the state-of-the-art is constantly evolving. Failure to provide data accessibility features greatly limits the ability to perform analytics across multiple datasets, which is critical for scientific reproducibility and advancement.

Specific recommendations should be made that outline the qualities of a desirable repository, including programmatic access, specific metadata, documentation and data format standards, and remote query access. This will ensure the greatest usability for both the data and the associated repository.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

A suitable data repository should satisfy the following criteria:

- Highly available – Scientists can rely upon accessing the data.
- Highly durable – The chance of data loss is so small as to be negligible.
- Secure – Data (for example, PHI) is distributed only to those who should have access.
- Accessible programmatically – Pipelines can be built that do not need to download data to local infrastructure to process it
- Discoverable – Data exploration can be conducted without downloading.
- Integrable – The value of data increases nonlinearly with its size, as this allows study of individual differences and personalized medicine.

These characteristics are not typically achievable at a single institution, but are typically characteristics of public cloud platforms.

In addition to general descriptions of the data and volume, we recommend you consider requiring a complete data dictionary, including metadata, with information about the data pedigree and lineage. As there is a two-page limit, this should be delivered as an appendix. This data dictionary should include field names, data types, description and purpose, and relationships to other data items.

The current recommendations in the draft policy reference timelines that detail where and when data will be made available. It is also important to include information about the data history to ensure that the pedigree and lineage is preserved. Information about why data was collected, why it was formatted and stored in a particular format, and why specific fields were selected is critical to determining how the data can be used in future analyses. Staff turnover in particular can mean that this descriptive data is lost, requiring its collection ensures the information is recorded and available.

In addition to information about data sharing limitations, we recommend that you require information about how the data may or may not be combined with other data sources. Aggregate data analytics improve a researcher's ability to develop and test hypotheses; however, data can only be integrated in specific circumstances. These criteria should be outlined as part of the data management plan.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

AWS Response to NIH Request for Public Comments.pdf

Description:

AWS Response to NIH Request for Public Comments



Amazon Web Services Response to National Institutes of Health (NIH) Request for Public Comments for Data Management and Sharing

January 10, 2020

Submitted By:

Amazon Web Services, Inc.
12900 Worldgate Dr. Suite 800
Herndon, VA 20170

Cage Code: 66EB1
DUNS Number: 965048981
NAICS: 518210

Eric Egan
Senior Account Manager
ericegan@amazon.com

Submitted To:

Andrea Jackson-Dipina, Dr.PH, Director
of the Division of Scientific Data Sharing
Policy, Office of Science Policy, NIH

6705 Rockledge Drive, Suite 750,
Bethesda, MD 20892, 301-496-9838,
jacksondipinaac@od.nih.gov

Table of Contents

1.0	Executive Summary	1
1.1	Comments on Draft NIH Policy for Data Management and Sharing.....	1
1.1.1	Data Management and Sharing Plans	1
1.2	Comments on Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing	1
1.2.1	Curating Data and Developing Supporting Documentation	1
1.2.2	Preserving and Sharing Data through Established Repositories	2
1.3	Comments on Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan.....	2
1.3.1	Data Type	2
1.3.2	Data Preservation, Access, and Associated Timelines	3
1.3.3	Data Sharing Agreements, Licenses, and Other Use Limitations.....	3

1.0 Executive Summary

Amazon Web Services, Inc. (AWS) is pleased to respond with comments on the Draft NIH Policy for Data Management and Sharing.

AWS, as a leading cloud service provider, provides a highly reliable and scalable cloud infrastructure that is frequently used by both US federal agencies and researchers to share data for societal benefit and scientific advancement. We support both the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability ([STRIDES](#)) and [Open Data](#) initiatives, both designed to enable access to scientific datasets. Making data available to the research community, for use in related topics as well as independent validation and verification, is essential to accelerating research. However, scientific research is also a competitive business. Balancing these competing priorities is an important part of any data sharing policy. Our response provides AWS's perspective regarding advances in data management and sharing and how those advances can enhance data analysis and research efforts that support the American public.

1.1 Comments on Draft NIH Policy for Data Management and Sharing

AWS has provided comments on the data management and sharing plans below.

1.1.1 Data Management and Sharing Plans

The current language in the draft policy encourages the use of established repositories of scientific data. There are several concerns with this recommendation, as the current repositories do not always meet the data accessibility and protection requirements outlined in the draft policy. We recommend requiring new repositories and a data migration plan to move current datasets into new repositories. While these efforts will incur additional expenses, they are in line with related efforts to modernize the existing data infrastructure ecosystem.

1.2 Comments on Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

AWS has provided comments on curating data and developing supporting documentation and preserving and sharing data through established repositories below.

1.2.1 Curating Data and Developing Supporting Documentation

Costs associated with effective data sharing can be substantial and include personnel to curate and convert data into appropriate archival formats, storage or storage services, and training researchers in the use of data sharing technologies. However, the data sharing plan is to be submitted just in time for extramural proposals, separate from budget line items. This may result in questions during grant review about whether the budget is appropriate, especially if cloud services are included without explanation.

Further, data sharing is an increasingly important factor in evaluating a grant proposal's impact. Data shared as part of a large, harmonized repository that

facilitates advanced analytics amplifies any impact for the project collecting it. This practice should be encouraged despite additional costs. Submission of a plan JIT for extramural proposals means that neither the data sharing plan nor its budget can be considered as part of the proposal.

We recommend that the budget for additional data management and sharing costs be considered in conjunction with the data sharing plan as an additional and separate component and that additional funding be allocated to support data sharing.

1.2.2 Preserving and Sharing Data through Established Repositories

AWS recommends preserving and sharing data through established repositories. Unfortunately, many popular repositories lack programmatic access through application programming interfaces, do not enforce adequate metadata or documentation, require download of data to perform simple analytics on the content, and in many other ways do not represent the state-of-the-art in data warehousing and storage. Moreover, the state-of-the-art is constantly evolving. Failure to provide data accessibility features greatly limits the ability to perform analytics across multiple datasets, which is critical for scientific reproducibility and advancement.

Specific recommendations should be made that outline the qualities of a desirable repository, including programmatic access, specific metadata, documentation and data format standards, and remote query access. This will ensure the greatest usability for both the data and the associated repository.

1.3 Comments on Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan

A suitable data repository should satisfy the following criteria:

- Highly available – Scientists can rely upon accessing the data.
- Highly durable – The chance of data loss is so small as to be negligible.
- Secure – Data (for example, PHI) is distributed only to those who should have access.
- Accessible programmatically – Pipelines can be built that do not need to download data to local infrastructure to process it
- Discoverable – Data exploration can be conducted without downloading.
- Integrable – The value of data increases nonlinearly with its size, as this allows study of individual differences and personalized medicine.

These characteristics are not typically achievable at a single institution, but are typically characteristics of public cloud platforms.

1.3.1 Data Type

In addition to general descriptions of the data and volume, we recommend you consider requiring a complete data dictionary, including metadata, with information about the data pedigree and lineage. As there is a two-page limit, this should be delivered as an appendix. This data dictionary should include field names, data types, description and purpose, and relationships to other data items.

1.3.2 Data Preservation, Access, and Associated Timelines

The current recommendations in the draft policy reference timelines that detail where and when data will be made available. It is also important to include information about the data history to ensure that the pedigree and lineage is preserved. Information about why data was collected, why it was formatted and stored in a particular format, and why specific fields were selected is critical to determining how the data can be used in future analyses. Staff turnover in particular can mean that this descriptive data is lost, requiring its collection ensures the information is recorded and available.

1.3.3 Data Sharing Agreements, Licenses, and Other Use Limitations

In addition to information about data sharing limitations, we recommend that you require information about how the data may or may not be combined with other data sources. Aggregate data analytics improve a researcher's ability to develop and test hypotheses; however, data can only be integrated in specific circumstances. These criteria should be outlined as part of the data management plan.

Submission ID: 1363

Date: 1/10/2020

Name: Lisa Arafune

Name of Organization: Coalition for Academic Scientific Computation (CASC)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All types

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: Membership Organization

Domain of Research Most Important to You or Your Organization:

All academic research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

CASC endorses the purpose of the proposed policy. Standards are important to reach the next level of data driven science. This policy helps challenges within the NIH mission. While standards are crucial, it is recognized elsewhere in the Policy that it must strike a balance of enforcement to foster standards with flexibility to deal with innovation that is in the very nature of research. In our view this Policy successfully hits such a balance. There is a clear and helpful statement that the Plan may be amended during the period of funded research. This addresses the fact that research projects by their very nature cannot be fully planned out. The supplemental guidance is good and enables the research community to develop and reach standards without creating an inflexible framework.

Section II: Definitions:

The definitions provide adequate clarity of the terms used in the Policy, without becoming overly detailed and technical.

Section III: Scope:

The statement of scope is clear and unambiguous. As addressed elsewhere in the Policy, it is possible to provide details of the Data Management and Sharing plan just-in-time so that the burden of preparing the plan is minimal and can be postponed until after it is known that funding of the proposed effort is highly likely.

Section IV: Effective Date(s):

The different cases to consider when the Policy will take effect are outlined with adequate precision to cover the existing NIH funding mechanisms and processes.

Section V: Requirements:

The requirements of the Policy are clear and simple, while still allowing for the required flexibility needed in research as accomplished by stipulating compliance with the NIH ICO, which may include some negotiation, and by providing supplemental guidance on allowable cost.

Section VI: Data Management and Sharing Plans:

The Policy recognizes the need for a data management and sharing plan as a requirement but allows for exceptions in both what data and when the sharing is to occur. This is an important consideration in the context of research, whereby the very nature of the activity, not everything can be predicted. The explicit listing of Plan Elements and Plan Assessment is crucial to make the Policy clear so that compliance is possible without placing undue burden on the researchers.

Section VII: Compliance and Enforcement:

The Policy describes the process for reaching compliance. It also recognizes that providing data management and sharing services after the funding or support period may place a burden on the institution where the research was carried out by providing the possibility to include some of the cost in the project budget as described in supplemental guidance.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The guidance on allowable cost provides valuable advice on planning of and budgeting for the sharing process and associated support activities. We urge that in practice, as implementation of this policy evolves, the data management and sharing costs are regarded as important components of the budget, and that proposers and reviewers all support budgeting reasonable costs for data management and sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

By providing the details on data elements and related tools, standards, data preservation and access timelines, data sharing agreements and licenses, and accountable person(s) in a supplemental guidance document instead of in the body of the Policy, the Policy retains the necessary flexibility to adapt without undue burden to special cases that may, and will, arise in the context of research activities.

Other Considerations Relevant to this DRAFT Policy Proposal:

The CASC response was prepared by a committee chaired by Erik Deumens, University of Florida.

Attachment:

CASC Response to NIH Jan 2020.pdf

Description:

PDF of CASC Response to NIH Jan 2020



January 10, 2020

**Coalition for Academic Scientific Computation
Response to
National Institutes of Health**

DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidelines

The policy and supplemental docs are

[https://osp.od.nih.gov/wp-content/uploads/Draft NIH Policy Data Management and Sharing.pdf](https://osp.od.nih.gov/wp-content/uploads/Draft%20NIH%20Policy%20Data%20Management%20and%20Sharing.pdf)

[https://osp.od.nih.gov/wp-content/uploads/DRAFT Supplemental Guidance Allowable Costs.pdf](https://osp.od.nih.gov/wp-content/uploads/DRAFT%20Supplemental%20Guidance%20Allowable%20Costs.pdf)

[https://osp.od.nih.gov/wp-content/uploads/Supplemental DRAFT Guidance Elements NIH Data Management and Sharing Plan.pdf](https://osp.od.nih.gov/wp-content/uploads/Supplemental%20DRAFT%20Guidance%20Elements%20NIH%20Data%20Management%20and%20Sharing%20Plan.pdf)

Comments must be uploaded by Jan 10, 2020 using the portal at

<https://osp.od.nih.gov/draft-data-sharing-and-management/>

Chair: Sharon Brode Geva, University of Michigan • **Vice Chair:** Neil Bright, Georgia Institute of Technology
Secretary: Craig Stewart, Indiana University • **Treasurer:** Scott Yockel, Harvard University
Director: Lisa Arafune

1155 F St., NW, Suite 1050 • Washington, DC 20004 • (202) 930-2272 • <http://casc.org>



CASC Response

=====

Section I Purpose

CASC endorses the purpose of the proposed policy. Standards are important to reach the next level of data driven science. This policy helps challenges within the NIH mission. While standards are crucial, it is recognized elsewhere in the Policy that it must strike a balance of enforcement to foster standards with flexibility to deal with innovation that is in the very nature of research. In our view this Policy successfully hits such a balance. There is a clear and helpful statement that the Plan may be amended during the period of funded research. This addresses the fact that research projects by their very nature cannot be fully planned out. The supplemental guidance is good and enables the research community to develop and reach standards without creating an inflexible framework.

Section II Definitions

The definitions provide adequate clarity of the terms used in the Policy, without becoming overly detailed and technical.

Section III Scope

The statement of scope is clear and unambiguous. As addressed elsewhere in the Policy, it is possible to provide details of the Data Management and Sharing plan just-in-time so that the burden of preparing the plan is minimal and can be postponed until after it is known that funding of the proposed effort is highly likely.

Section IV Effective Dates(s)

The different cases to consider when the Policy will take effect are outlined with adequate precision to cover the existing NIH funding mechanisms and processes.

Section V Requirements

The requirements of the Policy are clear and simple, while still allowing for the required flexibility needed in research as accomplished by stipulating compliance with the NIH ICO, which may include some negotiation, and by providing supplemental guidance on allowable cost.

Chair: Sharon Broude Geva, University of Michigan • **Vice Chair:** Neil Bright, Georgia Institute of Technology
Secretary: Craig Stewart, Indiana University • **Treasurer:** Scott Yockel, Harvard University
Director: Lisa Arafune



Section VI Data Management and Sharing Plans

The Policy recognizes the need for a data management and sharing plan as a requirement but allows for exceptions in both what data and when the sharing is to occur. This is an important consideration in the context of research, whereby the very nature of the activity, not everything can be predicted. The explicit listing of Plan Elements and Plan Assessment is crucial to make the Policy clear so that compliance is possible without placing undue burden on the researchers.

Section VII Compliance and Enforcement

The Policy describes the process for reaching compliance. It also recognizes that providing data management and sharing services after the funding or support period may place a burden on the institution where the research was carried out by providing the possibility to include some of the cost in the project budget as described in supplemental guidance.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

The guidance on allowable cost provides valuable advice on planning of and budgeting for the sharing process and associated support activities. We urge that in practice, as implementation of this policy evolves, the data management and sharing costs are regarded as important components of the budget, and that proposers and reviewers all support budgeting reasonable costs for data management and sharing.

Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan (Plan)

By providing the details on data elements and related tools, standards, data preservation and access timelines, data sharing agreements and licenses, and accountable person(s) in a supplemental guidance document instead of in the body of the Policy, the Policy retains the necessary flexibility to adapt without undue burden to special cases that may, and will, arise in the context of research activities.

Other Considerations Relevant to the DRAFT Policy Proposal

The CASC response was prepared by a committee chaired by Erik Deumens, University of Florida.

Chair: Sharon Broude Geva, University of Michigan • **Vice Chair:** Neil Bright, Georgia Institute of Technology
Secretary: Craig Stewart, Indiana University • **Treasurer:** Scott Yockel, Harvard University
Director: Lisa Arafune

Submission ID: 1364

Date: 1/10/2020

Name: Anurupa Dev

Name of Organization: Association of American Medical Colleges

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: N/A

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role:

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The AAMC concurs with the NIH's assertion that increased access to research data advances biomedical research by enabling further validation of scientific results, facilitating reuse of hard-to-generate data, catalyzing new research, and generally promoting more responsible stewardship of federal resources. These advantages can be realized through meaningful data sharing and the development of community-wide norms, as well as ensuring that accessed data are used for the advancement of discovery and in furtherance of rigorous scientific discourse.

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

The NIH should institute an implementation timeframe that allows for researchers to fully understand the policy requirements, and for institutions to develop the necessary training, resources and infrastructure. Because this is such a wide-ranging policy that impacts the way research is conducted and includes every NIH-funded investigator and project, we recommend a minimum implementation date of one year after the release of the final policy, with a delay in enforcement actions for at least one year after the implementation deadline. Any determination of non-compliance should follow well-defined and transparent criteria.

Section V: Requirements:

The draft policy states that researchers must submit a Data Management and Sharing Plan (hereinafter "Plan") to NIH, as well as comply with the final NIH ICO (Institute, Center, and Office)-approved Plan, leading many in the research community to the concern that over time there may be 27 different data sharing policies at the NIH for which investigators are responsible, an overarching policy and one from each ICO. While ICOs may have additional expectations for data management and sharing above the base NIH policy, these additional ICO requirements should be narrow and rare, with a priority placed on standardization across the agency whenever possible. Further, NIH should make publicly available the process (or at a minimum, basic criteria) ICOs will use to establish these requirements and explain the role of the Data Science Governance Council in making these decisions. We understand the need for special, large-scale projects to have specific data management and sharing requirements, but stress that these should be put into place through a transparent and deliberate process.

In order for researchers to develop an effective and executable Plan, there should be clarity about the evaluation criteria and the assessment that will be used by program officers. If NIH can make public any relevant tools that program officers are using, that would be very helpful to the research community and future applicants. We also recommend that any guidance provided to program officers regarding requirements for or evaluation of the adequacy of data management and sharing plans be developed in collaboration with external experts. With the Plan submission proposed to be submitted as a Just-in-Time requirement, the Plan will no longer receive feedback from peer reviewers with expertise in the field and instead will be added to the application after the researcher initially creates the grant budget. As the policy is implemented, the agency should evaluate if this is the most effective timing for submission of the Plan. As each ICO will be responsible for communicating with the researchers as they develop and comply with their Plans, there should be clear points of contact at each ICO for questions, including where researchers can go for assistance if they are unable to receive it from their designated Program Officer.

The policy currently states that NIH may make data management and sharing plans publicly available. It is critical that a Plan functions to help researchers manage data and clearly lay out their obligations to the agency. Given that researchers may have hesitations about making these plans publicly available, including concerns about privacy or progress of the research, the agency should consider an embargo period or exceptions for this requirement. However, there are also clear benefits to making data management plans broadly available, in allowing researchers and the public to be able to find the data associated with a particular grant, as well as provide examples of effective Plans to NIH investigators. We recommend that the NIH find a mechanism to make the data location element and when/whether the data will be shared publicly available (e.g. one or more dataset PIDs included as part of a RePORTER listing), and

secondarily that the NIH commit to creating and making available a collection Plans that have been submitted to and approved by the agency.

Section VI: Data Management and Sharing Plans:

The AAMC suggests that NIH define clearly a set of minimum requirements that researchers should include in the Plan submitted to the agency. The draft supplemental guidance on elements of a Plan currently contains a number of options for researchers to include in a Plan, with no indication of the relative importance or hierarchy of these elements. While it is understandable that unique projects may have different priorities and needs for inclusion in a Plan, we recommend that the policy define minimum requirements for researchers to include in a Plan, such as data type, standards and metadata, plans for data preservation, and projected data accessibility. Researchers should also be required to indicate in the Plan whether the project will involve data derived from human participants or specimens, and if so, include strategies for maintaining privacy, rights, and confidentiality. In the absence of sample templates and/or further guidance for the level of detail, each institution (or researcher) will create their own guidelines and tools, which may or may not meet the objectives of the policy. Providing greater guidance about the expected content of a Plan will better serve both the researchers assembling the document and the goals of the agency.

We would also recommend that the agency reconsider the currently proposed limit of 2 pages for a Plan. Many researchers who actively practice data sharing and frequently prepare DMPs have suggested to us that this length is insufficient to include all of the necessary information for the Plan to be a useful document with the appropriate level of detail. We suggest that NIH increase this limit to 4 pages and, in its ongoing evaluation of the policy's impact and effectiveness, determine whether this is an appropriate limit after the policy goes into effect.

There are a number of resources the agency will need to develop to facilitate researchers both creating and implementing a Plan. We recommend that NIH create and maintain an online clearinghouse that lists data elements and metadata for common data types for which best practices exist in the scientific community, as well as other existing resources such as DMPTool. The development of this policy presents an opportunity to amplify and disseminate efforts for good data management and sharing, particularly for certain disciplines, such as neuroimaging, or data types, such as microarrays or sequencing, which have well-defined standards and formats. This will provide a basis for standardization in the data that is submitted by researchers, and hopefully increase the usability of NIH-funded data.

We recommend that NIH identify key characteristics of suitable data repositories and additionally provide lists of accepted repositories for scientific disciplines where they have been

well established (see current efforts from Springer Nature/FAIRsharing/DataCite). In order to meet the presumed expectations that most or all data from NIH-funded research will need to be stored and made available for others to use, many institutions are planning to expand and use their own repositories. Without guidance from the agency on standards for data storage and discoverability as well as some level of centralized infrastructure or coordination, holding data in such disparate platforms and systems will place a significant technical burden on anyone who wants to reuse the data, thwarting the agency's laudable goals to increase and improve data re-use.

We appreciate that the draft policy acknowledges that valuable data are not always used to support a scholarly publication—this understanding is essential to recognizing data as a first-class research object and promoting a data-centric model of research. We also agree that investigators should have the opportunity to provide a rationale for decisions about which scientific data will not be made available for sharing. In order to accommodate this flexibility and also push forward the desired result of increasing sharing, the agency could consider setting a baseline for data that should be shared, such as the minimum underlying data to replicate and validate published findings, while still providing the researcher the ability to justify whether or not this is reasonable for a given study.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Data management and preservation will require significant infrastructure investment on the part of the institution; however, the allowable costs as currently defined specifically exclude infrastructure costs typically included in institutional overhead. We would recommend that if these costs are not permitted on a grant-by-grant basis, that the agency offers additional supplemental funding to institutions to develop this infrastructure.

The guidance on costs also should have additional clarity around what constitutes an "established repository," and particularly whether institutional repositories may fit this role and be included in the grant budget. While costs for deposition and storage in an established and/or commercial repository may be more well-documented, it can be difficult to define the costs for an institutional resource in the same way. The current statement that researchers can request funds for "unique and specialized information infrastructure" would benefit from examples on what this includes.

Increasing data management and sharing activities often requires significant support from personnel outside of the traditional laboratory environment, including librarians and data

scientists, to provide the necessary expertise and guidance needed to comply with a data sharing policy and build good data management practices into an investigator's research process. NIH should strongly consider including these additional staff as part of the allowable costs. Again, if this is not doable, it will be necessary for the agency to provide supplemental funding to institutions in building up and maintaining services that support scientific data sharing.

Finally, the draft guidance does not instruct grantees on what happens after a grant period comes to an end and whether additional funding would be available at this juncture, when much of the data preservation and storage will take place. It is critical that the agency specify how it plans to support these costs that will occur after the normal grant period ends and indicate whether there will be additional funding available specifically for this purpose.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The AAMC suggests that NIH define clearly a set of minimum requirements that researchers should include in the Plan submitted to the agency. The draft supplemental guidance on elements of a Plan currently contains a number of options for researchers to include in a Plan, with no indication of the relative importance or hierarchy of these elements. While it is understandable that unique projects may have different priorities and needs for inclusion in a Plan, we recommend that the policy define minimum requirements for researchers to include in a Plan, such as data type, standards and metadata, plans for data preservation, and projected data accessibility. Researchers should also be required to indicate in the Plan whether the project will involve data derived from human participants or specimens, and if so, include strategies for maintaining privacy, rights, and confidentiality. In the absence of sample templates and/or further guidance for the level of detail, each institution (or researcher) will create their own guidelines and tools, which may or may not meet the objectives of the policy. Providing greater guidance about the expected content of a Plan will better serve both the researchers assembling the document and the goals of the agency.

We would also recommend that the agency reconsider the currently proposed limit of 2 pages for a Plan. Many researchers who actively practice data sharing and frequently prepare DMPs have suggested to us that this length is insufficient to include all of the necessary information for the Plan to be a useful document with the appropriate level of detail. We suggest that NIH increase this limit to 4 pages and, in its ongoing evaluation of the policy's impact and effectiveness, determine whether this is an appropriate limit after the policy goes into effect.

There are a number of resources the agency will need to develop to facilitate researchers both creating and implementing a Plan. We recommend that NIH create and maintain an online

clearinghouse that lists data elements and metadata for common data types for which best practices exist in the scientific community, as well as other existing resources such as DMPTool. The development of this policy presents an opportunity to amplify and disseminate efforts for good data management and sharing, particularly for certain disciplines, such as neuroimaging, or data types, such as microarrays or sequencing, which have well-defined standards and formats. This will provide a basis for standardization in the data that is submitted by researchers, and hopefully increase the usability of NIH-funded data.

We recommend that NIH identify key characteristics of suitable data repositories and additionally provide lists of accepted repositories for scientific disciplines where they have been well established (see current efforts from Springer Nature/FAIRsharing/DataCite). In order to meet the presumed expectations that most or all data from NIH-funded research will need to be stored and made available for others to use, many institutions are planning to expand and use their own repositories. Without guidance from the agency on standards for data storage and discoverability as well as some level of centralized infrastructure or coordination, holding data in such disparate platforms and systems will place a significant technical burden on anyone who wants to reuse the data, thwarting the agency's laudable goals to increase and improve data re-use.

We appreciate that the draft policy acknowledges that valuable data are not always used to support a scholarly publication—this understanding is essential to recognizing data as a first-class research object and promoting a data-centric model of research. We also agree that investigators should have the opportunity to provide a rationale for decisions about which scientific data will not be made available for sharing. In order to accommodate this flexibility and also push forward the desired result of increasing sharing, the agency could consider setting a baseline for data that should be shared, such as the minimum underlying data to replicate and validate published findings, while still providing the researcher the ability to justify whether or not this is reasonable for a given study.

Other Considerations Relevant to this DRAFT Policy Proposal:

The AAMC supports NIH's efforts to integrate data management into the research review and funding process, to increase sharing and re-use of scientific data generated through NIH-funded research, and to develop a clearly defined policy to accomplish these objectives. In addition to responding to the specific areas for which NIH has requested information, AAMC provides the following high-level comments on the draft policy:

- As NIH moves forward in the policy development process, we encourage the agency to consider the type of policy that will lead to meaningful and positive, rather than compliance-

based, data management and sharing practices. When deciding how proscriptive to make the policy's requirements, NIH's focus should be on feasibility of consistent implementation and on encouraging the sharing of data that are scientifically valuable, discoverable and reusable.

- The agency can further incentivize the goal of increased data sharing through encouraging the use of persistent identifiers (PIDs) so researchers can track and receive credit for their data, as well as issuing funding opportunities focused on data reuse.
- It is critical to have as much as harmonization and standardization as possible across the NIH in both the policy requirements and implementation. This includes all grantees as well as consistency in evaluation of compliance and in institute-specific requirements.
- We appreciate that the draft policy does not require researchers to share all scientific data, since requiring the sharing of all data without considering its usefulness or likelihood of re-use does not contribute to scientific progress and would constitute a substantial burden on the researcher and institution.
- Given the scope of this new policy, incorporating flexibility is appreciated by the research community. However, throughout the draft policy there are many optional elements and very few requirements, which may lead to overcompliance or an ineffective or inconsistently implemented policy.
- If NIH or the ICOs have specific but unstated expectations for any aspects of data management and sharing, such as what types of data should always be shared, how accessible that data should be, or a timeline for data sharing, those expectations should be included in the policy or otherwise explicitly stated.
- Successful implementation of this policy will require additional resources from both the NIH and grantee institutions. In addition to these resources, grantees will need substantial guidance from the NIH.
- We understand that NIH is intending to undertake ongoing evaluation of the costs and impact of this policy as implemented. We encourage NIH to treat the implemented final policy as a robust pilot initiative and recommend that a strong and detailed statement regarding the evaluation and revision process be included in the policy itself.

In AAMC's response to the proposed key policy provisions, we noted that a policy alone will not be sufficient to reach the stated goal of increasing scientific data sharing, and that the agency must provide "adequate training, education, and guidance, increasing available financial resources, and leading the development of tools and infrastructure in order to enable and facilitate policy implementation." The research community has expressed concern about the lack of clarity regarding which resources NIH will provide to implement this policy, including options for data storage and additional funding mechanisms. It is important to acknowledge

that the significant culture change that will be required in a move to a data sharing ecosystem will involve many factors, such as incentives and community support, in addition to any policy or mandate.

We appreciate NIH's intent to create a policy that is flexible, responsive to researcher feedback, and able to keep pace with the state of biomedical science. Plans to evaluate the impact of the policy should be described and implemented prior to its effective date to align agency and community expectations about the metrics that will be evaluated. A feedback loop between the agency and researchers, and clear communication are key, but without detailed guidance from the NIH, there will be a wide range of interpretations and policy implementation that doesn't necessarily serve the end goal. As NIH develops this guidance, we encourage the agency to refer to established criteria and policies from other funders and federal agencies, journals, and scientific societies, as well as consider the impact of any given requirement on how institutions are already complying with existing NIH policy. AAMC would be happy to work with institutions to provide the agency with examples of how they are affected by and complying with varying NIH policies.

Finally, as the policy is put in place, NIH should engage specifically with working groups consisting of researchers who generate data, librarians and other data science support at institutions, and labs that have research programs based on sharing and re-using scientific data, to ensure that the policy is responsive to the needs and concerns of different stakeholders and supports the scientific community as effectively as possible while meeting its desired goals. AAMC would be glad to assist in identifying these partners from our member medical institutions.

Attachment:

Description:

AAMC letter in response to NIH draft policy

Submission ID: 1365

Date: 1/10/2020

Name: David Carr

Name of Organization: Wellcome Trust

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All biomedical

Type of Organization: Other

Type of Organization - Other: Research funder - foundation

Role: Other

Role - Other: Funder

Domain of Research Most Important to You or Your Organization:

Biomedical science, humanities and social sciences

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The Wellcome Trust is a global research foundation dedicated to improving health for everyone by helping great ideas to thrive. Like NIH, Wellcome is an advocate and champion of open research, and is committed to working with other funders to maximise the value of research outputs. Also in common with NIH, Wellcome has had a long-standing policies requiring that the researchers we fund maximise the availability of research data and other research outputs with as few restrictions as possible – ensuring that these outputs can be accessed and used in ways that will advance research and its application to improve health.

In 2017, Wellcome published an updated policy on managing and sharing data, software and materials (<https://wellcome.ac.uk/funding/guidance/data-software-materials-management-and-sharing-policy>) – which extended our long-standing policy on data management and sharing to also cover research software and materials.

We fully support the purpose and goals of NIH's new draft policy. We wanted to share a few specific comments based on our experience of implementing our own policy over the years and highlight a few areas where we would be keen to build on our existing partnerships with NIH to share experience and good practice.

Section II: Definitions:

These looked largely appropriate. Either as part of the definition of scientific data, or elsewhere, we think it would be worth explicitly encouraging researchers to share null and negative results as well as results underpinning research findings. This could usefully highlight some areas where there is a particular imperative to do so – such as clinical trials and studies involving the use of animals.

Section III: Scope:

NIH should give serious consideration to including original software outputs alongside research data in the scope of the policy. In our view, research data and software are inextricably linked – and both equally vital in enabling other researchers to scrutinise and replicate research findings. We would argue that there is value in encouraging researchers to think about these two key outputs of research together, and plan for how they will maximise their value.

In expanding the scope of Wellcome's own policy to cover software and materials, as well as data, Wellcome moved to requiring an "outputs management plan" rather than a data management and sharing plan. We wanted the researchers we fund to consider their outputs holistically – we believe this approach has had value for our researchers, as well as ensuring the value of a range of outputs is recognised.

Section IV: Effective Date(s):

No comments

Section V: Requirements:

We would recommend a specific requirement that research data and software underpinning published research articles is made available to other researchers at the time of publication. There should be an expectation that these are made open wherever they can, recognising that in some cases there will need to be controls and limits on access. At the very least, we'd suggest requiring that all original research papers resulting from NIH funding should have a clear statement indicating how underlying data and code can be accessed by others.

Section VI: Data Management and Sharing Plans:

We felt several elements of this section and the guidance should be stronger and more specific expectations. In addition to requiring the sharing of data at the point of publication, we would suggest that the use of recognised community repositories (where they exist for a particular data type and including those that the NIH directly operate) should be expected rather than encouraged. Similarly, the use of persistent identifiers for data could usefully be much more strongly encouraged given their core importance in discoverability.

It was not clear to us from this section or the supplementary guidance whether NIH is proposing to introduce a formal template for data management and sharing plans. While this is not something Wellcome has done to date, it might be worth considering as a basis to ensure plans can be assessed by programme staff on a more consistent basis and that some pieces of key information (such as the chosen repository) are readily identifiable, and that plans can be machine-actionable.

We were interested to see that NIH may consider making plans public - this is something we are also keen to explore and we would be keen to discuss NIH's plans further as they take shape.

Section VII: Compliance and Enforcement:

We think the approach is right in principle and this is another area in which we'd be keen to stay in touch and share experience and good practice. Our experience in this space would suggest that monitoring compliance with the original plan (even if you successful in getting researchers to actively update their plans as they proceed), could be challenging and potentially resource-intensive.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

In addition to our comments above, we wondered whether the guidance should refer to the perceived tension between data sharing and protection of intellectual property. Our policy is that researchers should adopt the approach that will maximise health benefit – where this involves securing intellectual property protection, this can be a legitimate reason to limit or delay data sharing and this should be justified in the plan.

NIH may wish to consider setting criteria for what constitutes an appropriate repository or standard, or otherwise link to authoritative resources (for example, FAIRsharing.org) that can guide researchers to options for their data type and field.

We'd suggest that for data which require controlled access mechanisms, description of those mechanisms and how they will ensure legitimate requests for access are granted should be a core element of the plan, and not just something that researchers should consider.

The section on agreements and licenses seems to stop short of recommending that researchers apply a suitable license to their data.

Finally, and very importantly, consideration of the resources required should be a core element of the plan itself.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We think this guidance is very important. Our experience would be that many do not adequately plan or account for the costs of data sharing and it is important to give clear guidance on the types of costs they could consider. While useful, our view is that this guidance should be further developed to highlight potential cost areas and whether these can be requested as part of a funding application. For example, the guidance is not clear on whether support from specialist data managers and data scientists can be included as a cost.

Other Considerations Relevant to this DRAFT Policy Proposal:

We would suggest that the policy should refer to the responsibilities of data users and data generators. While data generators are expected to make data available to potential users in line with the FAIR principles, data users have a core responsibility to use the data in accordance with the terms under which it was accessed and to acknowledge the data generator appropriately, and in line with good practice for data citation and other community norms. NIH could consider how it will encourage good practice in data use (as well as sharing) among the researchers it supports.

Alongside the policy, NIH should actively consider how it will incentivise researchers to make their data available and recognise those who do it well - including, for example, actively encouraging researchers to highlight the generation and sharing of high quality datasets as a core research output and instructing reviewers to take these into account in line with the principles of DORA and related declarations.

Attachment:**Description:**

Submission ID: 1366

Date: 1/10/2020

Name: Andrew Smith

Name of Organization: ELIXIR

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All forms of bioinformatics data

Type of Organization: Other

Type of Organization - Other: Intergovernmental organisation - research infrastructure

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

All application areas of bioinformatics data

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

This submission represents the response of ELIXIR Europe, the pan-European research infrastructure for biological data. ELIXIR is the initiative to coordinate, sustain and integrate Europe's life science bioinformatics resources, providing a platform for scientific discovery in the life sciences.

ELIXIR is a distributed infrastructure with a central Hub – with the primary function of coordination - located on the Wellcome Genome Campus, Hinxton, and national Nodes – with the primary function of service delivery - in each participating Member State across Europe. The following countries and EMBL are Members of ELIXIR: Belgium, Czech Republic, Denmark, Estonia, France, Finland, Germany, Greece, Hungary, Ireland, Israel, Italy, Luxembourg, the Netherlands, Norway, Portugal, Slovenia, Sweden, Switzerland, Spain and the UK. Cyprus is an Observer.

This submission is aligned with the institutional response of EMBL-EBI, which is an ELIXIR Node, and the submission of the Global Biodata Coalition, of which ELIXIR is a partner.

ELIXIR welcomes the opportunity to submit comments on the Draft NIH Policy for Data Management and Sharing. The main suggestion that individual or institutional recipients of NIH-funding should be encouraged to develop a Data Management and Sharing Plan is supported by ELIXIR.

The stress placed on the 'sharing' of data, in addition to 'data management', is a welcome inclusion, as is the statement on NIH's encouragement towards the FAIR principles.

Section II: Definitions:

ELIXIR supports the definitions in the corresponding section and has no comments to add.

Section III: Scope:

We support the proposed broad scope of the policy in applying to all 'NIH-funded research that results in the generation of scientific data', regardless of the size of the grant.

Section IV: Effective Date(s):

We have no specific suggestions to make in relation to the date by which the policy should become effective. However, once agreed dates have been established, efficient communication of these (along with any updates to the policy) to various NIH stakeholders will be necessary to ensure effective implementation by recipients of NIH funding.

Section V: Requirements:

It would help to have further information here on possible reasons and circumstances behind acceptable exemptions to the policy.

Section VI: Data Management and Sharing Plans:

ELIXIR recommends that the policy should go beyond published data to include all aspects of scientific data generated from the project.

Good data management also needs to consider analytics tools (e.g. the programming codes used to analyse the data) along with the standards used to describe the data's metadata, to help ensure that the data remain useful in the long-term.

ELIXIR welcomes the references to needing appropriate plans and strategies concerning sensitive data and supports the emphasis placed on securing identifiable data.

Section VII: Compliance and Enforcement:

This section is vague in relation to consequences of not complying. It would be useful to have a high-level statement in relation to 'enforcement actions' that may result from not complying, with the caveat that exact consequences may depend on the type of funding scheme.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We welcome the inclusion of additional supplementary guidance relating to allowable costs (notably in relation to fees for commercial repositories), but wonder if this may encourage funding recipients to use these rather than other available options. Indeed, we would prefer a greater emphasis on the fact that most data management needs are already catered for by public-, and otherwise non-commercially, funded resources, which bear no cost (to the user) at the point of use.

In fact, a number of these internationally renowned and widely open resources receive significant funding from the NIH. Our recommendation would be to point users to lists/compilations/registries of recommended services and resources, to help ensure compliance with the policy. One notable example would be ELIXIR's Deposition Databases (<https://elixir-europe.org/platforms/data/elixir-deposition-databases>), which includes many resources that are part-funded through NIH such as Metabolites.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

As previously stated, explicitly recommending the use of appropriate publicly-funded resources (for example, for data deposition) will make it easier for awardees to deposit their data in the relevant database, thereby supporting reuse.

Beyond data deposition, resources that support interoperability are a crucial component of good data management practice, so referencing ELIXIR's Recommended Interoperability Resources (<https://elixir-europe.org/platforms/interoperability/rirs>) or other such recommendations may also be helpful.

In relation to Section 3 "Standards", we note that the example pool of existing standards is rather narrow. We suggest referencing the FAIRsharing registry (<https://fairsharing.org>) to help researchers to find the standards relevant to them and those that are also implemented by the repositories.

The FAIR principles are embedded within the text of the guidance, though not spelled out as explicitly as they could be. We suggest that the scope of the data sharing management plan

includes aspects of 'research sustainability': e.g. considerations of (1) will tools used in data processing and analysis during/after the project remain available under clear stated terms of use? (2) are the temporal evolutions of data, metadata, methods of data generation and analysis adequately addressed? (3) are standards that persist in the long-term preservation plan being sufficiently considered?

Other Considerations Relevant to this DRAFT Policy Proposal:

In addition to research data, software is a crucial output and public good from research projects. Whilst separate NIH consultations have covered the subject of software, ensuring that there is synergy between recommendations on research software and the final data management policy will be important.

ELIXIR is grateful to the NIH for providing financial support to many open bioinformatics resources, thereby helping to ensure that international research efforts are more efficient and useful to many. It is our hope that the data management plans in the context of NIH funded projects will complement this existing effort, rather than leave too much freedom to data producers in terms of what they do with data funded by the public purse.

Attachment:

Description:

Submission ID: 1367

Date: 1/10/2020

Name: Sarah Greene

Name of Organization: Health Care Systems Research Network

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Epidemiology, Health Services, Comparative Effectiveness, Behavioral Science,
Pharmacovigilance, Genomics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Statement in Response to NIH RFI - Data Sharing Jan 2020 final.docx

Description:



**Statement in Response to NIH Request for Information:
Department of Health and Human Services (HHS) 84 FR 60398:
DRAFT NIH Policy for Data Management & Sharing and Supplemental DRAFT Guidance**

On behalf of the Health Care Systems Research Network, we offer the following input on the recently released draft NIH Policy for Data Management and Sharing, including Supplemental Draft guidance on allowable costs, and recommended elements of a data sharing plan.

As background, the Health Care Systems Research Network (HCSRN) is a voluntary coalition of 18 members, each of which is a research center embedded in a health care delivery setting. Established in 1994, the HCSRN has a long history of research collaboration. Our researchers are fortunate to be able to utilize health care data generated by the members and patients who receive care from the participating systems. We are prudent stewards of the data we use in our studies, including primary data collected from patients themselves, and secondary data captured and stored in electronic health records and insurance claims in the course of patient care. Moreover, all of the HCSRN members are committed to placing findings in the public domain research so that our work benefits the greater good. To this end, we seek to balance responsible data use, management and sharing; safeguarding our patients' trust in their health care providers and systems; and ensuring that our research can benefit the broader population.

We fully support the goals and principles of the NIH Policy for Data Management and Sharing and agree with the intention to maximize the utility and usability of data collected under the auspices of NIH funding. We offer the following additional points for consideration.

1. Acknowledging Differences across Key Sources of Data Used in Research

Myriad sources of data are now available to researchers, including routinely collected data stored in electronic health records, and information from wearables, smartphones, and devices. By nature, data from these sources are different from prospectively collected experimental data for which participants authorize data sharing through their informed consent. Moreover, data collected and held by health systems may entail different constraints imposed by the systems, based on legitimate proprietary, security/re-identifiability, data ownership, and other business concerns. For many health systems, it would be untenable to agree to participate in a study, if doing so meant committing to unspecified future uses of data.

Given that electronic health data is qualitatively different from experimental data with explicit participant agreement to sharing, we underscore the importance of Section 5 of the Supplemental Draft Guidance for Elements of a Data Sharing and Management Plan. We encourage NIH to explicitly acknowledge that data sharing plans can impose a variety of restrictions, such as a requirement that secondary analyses be performed in a data enclave controlled by the original data holders, and that the original data holders be allowed oversight of the kinds of secondary analyses performed. Finally, we encourage NIH and others in the research community to leverage newer privacy-preserving analytic techniques, including distributed analysis methods. By design, these methods avert the need to create patient-level datasets or export large quantities of EHR data. Toh and others have developed principles and practices for sharing the minimum necessary data to perform analyses with precision (see: <https://doi.org/10.1097/MLR.000000000000147>)

2. Collaborative Research, the Revised Common Rule, and Informed Consent:

Beginning January 20, 2020, the revised Common Rule stipulates that for collaborative, multisite research, use of a single IRB will be required, unless otherwise properly justified and approved by the NIH Institute/Center Official. We support the efficiency and administrative streamlining that will result from this new requirement. However, we note that an IRB of record may be approving collaborative research studies on behalf of multiple sites, and these studies may have complex data use, sharing and management plans for their multisite context. There may be individual variation in site-specific data sharing policies. While an IRB of record for a collaborative study will have some cognizance of site-specific considerations around disclosing data, that IRB may be making assessments about the sufficiency of data sharing language in a protocol or consent form that might not fully represent the array of concerns and requirements of all study sites. Hence, additional guidance to IRBs, jointly prepared by NIH and OHRP, would be a useful adjunct to the NIH Policy on Data Management and Sharing.

The revised Common Rule also stipulates concision and readability of informed consent documents. Helping study participants understand future/secondary uses of data, once data are shared in an enclave, public use data set, or other format, will be imperative. That said, researchers who deposit and share data via a repository or enclave may not be fully able to anticipate all future uses or future users. Thus, it could be an opportune moment for NIH to launch a public education effort regarding the importance of research and health data, and how broader sharing of research data will accrue benefits to the general public.

3. Timing and Costs of Support for Data Sharing after Study Funding has Concluded

Health system data from electronic health records and administrative claims are increasingly used for research, but are not “research ready” at the time they are entered into an electronic health record. Understanding nuances related to data provenance, quality, and validity is its own robust part of the research process. In our Network, we have developed and deployed a common data model—the HCSRN Virtual Data Warehouse—to support data standardization, curation and quality assurance. Data documentation is an integral part of our research infrastructure. Over the HCSRN’s 26-year history, our data analysts have honed expertise about data provenance, quality, and evolution (e.g., changes to underlying native data that can have implications for their use and interpretation in research). Interpretation issues are critical, as is local knowledge. In our experience, new idiosyncrasies in complex multi-year data sets are sometimes discovered during the performance of secondary analyses.

Clear processes and support for making updates to data and metadata will be a critical aspect of executing this policy successfully, and as such, we appreciate that this NIH policy includes supplemental guidance regarding “Allowable Costs for Data Management and Sharing.” Given that such discoveries may occur after funding has ended, we urge NIH to consider simplified administrative supplements or other streamlined mechanisms to support study staff in making substantive updates to datasets, metadata and related documentation.

4. Affirming Public Trust in Research Data Use and Sharing

Public attitudes toward science, data, ownership, privacy, and security should be a paramount concern for everyone working in biomedical research. Data breaches that affect one of us affect all of us. As data sharing policies and processes become more widespread and ingrained in the

entire research enterprise, it will be imperative to ensure that the general public is educated on the importance of data to advance human health and improve health care.

At an August 2019 meeting of the National Academy of Medicine's Clinical Effectiveness Research Innovation Collaborative, stakeholders—including researchers, patients/families, funding agencies, and health system leaders—recommended taking a specific action to convene a national task force that would publicly affirm a set of principles and commitments on the collective benefits of data as a public good. We hope that this NIH Policy can also galvanize a national conversation on the vital need to share data to accelerate progress and maximize our collective investment.

HCSRN Statement in Response to NIH RFI - Data Sharing & Management

Submission ID: 1368

Date: 1/10/2020

Name: Barbara E. Bierer MD

Name of Organization: Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard (MRCT Center)

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: Health Care Delivery Organization

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Clinical research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

MRCT Center NIH Data Sharing Comments 10Jan 2020 .pdf

Description:

January 7, 2020

Francis S. Collins, MD, PhD
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892
Submitted electronically: <https://osp.od.nih.gov/draft-data-sharing-and-management/>

RE: DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Dear Dr. Collins:

The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard (MRCT Center) appreciates the opportunity to comment on the National Institutes of Health (NIH) draft NIH "Policy for Data Management and Sharing and Supplemental DRAFT Guidance" (hereinafter the "Policy"), published in the Federal Register Vol. 84, No. 217 on November 8, 2019.

The MRCT Center is a research and policy center that addresses the ethics, conduct, oversight, and regulatory environment of international, multi-site clinical trials. Founded in 2009, it functions as a neutral convener to engage diverse stakeholders from industry, academia, patients and patient advocacy groups, non-profit organizations, and global regulatory agencies. The MRCT Center focuses on pre-competitive issues, to identify challenges and to deliver ethical, actionable, and practical solutions for the global clinical trial enterprise. Over the last five years, the MRCT Center has been intimately involved in data sharing, including (1) developing guidance for sharing aggregate plain language summaries for participants and the public, (2) developing guidance for sharing individual results with participants, (3) promoting principles of individual participant data (IPD) sharing including protections of patient/participant confidentiality and privacy and of confidential commercial information, (4) developing template data use agreements and data contributor agreements for IPD and other data sharing, (5) crafting informed consent language to promote participant understanding of the implications of sharing de-identified data, (6) launching Vivli, a platform for global data sharing of IPD data, and (7) furthering the establishment of credit for data sharing for those individuals who choose to share their data, among other efforts. Of note, the responsibility for the content of this document rests

with the leadership of the MRCT Center, not with the its collaborators, nor with the institutions affiliated with the authors.¹

The MRCT Center strongly endorses the NIH draft policy and the importance that it places on data management and data sharing. This draft policy demonstrates an ongoing appreciation by the NIH of the utility and value of previously collected data and metadata not only for replication but for new discoveries. Further, proper stewardship of data is important, and the requirement for the submission of data management and data sharing plans prior to initiation of the research will be helpful in that regard. We are enthusiastic that NIH has taken this further step to include all scientific data (and metadata) as defined, of all data types and all sizes, and for all research funded by the NIH. We also understand that the NIH has outlined only the minimum expectations for NIH-wide Plans, and that the NIH ICOs may add additional requirements or expectations. We believe, however, that the NIH policy should be stronger, while nevertheless still permitting some flexibility.

We feel strongly that the **NIH should require data sharing, unless there is an ethical, scientific, or other defensible reason not to do so.** There should be a rebuttable presumption to share data; the burden should be on the investigator to provide cogent reasons that the data should not or cannot be shared. Subjective evaluations by investigators of potential data utility to the research community or the public should not be considered a sufficient reason not to share data.

There are risks to data sharing, including that of participant and patient privacy for studies that involve human participants and their data or biospecimens. Not all data need be downloadable and freely accessible: **measures to protect privacy and confidentiality** should be required. Those measures include de-identification, as mentioned in the draft policy, but also include other risk mitigation strategies: physical and technical security measures (e.g. data maintained in a repository, in a fit-for-purpose compute environment and not downloadable), controlled data access by qualified users, and other more novel methods (e.g. differential privacy, block chain technologies, etc.). We encourage the NIH to invest in the development and dissemination of these technologies to promote data sharing of sensitive data, and to issue appropriate guidance for their use. We further encourage the NIH to require disclosure of—and explanation of—data sharing plans to research participants during the informed consent process.

We encourage the NIH to provide **minimum expectations** for data management and scientific data, either within the policy or as additional guidance. The breadth of research and data acquisition supported by the NIH is expansive, covering different disciplines and including the spectrum of basic, translational, and clinical research. Guidance is needed to

¹ Brigham and Women's Hospital, Rope & Gray LLP, Harvard Medical School, Harvard University, and Yale Law School.

assist investigators and institutions, many unfamiliar with optimal data management and data sharing approaches.

Specific, required elements of the Plan should be developed, and an approximate (or “not to exceed”) **time frame** regarding when the data will be made available should be stated. The completeness and sufficiency of the Plan will only be encouraged by written detail.

We appreciate the development of the Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (Plan). While the descriptions of the specific data elements provide the reader with guidance on the development of a Plan, we encourage NIH to further complement this guidance with examples of (potential) comprehensive data sharing plans for different data types.

We also encourage NIH to provide **minimum expectations for data repositories and data sharing platforms** that meet requirements of the policy. We encourage NIH to develop and **maintain a database** that recognizes those repositories and platforms.

The policy states that “NIH may make Plans publicly available.” We believe that the **NIH should affirm its commitment to make available to the public the Plans** of funded research proposals and contracts. Public visibility of the Plans will be informative and educational, permit tracking, and encourage compliance. ClinicalTrials.gov should be used to disseminate the Plans for registered clinical trials, and the Plans should be posted prior to study initiation. Additional repositories can be used for other types of research, or the NIH can simply publish the Plan as an additional field linked to or hosted on the NIH RePORTER.

Data holders and data contributors should be encouraged to apply **data tags (i.e. metadata) that describe how the data can be used**—and applicable restrictions to its use—to reflect any contractual terms (e.g. licensing, copyright), informed consent parameters, and institutional, state, and federal policies. Metadata that describe the terms of use will help ensure the appropriate and compliant use of the data in the future. Further, NIH should invest in developing a universal language or library for such data tags and tools to render such metadata machine-readable.

The burden of managing and sharing data does not rest solely on the data contributor but equally on the data scientists and researchers who have access to the data. **Strict policies with enforcement provisions should be communicated to those who access the data**, and data use agreements employed as appropriate. Data tagging as described above will make compliance both easier for the user and auditable if necessary.

The data management and sharing plan should be an important and determinative part of any NIH proposal, and the **Plan should be reviewed and scored by the study section** (or contracting entity). The Plan should not be relegated to a “Just-In-Time” submission but should affect whether a proposal is prioritized for funding. Consideration of data

management, integrity, and stewardship (and, later, sharing) is an integral part of study design and quality.

We believe further that **no two-page limitation** should be imposed on the Plan. The prospective description of data management and sharing of data and metadata should be as long as necessary to describe all important details. To support its significance, the Plan should not “count” against the page limitations of the proposed science.

Finally, given that a principal goal of the NIH policy is to “serve the public,” we believe strongly that this is a time when the **NIH should require return of aggregate study results** to participants, at least for the results of clinical research, and in plain language understandable to an individual. Absent a cogent reason, these aggregate results should be available to the public. While there are many issues with return of individual results to a participant that require consideration and analysis, summary results of clinical trials and clinical research should be widely available and understandable—and may help to promote public engagement and public trust in the research and scientific enterprise.

Thank you again for the opportunity to comment on this important issue. We believe that the NIH is in a unique position to harness the power of data sharing for the public good, but only if it uses this opportunity to advance the culture of, and infrastructure to support, data sharing.

We are available to discuss our comments with you if that would be helpful and would be happy to work with you on any of the aforementioned items. Please feel free to contact the MRCT Center at bbierer@bwh.harvard.edu, sawwhite@bwh.harvard.edu, and mark.barnes@ropesgray.com.

Respectfully submitted,

Barbara E Bierer, MD
Sarah A White, MPH
Mark Barnes, JD, LLM

MRCT Center Comments on Draft Policy

Submission ID: 1369

Date: 1/10/2020

Name: Stephanie J. Lee, MD, MPH

Name of Organization: American Society of Hematology

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: ASH President

Domain of Research Most Important to You or Your Organization:

hematology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

ASH Comments on NIH Data Sharing Policy_LH_01_10_20_.pdf

Description:

ASH Comments on NIH Data Sharing Policy and Supplemental Draft Guidance



January 10, 2020

2020**President**

Stephanie Lee, MD, MPH
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue N, D5-290
PO Box 19024
Seattle, WA 98109
Phone 206-667-5160

President-Elect

Jane N. Winter, MD
Northwestern University
Robert H. Lurie Comprehensive Cancer Center
676 N. Saint Clair Street, Suite 850
Chicago, IL 60611

Vice President

Martin Tallman, MD
Memorial Sloan-Kettering Cancer Center
1275 York Avenue
Howard Building 718
New York, NY 10065
Phone 212-639-3842

Secretary

Robert Brodsky, MD
Johns Hopkins University
Ross Building, Room 1025
720 Rutland Avenue
Baltimore, MD 21205
Phone 410-502-2546

Treasurer

Mark Crowther, MD
McMaster University
50 Charlton Avenue East
Room L-301
Hamilton, ON L8N-4A6
Canada
Phone 1-905-521-6024

Councillors

Alison Loren, MD, MS
Bob Lowenberg, MD
Belinda Avalos, MD
John Byrd, MD
Cynthia Dunbar, MD
Arnold Ganser, MD
Agnes Lee, MD, MSC, FRCPC
Joseph Mikhael, MD, FRCPC, Med

Executive Director

Martha Liggett, Esq.

Andrea Jackson-Dipina, Dr.PH
Director of the Division of Scientific Data Sharing Policy
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

NOT-OD-20-013, “Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance”

Dear Dr. Jackson-Dipina:

The American Society of Hematology (ASH) appreciates the opportunity to provide comments to the National Institutes of Health (NIH) in response to NOT-OD-20-013, *Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance*.

ASH represents more than 17,000 clinicians and scientists worldwide, who are committed to the study and treatment of blood and blood-related diseases. These disorders encompass malignant hematologic disorders such as leukemia, lymphoma, and multiple myeloma, as well as non-malignant conditions such as sickle cell disease, thalassemia, bone marrow failure, venous thromboembolism, and hemophilia. In addition, hematologists are pioneers in demonstrating the potential of treating various hematologic diseases and continue to be innovators in the field of stem cell biology, regenerative medicine, transfusion medicine, and gene therapy. ASH membership is comprised of basic, translational, and clinical scientists, as well as physicians providing care to patients.

After reviewing the draft policy and supplements, the Society is fully supportive of the NIH policy for data management and sharing. We believe that such a policy will provide an important foundation to improve the reproducibility and reliability of research findings and to promote collaborative interactions. We are especially supportive of NIH’s proposal to collect data management and sharing plans as part of “Just-in-Time” documentation for extramural awards. Allowing the applicant to submit the plan later in the process instead of in the initial proposal will greatly reduce administrative burden for applicants and reviewers. In addition, having NIH staff review the plans will allow for a more uniform and streamlined process. We look forward to working with NIH to implement the final version of the policy, for example through workshops at our annual meeting.

The Society would like to highlight some specific issues related about the proposed policy’s scope and implementation.

While we fully support data sharing of almost all types, the draft policy is not clear about exactly which types of data NIH expects to be shared. While NIH is relatively clear on what is *not* expected to be shared, there may be benefit to NIH on being specific about which data *are to be* shared. The draft policy suggests the “incorporation of principles that respect the autonomy and privacy of research participants and protection of confidential

data," but the supplement suggests that data from human participants might be shared in an aggregated or summarized form and that each institute or center would have authority to determine which data ultimately must be shared. More precision in this area would be helpful to researchers; specificity will allow investigators to know exactly what is expected and will prevent the submission of data that is not wanted. As an example, there is information that should not or cannot be shared, such as PET scans from lymphoma clinical studies.

The Society is also concerned that data sharing plans will vary quite a bit from institution to institution and across NIH institutes and centers and thus will impact our members differently. As such, we recommend that NIH ensure that reasonably uniform standards are being applied across all entities. This will also help Institutional Review Boards craft consistent and acceptable consent forms.

Patient privacy, confidentiality, and institutional responsibility are not defined or discussed in the draft policy. Upon implementation of the data sharing policy, we are concerned that the ambiguity about privacy/confidentiality issues might be a barrier to the deposition of patient data. For example, the current genomic data sharing plans, that attempt to include patient consent for future data sharing, have proven difficult to implement because of these concerns. ASH recommends that NIH provide a model for how patient data is to be obtained with informed consent about deposition. While outside the scope of the draft data sharing policy, ASH would like to work with NIH in the long-term to address the legal issues regarding the public deposition of patient data.

ASH very much appreciates NIH's recognition of the effort and costs associated with data deposition and sharing and the supplemental guidance defining possible allowable costs. However, the draft policy is not clear about where the resources will come from to collate, submit, and store all of these data. Furthermore, the guidance only addresses those costs incurred during the term of the award but does not address costs associated with long-term data retention, stewardship and accessibility. Additional information and guidance from NIH on these points are recommended.

Finally, an important addition to the guideline would address the difficulty of depositing or using data in central repositories such as dbGaP. We feel that a commitment on the part of NIH to make central repositories more user-friendly would be an important addition to the NIH data sharing policy and would enhance acceptance of the final data sharing and management policy by our members.

Thank you again for the opportunity to submit comments. Please contact Suzanne Leous, Chief Policy Officer (sleous@hematology.org or 202-292-0258) if ASH can provide additional expertise as the data sharing and management policy is finalized.

Sincerely,



Stephanie J. Lee, MD, MPH
President

Submission ID: 1370

Date: 1/10/2020

Name: Jennifer Graff

Name of Organization: National Pharmaceutical Council

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Comparative effectiveness research, health services research, care delivery and reimbursement, and medical innovation

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Please see attached comments.

Attachment:

NPC- NIH Data Management Comments- Final.pdf



1717 Pennsylvania Avenue, NW, Suite 800, Washington, DC 20006 Phone: 202.827.2100 Fax: 202.827.0314 Web: www.npcnow.org

January 10, 2020

Principal Deputy Director Lawrence Tabak
National Institutes of Health
9000 Rockville Pike
Bethesda, MD 20892

Dear Director Tabak,

Thank you for the opportunity to comment on the Draft NIH Policy for Data Management and Sharing and Supplemental Draft Guidance. The National Pharmaceutical Council (NPC) shares the NIH's goals of promoting effective and efficient data management and ensuring research results and accomplishments are shared with the public.

NPC is a health policy research organization dedicated to the advancement of good evidence and science and to fostering an environment in the United States that supports medical innovation. NPC is supported by the major U.S. research-based biopharmaceutical companies. We focus on research development, information dissemination, education and communication of the critical issues of evidence, innovation and the value of medicines for patients. Our research helps inform important health care policy debates and supports the achievement of the best patient outcomes in the most efficient way possible.

As the National Institutes of Health (NIH) finalizes guidance on data sharing, we encourage the agency to ensure that publicly-funded research results and data are available for research purposes. The NIH should build on existing policies, including those established under the OPEN Government Data Act and the Health Insurance Portability and Accountability Act, ensuring that the information needed to conduct productive research is accessible, but personally identifiable health information is not publicly available. Health data from federal agencies and publicly-funded research is key to helping stakeholders accomplish many of our nation's most ambitious health goals. Promoting transparency and consistency in data sharing policies can lead to increased public accountability, promotion of research rigor, and an increase in the generalizability of knowledge gained from such data,¹ ultimately creating better outcomes for public health broadly while still ensuring personal health information remains confidential.

Further, data access plays a vital role in supporting consumer decision-making and improving overall population health. For example, NIH efforts such as the All of Us Research Program hold great promise to accelerate personalized medicine, understand disease progression, and improve health. Developing comprehensive and accessible data policies will be key to ensuring All of Us core values are maintained, including transparency and data available for research purposes.² In addition, increasing the availability of research-identifiable files from

¹ Doshi, JA, Hendrick, F, Graff, J, and Stuart, B. Data, Data Everywhere, But Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-level Health Care Data in the United States. *eGEMS (Generating Evidence & Methods to improve patient outcomes)*, 2016; 4(2).

² National Institutes of Health All of Us Research Program "Core Values." <https://allofus.nih.gov/about/core-values>. Accessed January 7, 2020.

both past and current research efforts can help further our understanding of precision medicine.³ In an aligned, high-functioning health-care system, everyone should be able to benefit from effective use of this and other data in order to improve quality and efficiency across the health care and public health landscapes of this country.

Federal and state agencies commonly release data at the aggregate health level, often called “public use files” or PUFs. While valuable for many research purposes, these files limit the data elements available or do not link to other files. For many research questions, data needs to be 1) available at the individual, rather than aggregate level, 2) longitudinal to distinguish patterns of care, 3) include dates of diagnoses, treatments and outcomes to accurately describe the order of events, 4) encompass fine-grained geographic detail to assess environmental or socioeconomic factors that may affect health outcomes, and 5) linkable to other data files such as provider characteristics, lab or genetic information, etc.⁴

To balance these research needs while maintaining individual privacy and confidentiality, many groups have addressed the policy dilemma using two approaches. First, other federal agencies have developed limited data sets (LDS) versions of files which limit the geographical information or small cell information. The NIH/National Cancer Institute Surveillance, Epidemiology, and End Results (SEER)-Medicare Linked Database is an example of an LDS. Second, other data sets are available only to researchers at governmental agencies or researchers in academic and nonprofit organizations. The NIH National Institute on Aging Health and Retirement Study (HRS)-linked to Medicare enrollment claims is an example of a dataset limited by research affiliation.

Deep scientific and analytic expertise resides within organizations that are often excluded from access to publicly-funded data. Many of these organizations already safely hold and analyze data collected through the delivery of healthcare operations. Ultimately, any standard that bars access to important data is detrimental to the larger goals of our healthcare system and the evolution of that system. Expanding access to federal and publicly-funded data to all researchers will dramatically increase the bandwidth for research, leading to increased quality of care, system efficiency, and patient satisfaction.

All researchers, no matter their affiliation, should be granted similar access to publicly-funded data. Financial benefit and profit status of an organization should not overlay the criteria by which access to data or a research proposal are evaluated. NIH data sets and funded data such as the SEER and HRS linked databases are valuable tools for researchers from all organizations and it will be important to ensure that future data, including that information available in the All of Us Workbench and Hub,⁵ are accessible and provide data that go beyond just summary information. The quality and efficiency of all physician groups, health plans, hospital systems, suppliers, and manufacturers can be enhanced using data. Therefore, the quality of research and its potential to improve health should instead be the standard. Placing a greater emphasis on research quality and intent, rather than simply the investigator’s affiliation could create greater opportunities while protecting confidential patient information.⁶

³ Doshi, JA, Hendrick, F, Graff, J, and Stuart, B. Data, Data Everywhere, But Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-level Health Care Data in the United States. *eGEMS (Generating Evidence & Methods to improve patient outcomes)*, 2016; 4(2).

⁴ Ibid.

⁵ National Institutes of Health All of Us Research Program. “Workbench.” <https://www.researchallofus.org/workbench/>. Accessed January 8, 2020.

⁶ Doshi, JA, Hendrick, F, Graff, J, and Stuart, B. Data, Data Everywhere, But Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-level Health Care Data in the United States. *eGEMS (Generating Evidence & Methods to improve patient outcomes)*, 2016; 4(2).

Overall, NPC encourages NIH to ensure that publicly-funded research is available to all interested and qualified researchers. Consistent and straight-forward policies on data management throughout the government has the potential to enhance research and promote more innovation across this country's health care system. We thank you for consideration of our comments and would be happy to discuss these ideas further.

Sincerely,

A handwritten signature in black ink, appearing to read "Jennifer S. Graff". The signature is fluid and cursive, with the first name being the most prominent.

Jennifer S. Graff, PharmD
Vice President, Comparative Effectiveness Research
National Pharmaceutical Council

Description:

NPC- Comments on NIH Data Management and Sharing

Submission ID: 1371

Date: 1/10/2020

Name: Chuck Cook

Name of Organization: Global Biodata Coalition

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All biodata resources

Type of Organization: Other

Type of Organization - Other: Non-governmental organization

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

The Global Biodata Coalition (GBC) is a forum created by and for biomedical and life sciences funders to aid those funders in better coordinating support for biodata resources and to ensure sustainable funding for the global infrastructure of biodata resources worldwide.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

This primary purpose of the NIH data management policy is to encourage researchers to manage their research data as responsibly as possible. The data generated from publicly funded research should whenever possible be made FAIR so that other researchers can benefit from previous work. The Global Biodata Coalition strongly supports this policy.

Section II: Definitions:

The definitions in section II are accurate. No other comments on this section.

Section III: Scope:

We fully support the wide scope of this policy in applying to virtually all NIH-funded research that results in the generation of scientific data, regardless of the funding amount.

Section IV: Effective Date(s):

The Global Biodata Coalition has no specific suggestion with regard to the effective date of the policy. However, it is important to ensure that all stakeholders and funding recipients are clearly aware of this policy as early as possible before it is implemented, and to support applicants in completing data management plans when the policy is first implemented.

Section V: Requirements:

Requirements for submission of and compliance with a data management plan are reasonable. However, there are a number of other issues that might be addressed:

How will compliance with the data management plan be assessed?

Will there be a requirement to describe data management as part of final grant reporting?

Will failure to implement a data management plan adversely affect future funding applications?

Section VI: Data Management and Sharing Plans:

This section as written is reasonable and unremarkable.

The supplemental draft guidance makes it clear that the "data" to be managed include not just raw data, such as nucleotide sequences, but also software/analytical tools, pipelines, workflows, metadata, and relevant community standards. This inclusive definition of scientific data should also be included in the main data management policy description.

Section VII: Compliance and Enforcement:

This section has few details. We take it to mean that compliance with the data management plan will be part of the normal process of reviewing funding during the funding period and as part of any final report.

The largest penalty for not implementing a data management plan appears to be reducing the rank of future applications or even disallowing future funding. NIH may wish to ensure that mechanisms are in place to monitor data management, to record data management failures, and to penalize those failures in future funding rounds.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

This additional supplementary guidance relating to allowable costs is welcome.

Section 2 references eligibility of fees for commercial repositories. This might encourage researchers to seek fee-paying repositories when, in fact, most data repositories are publicly-funded and do not incur charges for data deposition or long-term storage. Many such repositories are within NIH, such as those managed at NCBI, and many others are extramurally funded by various NIH institutes.

The text should be rephrased to ensure that researchers understand that there are many many open access data resources that will accept research data without incurring any charges, and that fee-requiring data resources are unusual and rare.

When costs are incurred for long-term storage—these could be commercial or simply costs for local infrastructure—how long will NIH support those costs: 5 years, 10 years, 20 years? This should be stated explicitly and, of course, researchers should be encouraged to seek long-term solutions that do not incur costs, such as deposition in public data resources.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The draft guidance should encourage the use of free open access data resources, most of which are publicly or charitably funded (see next section). Such resources provide the safest solution for long-term storage of research data: data are accessible to all researchers, even those with limited funding, and these resources often have long track records, suggesting that they will be sustainable into the future.

The guidance should allow, but discourage, deposition into data resources that charge fees to users even if they do not charge fees for data storage. Such resources make it difficult for other researchers, particularly those with limited funding, to access stored data for reuse.

The guidance should strongly discourage deposition into data resources that charge fees for long-term storage, and NIH should consider how it might fund such charges if they are incurred. Will NIH fund data storage in perpetuity after the end of a granting period?

The guidance should also strongly discourage researchers from setting up local systems for data storage and access. Such systems are very likely to fail due to loss of local funding or when a PI retires or moves to a new institution.

NIH should make sure to continue sustained long-term funding for data resources that act as repositories for managed data. These are both internal (e.g., NCBI resources) and external (for example the Model Organisms Databases and UniProt). These resources are crucial to the infrastructure of data deposition and will remain the primary repositories selected by NIH-funded researchers to support their data management plans.

Other Considerations Relevant to this DRAFT Policy Proposal:

The policy as written makes no recommendations with regard to where research data should be deposited, and the webinar on 16 December made it clear that such recommendations will be made separately from this data management policy.

We strongly encourage NIH to issue recommendations for repositories simultaneously with issuing this data management policy. Worldwide, there are many established repositories that accept and store most types of experimental data. Most of these resources are publicly or charitably funded, and include many funded by various NIH institutes both internally ((e.g., NCBI resources) and externally (for example the Model Organisms Databases and UniProt). These resources are crucial to the infrastructure of data deposition and will remain the primary repositories selected by NIH-funded researchers to support their data management plans.

Recommendations for which repository to use should also take into consideration other efforts to recommend depositories for research data. These include the ELIXIR Deposition Databases (<https://elixir-europe.org/platforms/data/elixir-deposition-databases>) and nascent efforts by FAIRsharing and publishers to collaboratively recommend resources for data deposition (<https://osf.io/m2bce/>).

Finally, long-term storage of research data relies upon the sustainability of the data resources that archive most of these data. These resources are funded by various public and charitable funding agencies worldwide, including the NIH. Any NIH policy to require long-term storage and management of research data should therefore be supported with a concomitant commitment by NIH to ensure sustained long-term support for the resources it funds and for the global biodata resource infrastructure as a whole.

Attachment:

Description:

Submission ID: 1372

Date: 1/10/2020

Name: Janis Geary

Name of Organization: Arizona State University

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

genomic data sharing

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Mention the CARE principles for data sharing. <https://www.gida-global.org/care>

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

The only mention of Tribal considerations is found here. What about data that researchers collate or collect outside of Tribal jurisdiction? There are examples happening currently where Tribal groups feel that their laws are being purposefully circumvented. The policy needs to be explicit about Tribal data, ensuring Tribal groups have absolute control over all Tribal data.

It is not enough to mention protecting individual privacy, as it misses the concept of group privacy.

Section VII: Compliance and Enforcement:

The NIH should tackle the problem of researchers misusing shared data. A lot of hesitancy regarding sharing data comes from researcher concerns about misuse.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The costs of community engagement to develop data management and sharing approaches (Tribal groups, patient groups, other stakeholders) should be an allowable expense. There should be an allowance for Tribal groups to develop their own data management infrastructure.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1373

Date: 1/10/2020

Name: Christopher Austin

Name of Organization: Johns Hopkins University

Type of Data of Primary Interest:

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

JHU Response to NIH Request for Comment on Data Sharing.pdf

Description:

JHU Response to NIH Request for Comment on Data Sharing



January 10, 2020

Andrea Jackson-Dipina, Dr.PH
 Director of the Division of Scientific Data Sharing Policy
 Office of Science Policy, NIH
 6705 Rockledge Drive, Suite 750
 Bethesda, MD 20892

RE: Response to Request for Public Comment on the DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance (the "NIH Proposed Policy"); 84 Fed. Reg. 60398 (Nov. 8, 2019)

Dear Dr. Jackson-Dipina,

Please accept this letter as the response of the Johns Hopkins University ("JHU") to the request for public comment on the above captioned NIH Proposed Policy. JHU is committed to the creation and dissemination of new knowledge for the improvement of health. JHU shares the National Institute of Health's ("NIH's") view that "increasing access to scientific data resulting from NIH-funded or conducted research...enabl[es] the validation of scientific results, allowing analyses to be strengthened by combining data, facilitating reuse of hard-to-generate data, and accelerating future research." 84 Fed. Reg. 60398 (Nov. 8, 2019). JHU shares in the broad goals that underlie the NIH Proposed Policy, and JHU has, to its knowledge, developed one of the first institutionally supported, centralized data management services group within a university, and has worked to create appropriate architecture and work flows to continue to support the ethical sharing of restricted health sciences data.

JHU joins generally in the comments submitted by the Council of Governmental Relations, the Association of American Medical Colleges, and the joint comments of the Association of American Universities and the Association of Public and Land-grant Universities, and writes separately to make the following, additional recommendations, which are informed by our institutional experiences.

I. The NIH Should Fully Leverage Existing Data Repositories and Require Institutes to Follow Consistent Guidance

The NIH Proposed Policy notes that NIH "encourages the use of established repositories." JHU urges the NIH to consider explicitly stating that a commitment to deposit in existing repositories (including posting to clinicaltrials.gov for any qualifying studies) is sufficient to satisfy the guidance, and requests that the NIH maintain a public list of those repositories that will be deemed to be acceptable. Rather than leaving broad latitude for individual institutes and even individual NIH staff to vet individual proposed data sharing plans, JHU believes the goals of ensuring sharing while protecting participant

Office of the Provost

265 Garland Hall 3400 N. Charles Street Baltimore, MD 21218 410-516-8070 <http://web.jhu.edu/administration/provost>

privacy and proprietary interests can be better served by focusing on supporting a limited number of well architected, well managed and adequately funded repositories to meet this goal. From a technology and resources perspective, it is worth noting that the NIH currently lists some 87 data archives for submission of different data types which are supported in whole or in part by the NIH. Given the large costs associated with effective sharing of research data, individual grant applications are not the most effective place to experiment with plans. JHU respectfully submits that fully leveraging existing repositories will be facilitated by NIH funding the repository function directly to a limited number of repositories, and then offering (perhaps as a checkbox option) that investigators indicate which of the NIH approved existing repositories they will utilize.¹ Institutes should be required to indicate which repositories they support, and there should be maximum effort from the NIH to limit the ability of funding institutes to require different or additional sharing.

To better understand why a limited number of repositories will better serve the NIH goals, JHU offers its recent experience in expanding and developing its existing data archives. JHU Libraries have spent the past several months working with the Johns Hopkins Medicine Data Trust (the entity responsible for governance of patient-related data) to develop support for the sharing of restricted health sciences data. The Libraries' team has included administrators, project managers, infrastructure software developers, user interface software developers, and systems administrators. In addition to numerous hours of effort from investigators, this team has spent over 2,000 hours toward the design of the new archive and the initial infrastructure development. This effort does not include the subsequent implementation and operation of the new archive. Clearly, the most efficient way for data sharing to occur is not to replicate the development and implementation of extensive infrastructures across all institutions.

NIH has experimented with cloud-based resources through the Data Commons, STRIDES, and other programs and the NIH Proposed Policy should expressly address cloud-based repositories as an option. JHU recommends that NIH determine whether cloud-based resources can be a viable option for research data sharing. Specifically, it will be important to understand whether the use of such cloud-based resources increases costs, especially if one attempts to mitigate technological lock-in, which can reduce flexibility and possibly introduce significant egress costs. Also, NIH is best positioned to negotiate appropriate and consistent terms and conditions, license agreements, cost models, etc. for cloud-based resources on behalf of NIH funded scientists to ensure cost-effective export and transfer of data.

II. The NIH Must Recognize the Revised Common Rule Changes in Definitions of Identifiable Data and Consider the Ethical Implications of Human Subject Data Sharing

The NIH notes in the NIH Proposed Policy that it “prioritizes the responsible management and sharing of scientific data derived from human participants,” but provides little specifics on how

¹ The list is currently available at https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html could be expanded to include other repositories that, while not directly NIH supported, are deemed sufficient to meet the data sharing goals. In this way, the function of data sharing and curation could follow the model set by the ATCC, which serves as an acceptable repository and dissemination model for biological materials.

individual investigators can manage the need for appropriate consent, particularly given the significant recent revisions to the Federal Common Rule, 45 CFR Part 462. The revised Common Rule defines a “human subject” as “a living individual about whom an investigator...[o]btains information or biospecimens through intervention or interaction with the individual...or...[o]btains, uses, studies analyzes, or generates identifiable private information or identifiable biospecimens.” 45 CFR 46.102(e)(1)(emphasis supplied). The Common Rule further defines “identifiable private information” as information from which the identity of the person “is or may readily be ascertained by the investigator” or “associated with” the information. 45 CFR 46.102(e)(5). In recognition of rapid changes in data science which make it possible to combine data to identify individuals in ways that were not previously possible, the Common Rule now requires the federal government to consult with “appropriate experts (including experts in data matching and re-identification)” on what it means for data to be “identifiable” 45 CFR 46.102(e)(7)(i). While the guidance called for by this regulation has not yet been generated, it is apparent that coordination between the NIH Proposed Policy and the new standards to be developed by the Office for Human Research Protections (“OHRP”) is critical. The NIH Proposed Policy must take into account the ability for the definition of identifiable private information to change under the revised Common Rule (and other applicable local laws and regulations) and must account for the fact that future research with shared data may require additional IRB review and approval.

JHU has on-going and active efforts to engage our research participant community in dialogue about the responsible use of research data and biospecimens in research. We are acutely aware that many individuals desire a fuller understanding how their data may be used in the future.² An ethical approach to data sharing that is respectful of participant concerns must consider the language of consents related to data sharing, including any potential limitations on data sharing that should be imposed based on the language presented when consent was obtained.³ JHU notes that the NIH has expressly excluded “completed case report forms” from the definition of “Scientific Data” that is subject to the sharing requirements. We strongly recommend that the NIH add to this exclusion “any human subject data that is determined to be individually identifiable, under applicable OHRP standards and guidances or other applicable law, for which express informed consent was not given to the sharing of the data” to the scope of what is excluded from sharing. While the NIH has experimented with forms of dynamic consent as part of the All of Us research project, JHU submits that the NIH be explicit that awarding components in institutes not attempt to mandate the content of informed consent for prospective studies, but to continue to engage with the research community and IRBs on ethical and appropriate approaches.

III. The NIH Proposal That Data Sharing Plans be Addressed as Part of the Just in Time Review is Impractical and Burdensome For Human Subject Data

² In particular, increased focus on use of personal data by technology companies has heightened patient awareness and concern about the use of their data without their explicit consent.

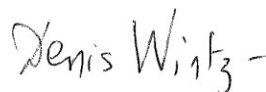
³ With respect to the NIH encouragement that “shared scientific data...be made available as long as it is deemed useful to the research community or the public,” JHU notes that the federal requirements for data retention are already set in federal laws and regulations, and that the NIH cannot, by policy or guidance, change those retention obligations, particularly after the award funding has ended.

The NIH Proposed Policy states that data sharing plans should be submitted as of just-in-time (“JIT”) reviews for extramural awards. Except for where the NIH has established and prospectively approved a particular existing data repository, JHU believes this requirement for submission at JIT review does not take into account the need for institutions to conduct the reviews of such plans for legal, ethical, security, and institutional policy compliance in the case of data collected from human subjects. As approved data sharing plans will be a condition of award, institutions would be forced to develop and institute new review processes for data sharing plans at the point of JIT notification to ensure the proposed plans align with institutional policies and are reflective of any consent or contractual limitations for data sharing.

This burden will be particularly challenging for human subject research data sharing plans. In 2019 alone, JHU processed over 100 “planning phase” applications, preliminary IRB applications designed to meet the JIT requirements. These applications often required review in less than 24 hours and on average were processed in 3 business days. Adding additional institutional reviews of data sharing proposals into this process would significantly impede our ability to meet JIT deadlines and jeopardize important funding opportunities. Except for cases where the data sharing is to a pre-approved repository (e.g. clinicaltrials.gov), institutions will not be able to respond with confidence that a legal, ethical, secure, and compliant plan is being proposed. JHU recommends that, if JIT review of plans becomes a binding conditions of an award, that NIH include a standard exemption provision in the condition of award that makes explicit that institutions are permitted to revise the type, amount, and form of data shared based on the final approval of the research.

Sharing of research data is a fundamental part of the on-going scientific dialog and JHU remains committed to that dialog. In order to build constructively on past NIH guidance on data sharing, JHU respectfully submits these comments for NIH consideration.

Very truly yours,

A handwritten signature in black ink that reads "Denis Wirtz". The signature is written in a cursive, slightly slanted style.

Denis Wirtz, Ph.D.
Vice Provost for Research
Johns Hopkins University

Submission ID: 1374

Date: 1/10/2020

Name: Sarah Wright

Name of Organization:

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Data of Primary Interest - Other:

Type of Organization: Not Applicable

Type of Organization - Other:

Role: Other

Role - Other: librarian

Domain of Research Most Important to You or Your Organization:

basic life sciences data

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

NIH_DMPolicyDraft_20200110.docx

Description:

January 10, 2020

Carrie D. Wolinetz, PhD
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Dear Dr. Wolinetz and the Office of Science Policy,

I am a Life Sciences librarian at Cornell University. In my position, I am committed to creating, maintaining, advancing, and teaching best practices for research data management, access, and preservation, and am actively engaged in assisting researchers with writing and complying with data management plans from NSF and other funding agencies.

With regards to the Draft Data Management and Sharing Policy the National Institutes of Health's Office of Science Policy has proposed for all NIH-funded research, I am writing to share my comments, incorporating feedback gathered from a very small sample of active NIH-funded researchers at the Ithaca campus of Cornell University. Feedback was gathered via a short survey that was mailed to a list of NIH-funded researchers provided by the university's OSP, garnering 19 responses. Although this is a small number, I believe that it still constitutes valuable feedback, and I have tried to honor all of the input received. I appreciate both the opportunity to share feedback and the Office of Science Policy's iterative approach, with obvious improvements between this and prior versions of the policy.

If you have any questions or would like to speak further about anything mentioned below, please contact me at sjw256@cornell.edu.

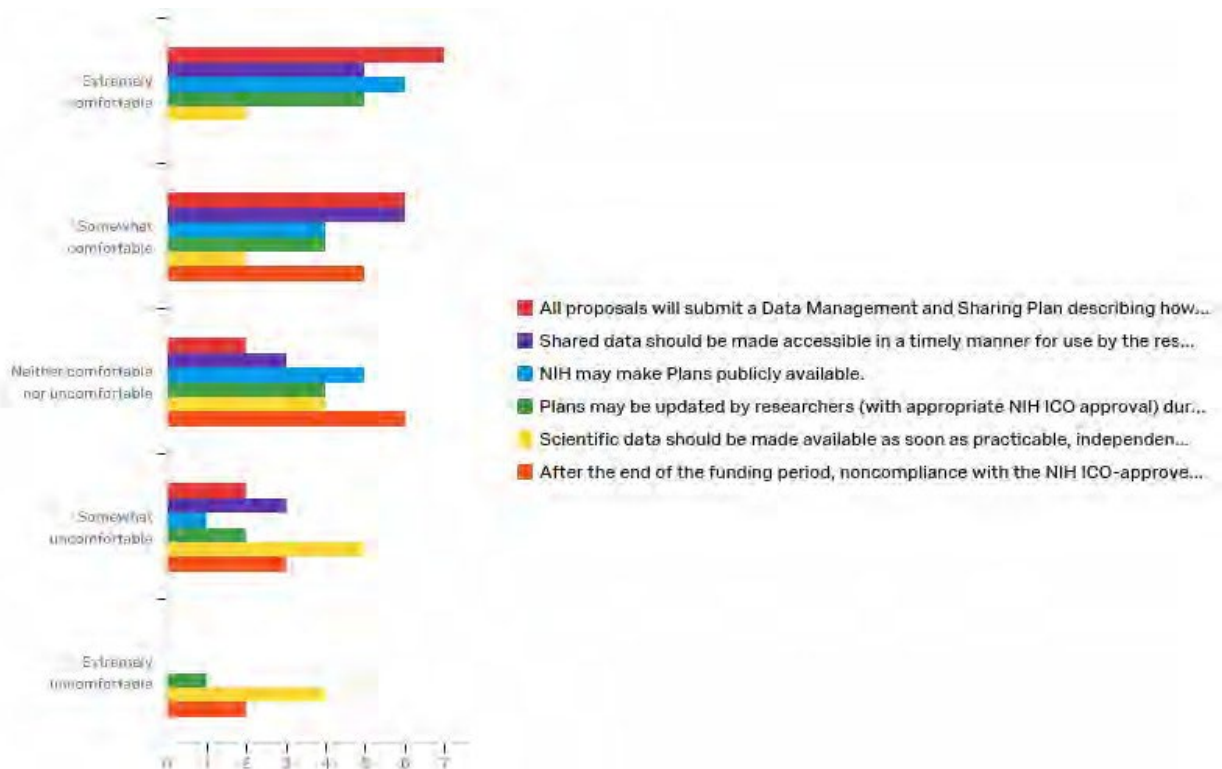
Sincerely,

Sarah Wright

Section I: Purpose

The Cornell University researchers that responded to the survey unanimously agreed with the statement that data sharing is important (albeit assigning data sharing varying degrees of importance). Of the 18 researchers that responded to the question, 15 indicated that they already have a data management plan requirement, and this is consistent with my estimates that about 50% of the Ithaca campus researchers are funded over \$500,000, and thus already required to have a data management plan. Furthermore, I believe that the majority of researchers are already satisfying good data management practices, and the implementation of this policy should not constitute an undue burden. My survey results also back up this statement, as Table 1 shows – most researchers were pretty comfortable with most of the aspects of the data management plan that we asked about. Only 2 of 17 respondents were uncomfortable with the statement “All proposals will submit a Data Management and Sharing Plan...,” and only 3 of 17 were uncomfortable with the statement that “Shared data should be made accessible in a timely manner for use by the research community and the broader public.” I agree with and applaud the NIH's “longstanding commitment to making the results and outputs of the research that it funds and conducts available to the public.”

Table 1: For each statement, please select your level of agreement with the following: “I am comfortable with this component of the draft policy.”



Section II: Definitions

Section III: Scope

Section IV: Effective Date(s)

Section V: Requirements

Section VI: Data Management and Sharing Plans

The NIH's "just-in-time" approach to data management and sharing is interesting. Many other funding agencies require plans at the grant application stage, and both ways have advantages and disadvantages. In general, I would recommend to keep funder requirements and RDM workflows as similar as possible to prevent confusion among researchers, many of whom are receiving funding from multiple funders. The "just-in-time" approach streamlines the grant submissions process by not requiring plans for projects that may never receive funding and prevents peer reviewers, most of whom are not data management experts, from needing to assess plans. However, it could result in missing infrastructure or lack of budgetary support for data management and sharing since those things may not be considered until well after project has been designed. Adopting this timeline will make it very important to ensure that the NIH staff reviewing the plans have adequate data management training and experience to help researchers anticipate such issues.

Additional comments found below under Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan below.

Section VII: Compliance and Enforcement

Responses indicate that Cornell researchers are also less comfortable with the statement that "After the end of the funding period, non-compliance with the NIH ICO-approved Plan may be taken into account by the funding NIH ICO for future funding decisions for the recipient institution," with almost a third of my responses (5 out of 16) indicating that they were somewhat or extremely uncomfortable (see Table 1). While I recognize that this method of enforcing compliance may be an unwelcome adjustment for some researchers, it's consistent with other funder's methods of enforcement, and I support the provision that the plan becomes a term and condition of the grant. I also recommend stronger wording around DMP updates: the NIH should indicate that it is expected for researchers to update their plan as their research project changes. Additionally, I recommend more guidance on how this accountability will be assured, both during and after the award (for example, guidance from the NSF Engineering directorate includes a detailed section on Post-award management: https://nsf.gov/eng/general/ENG_DMP_Policy.pdf).

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

While I do believe that the majority of researchers are already satisfying good data management practices, and the implementation of this policy should not constitute an undue burden, the researchers that responded overwhelmingly indicated some level of difficulty associated with many of the data management-related activities I asked them about (Table 3). Help with data curation seems especially important, and I am appreciative that this and other activities are explicitly included in allowable costs. However, the list of activities necessary for good data management and sharing are longer than those currently included in allowable costs, and require a substantial amount of human intervention beyond what may be easily sustainable with current infrastructure at most institutions, Cornell included. I recommend that NIH state explicitly that grant funds may be spent on research data management activities and personnel, showing that NIH understands the level of expertise, time and effort necessary

to properly manage research data, indicate to researchers that it is a worthy expenditure and clarify that this would not come at the expense of the core, funded research.

An additional challenge is that some researchers won't use the grant budget for data management and sharing costs –responses to my survey were evenly split between indicating likely vs. unlikely to budget for data management. Some of the reasons for not using the budget for data management include the following:

- A desire to prioritize resources for generating data
- Uncertainty about how much to budget for
- Concern that they are already requesting the maximum allowable funds with large human clinical study costs or genotyping costs
- Concern over payment mechanisms because sharing costs will be incurred much later, perhaps even after the grant is over depending on the grant mechanism. For example, an R21 is only a short-time period, and often the data are all barely collected by the time the grant ends.

Researchers also indicated some additional expenses that they would like to see as allowable costs in Table 2 below.

Table 2: Researchers indicated additional expenses that they would like to see as allowable costs.

What other expenses, if any, would you like to see included as an allowable cost associated with data management and sharing?
data storage
administrative costs
Storage of all types
Administrative assistance in this task
storage costs
additional costs outside of cap that are from other funding reserve to support this need
Back up of data

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan

I am happy that the NIH guidance aligns with requirements of other funding agencies, as this makes it easier for multidisciplinary researchers working with multiple sponsors, and supports the idea of general best practices that should apply to all research data. However, while not every detail may be finalized at the time of proposal, I recommend requiring that plans be updated as details change or develop, rather than allow details “to be determined”.

Under “1. Data Type” I have concerns about de-identification. My experience is that researchers don't have enough guidance and instruction concerning de-identification and I worry that both researchers and repositories will not be able to comply without better guidance and standards. More explicit advice

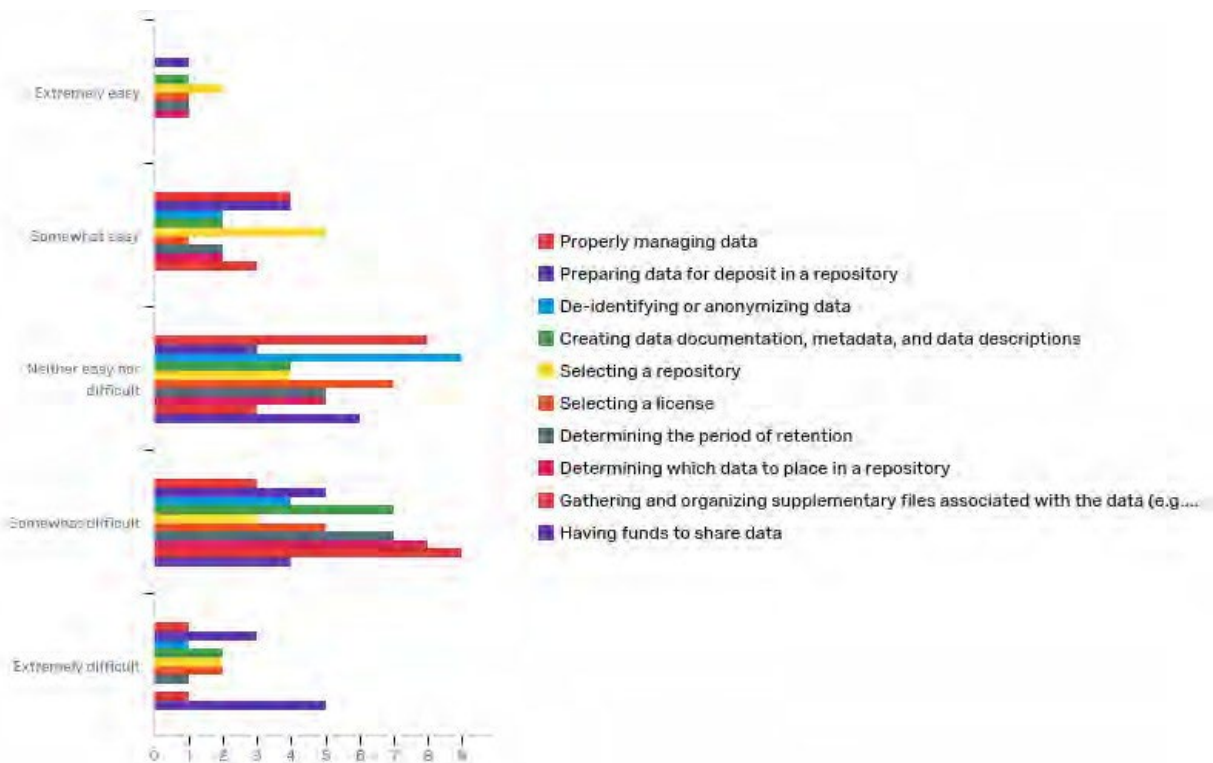
about which data can be publicly shared versus data that should not be shared, as well as more guidance about de-identification best practices (how to de-identify instead of what must be removed) would be most useful for compliance with this new policy.

Under “4. Data Preservation, Access, and Associated Timelines” researchers are slightly less comfortable with the statement that “In general, scientific data should be made available as soon as practicable, independent of award period and publication schedule,” (9 somewhat or extremely uncomfortable, Table 1) in contrast to the statement in the “I. Purpose” section, “Shared data should be made accessible in a timely manner for use by the research community and the broader public,” (only 3 somewhat uncomfortable, Table 1). My interpretation is that researchers are willing to share, however, many researchers still want to be able to complete the project and publish before sharing the data.

Other Considerations Relevant to this DRAFT Policy Proposal

While I do believe that the majority of researchers are already satisfying good data management practices, and the implementation of this policy should not constitute an undue burden, data management-related activities do present an additional challenge to already overloaded researchers. (Table 3). There will be a learning curve with some of these activities, and some institutions are better posed to help researchers than others. The library and other service providers on campus provide assistance with many of these activities, however we still see challenges around getting researchers connected with the services they need and face sustainability issues with not enough staff to provide those services. Many researchers indicated that they would spend a significant amount of time managing their data, or about 10% of their project time. Implementation of this policy will have a financial impact, and researchers indicated that they are already often at the maximum of the grant budget, so we urge the NIH to consider ways to support this policy, whether by raising the cap, providing other funding streams, providing support services (for example around data curation or other pain points for researchers), and/or providing tools and software to make the process easier and more straightforward for researchers.

Table 3: For each of the activities listed, please indicate the expected level of difficulty.



Submission ID: 1375

Date: 1/10/2020

Name: Nicole Capdarest-Arest

Name of Organization: University of California, Davis

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Head, Blaisdell Medical Library

Domain of Research Most Important to You or Your Organization:

translational research, basic science research, clinical research, big data, veterinary medicine, etc.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

"Shared data should be made accessible in a timely manner for use by the research community and the broader public." - will there be any minimal requirements or expectations related specifically to how data should be made accessible and what sort of reasonable time frame can be expected?

Section II: Definitions:

Should there be any definitions related to repositories? Given that this data will need to be stored somewhere, will there be any requirements and/or related definitions related to length of time data will need to be stored, security of repository, etc.? Will NIH be providing the repository for storage of this data or will it be up to principal investigators to secure adequate storage?

General note - not all abbreviations used in the Policy are defined in the Definitions section. This should be cross-checked.

Section III: Scope:

No comment.

Section IV: Effective Date(s):

No comment.

Section V: Requirements:

Will there be any requirements related to length of time for preservation of data, security/stability/ownership of repository, etc.?

Section VI: Data Management and Sharing Plans:

"NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." - who will decide whether said data is "useful" for the research community or the public? Although some data may be useful to preserve for historical purposes, due to likely advances in research methods, etc., and the reality that both "good" and "not-so-good" science occurs, how will researchers know when data should be made no longer accessible? Beyond the stated mention of approaches to ensure data security and privacy compliance, will there be any provisions related to requests for access or are researchers expected to make all of their data open for anyone, anywhere, at any time? In certain cases, where privacy and security of data are paramount, would providing metadata about the dataset meet the policy requirement?

"NIH encourages the use of established repositories for preserving and sharing scientific data" - will NIH be providing a list of such repositories and do those repositories need to meet certain standards for where they are based, security, capacity, etc.? Will NIH be spinning up a platform or a facet/data element (Field) of PubMed for data so that the public can easily search and access publications with attached/accessible datasets and data management plans?

Related to Plan Assessment - will a rubric or a standard be published so that grant applicants can be aware of the metrics by which the funding NIH ICO will assess each Plan? Guidelines for researchers should be made available so that they are aware of expectations, format, requirements, etc. for basic standards for Data Management Plans, along with rationale and context for such standards. It is not simply enough to require that people submit a plan, some transparent quality standards should be in place to help grant applicants aspire to a minimal level of quality. Such standards should also provide some basic consistency for NIH to be able to assess the effort on an aggregate level.

Section VII: Compliance and Enforcement:

Related to the comments in Section VI above, we again reiterate the need for goals and measures for a modern data sharing framework. "During the funding period, compliance with the Plan will be determined by the funding NIH ICO." - again, by what metrics will the funding

NIH ICO be considering Data Management Plans? It is not simply enough to submit a plan; some basic, transparent standards for such should be in place so that applicants are aware and NIH ICO can subsequently more objectively measure whether a Plan is "quality" or substandard and should not be approved or revisions requested.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

As mentioned in above comments, will there be any standards for data repositories? If they are hosted in private industry or by third-party vendors, will such repositories need to undergo any sort of risk assessment or provide NIH assurance of meeting a certain level of information security standards (e.g., CoreTrust Seal, ISO 16363)? From a risk management perspective, critical information technology infrastructure should be requisite for repositories preserving federally funded health-related data. Even if it is de-identified and privacy is not a concern, if preserved data were maliciously or even unintentionally altered, the risk of significant harm to further research drawing upon such data, leading to potential future faulty research and potential public harm.

We would also recommend that, given our current environment, including data analysis, data validation, data security as part of the listed allowable costs. Additionally, with regard to infrastructure, we recommend that NIH broaden this section to be more inclusive to data repository infrastructure which may not fit into this model. Infrastructure improvement that enhances our ability to preserve and share data should be included in allowable costs.

Lastly, with regard to local data management considerations, consider broadening the language related to local unique and specialized information infrastructure so that it would allow institutions who participate in standardized data sharing tools, resources, and information exchanges to support and incentivize investment in intraoperative and secure data sharing tools. This is especially important in healthcare where the costs of cybersecurity controls and security personnel is in great demand and expensive to implement and hire.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We are happy to see the list of requisite elements for a Data Management and Sharing Plan, however, as stated above, a transparent rubric for assessing each element should be made available. Furthermore, educational modules and training should be made available related to such expectations and recommended submission qualities/format for a successful plan. For example, making available some examples of quality plan submissions for various particular use cases (e.g., data management plan for study involving EHR data, device development, human specimens, clinical research, public/private).

Other Considerations Relevant to this DRAFT Policy Proposal:

Some proposed rubric elements for a modern data sharing framework:

- Does the data management plan name a metadata format or commit to maintaining and sharing a readme file about each dataset? Is the documentation clearly outlined and defined?

- Are the analysis methods outlined?

- Does the data management plan address storage security and privacy concerns regarding the data or clearly state that the dataset can be open?

- Is there a backup strategy?

- For longitudinal datasets, have appropriate data management practices been identified--use of database, data standardization, version control?

- Has conversion to non-proprietary file formats been considered?

- Has a repository for the shareable data been identified; is it suitable for the type of data? Has that choice been appropriately reflected in the budget?

- Have data management roles been assigned to specific team members?

Attachment:

Description:

Submission ID: 1376

Date: 1/10/2020

Name: Kevin McGhee

Name of Organization: New York Genome Center

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

neurodegenerative disease, neuropsychiatric disease, and cancer

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The New York Genome Center (NYGC) supports the proposed NIH Data Sharing and Management Policy. Broad sharing of scientific data generated with public support is beneficial to the public and scientific advancement, and this policy helps to further this widely accepted goal. However, we would like to ask NIH clarification concerning several issues that pertain to the sustainability of ongoing data sharing and management.

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

RFAs issued after the effective date of this policy should have sections explicitly outlining whether there are deliverables expected from the proposals that would specifically be subject to sharing, and whether there exist specific repositories in which the data are expected to be stored so that applicants can be sure to address these points as needed in their applications as well as their data management and sharing plans.

In addition, we ask NIH for clarity on expectations for primary storage of data by institutions that have submitted data to an NIH-supported shared repository. It is unclear whether basic data security requirements mandate that the institution maintain redundant archives of such data through the life of an award when such data are hosted publicly, or if institutions may rely solely on the publicly hosted copies as fulfillment of their data security obligations, and may delete their copies upon submission.

Section VI: Data Management and Sharing Plans:

We support the consolidation of data sharing elements of award proposals into a concise Data Management and Sharing Plan (Plan). This approach, along with the policy's provisions for enforcement and regular review of the Plan, will ensure that careful attention is given to these issues prior to, and throughout the life of, an award. The requirement for an explanation of protections for human subjects research participants helps to enforce early scrutiny of this issue by researchers and IRBs, better ensuring appropriate protections for human subjects while minimizing future problems with data sharing and use.

Section VII: Compliance and Enforcement:

It is unclear, from the proposed policy and guidance documents, whether the Data Management and Sharing Plan is meant to differ substantively from the Resource Sharing Plan required at the application stage, and how these plans could potentially differ, and if so, will the Data Management and Sharing Plan filed at JIT be subject to peer review, or held to the standard set by the Resource Sharing Plan during application?

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We have broad concerns about the future availability of NIH support for shared data repositories and ask clarification about the impact of this policy on those issues. Specifically, we are concerned about the potential creation of unfunded mandates for institutions that stem from long-term obligations for support of data sharing and management, whether these costs stem from mechanisms maintained by the researchers such as cloud-based web/ftp sites or through ongoing payment to third parties providing shared hosting at cost. While the proposed policy accounts for short-term support of budgets for such costs, it is unclear how institutions are meant to plan for long-term costs of data sharing and management that extend beyond the award period, or whether award terms would allow for pre-payment of such future costs.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We ask that the guidance on the Data Management and Sharing Plan ("Plan") require applicants to specifically address the sharing of aggregate data in the Plan, and whether it is expected that special sensitivities concerning the research subject population require placement of aggregate data under controlled access, in contrast to the November 1, 2018 Update to NIH Management of Genomic Summary Results Access (Notice Number: NOT-OD-19-123), which established a general expectation for unrestricted access for genomic summary results.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1377

Date: 1/10/2020

Name: Joanna Groden

Name of Organization: University of Illinois at Chicago

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: biomedical research

Type of Organization: University

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

biomedical research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Over the past 15 years, the importance of data sharing, curation, and preservation has grown more explicit. Datasets are increasingly recognized as independent scholarly objects that can be documented in reports to funding agencies.

We are pleased to see the DRAFT NIH Policy for Data Management and Sharing which documents the growing recognition of the need for data curation, preservation, and sharing to enhance biomedical research veracity, impact, and return on tax payer investment. We concur with the NIH's expansion of the data sharing requirement beyond those grants requesting 500K of direct costs each year and the inclusion of subcontracts. We support the increasing requirements for documentation of the responsibility needed for data management activities and the identification of the individuals who will be undertaking those duties.

We were pleased to see inclusion of archival repository storage and ongoing preservation costs, which has proved a significant barrier for researchers and their institutions. Allowing for prepayment of long term storage will be a benefit to allowing researchers to create data sets which can have ongoing from support rather than preservation only in fixed media or hard drives which may encounter failures.

We appreciate the recognition of the interest of the public in the research and scholarship funded by the NIH and their desire to better engage with it. We too were pleased to see the explicit collaboration with the tribal communities and the recognition of the need for protecting research from vulnerable populations.

The draft document points to the FAIR Standards. While the premise of these standards is sound, the current standard is hosted on a community

organization website as opposed to a recognized standards organization and could be subject to change in the future before there is a revision of the data management plan requirement document. This could mislead researchers in the future. Instead, we recommend that NIH work with standards organizations to define which principles specifically concern data management reusability rather than relying a potentially changing document.

Section II: Definitions:

While the document calls for the identification of individuals who will have responsibility for data management; this may incorrectly be defaulted to the primary investigator and may not appropriately reflect those who are responsible for the granular data management requirements. In addition to promoting greater transparency of the roles for data management and a clarification of what the definition of "responsible" means, we encourage a move towards standards of expertise and training to ensure that NIH funded researchers have appropriate data management support. Further, there continues to be a need for more systematic data management education for researchers at all levels.

Throughout the document are phrases which are likely to be misused or used as an excuse for noncompliance, such as "as soon as practicable" / "timely" / "as long as it is deemed useful" / "reasonable efforts". It is unclear how the NIH is planning to vet and promote best practices in data sharing, retention, and destruction. We encourage the ICOs to create guidelines which clarify appropriate disciplinary timelines for the datasets generated under their purview.

Section III: Scope:

As the NIH continues their efforts in promoting data reuse, sharing, and preservation, there continues to be a need for infrastructure to meet these demands at the agency and national level in order to prevent the ability to comply with data management and sharing requirements only to those most financially rich institutions who are able to offset the costs of infrastructure. The NIH is not alone in this, as documented in late 2018 by the NSF Bridging the Gap report, which detailed the need for this sort of infrastructure. The NIH has a clearly demonstrated history and current practice of developing infrastructure and mechanisms for specific types of

data sharing. We were encouraged to see the pilot collaboration with figshare and the opportunity for using Amazon Web Services for certain grants. However, there is ongoing need for a centralized repository of biomedical research, particularly that which contains human subject data and other sensitive data that may fall under HIPAA, Tribal or other privacy laws.

Further, as NIH enhances their efforts towards data preservation, reuse, and sharing, one specific charge made in the draft document is that scientific data should be findable. We agree and encourage the development of a supplemental tab to NIH Reporter or a central database which points to the final homes for datasets to enhance and improve discovery. This will allow for enhanced record linking to the National Library of Medicine's current discovery tools, particularly PubMed.

Beyond the discovery, however, is the need to protect the access to this taxpayer funded research data. Already, scientific research is frequently inaccessible to other researchers and the public whose funds supported it due to publisher paywalls. The NIH's Open Access Policy and the enforcement through PubMedCentral was critical in recognizing the inherent inequity of the current publishing standards and moving towards appropriate access. Research data runs similar or even increased potential for paywalls which will exploit access. While there will need to be restricted access to sensitive data and there should be recognition of the ongoing costs of maintaining and providing access to larger datasets, there should also be further guidance and policies from the NIH which will ensure the future accessibility of NIH funded data.

Section IV: Effective Date(s):

Section V: Requirements:

NIH does not presently identify requirements for the repositories where NIH funded and conducted data will be stored and we recommend the implementation of minimal standards for these repositories to encourage reusability. Among the standards should be ways for researchers to document or add new versions; techniques for the original researchers to obtain a copy of their data in the future; minimum metadata standards; fixed URLs (handles) or DOIs; mechanisms for verification of the protection of sensitive data; policies for when data might be deleted; or when more detailed data than what can be stored in a national repositories is required and when can others use.

Section VI: Data Management and Sharing Plans:

One particularly problematic phrase is the allowance for researchers to insert "to be determined" into their plans. This goes counter to what is earlier in the document, which explicitly states that ICOs have the authority to ask for more specific details, and also does not reflect the annual regular review that the NIH program officers will conduct. This phrase may

provide the opportunity for researchers to underestimate the likely costs of data management and preservation, which will impede their ability to create an accurate budget and appropriately account for the labor and costs of providing long term access for reuse and replicability. Researchers should be required to document their stated intentions in the data management plan at the beginning of the grant, knowing that during the annual regular review cycles they are likely to need to make adjustments.

One mechanism described for data sharing is through approval of the requestor by the original researcher. This will be harmful when coming to data sharing. As documented by Wallis et al, researchers approach data from a gift culture perspective – they were willing to share with colleagues they know and trust (Wallis 2013). This combined with other research which has affirmed that researchers are prone to discriminate in their communications, collaborations, and citations based on perceived race, gender, ethnicity introduces the issue that data may only be shared within small privileged circles. We support the use of repositories, including biorepositories, to provide an intermediate to handle the administrative burden, regulatory compliance, and navigation of access. We additionally recommend language to require sharing unless rather than only sharing if.

Section VII: Compliance and Enforcement:

In terms of noncompliance, we would like to see more clarification about data ownership and the obligations of ownership; as ownership is usually retained by the institution receiving the grant. Clarification on what is an acceptable length for embargoing data is also needed so that an embargo is not misused to prevent sharing and advancement of research solely for individual profit.

Within the description of the DMP, NIH acknowledges that legal, ethical, and technical issues might limit the ability of investigators to commit to data sharing. For instance, the likelihood of a breach of confidentiality through data sharing is higher for research that includes individuals from specialized populations such as those with rare diseases, uncommon disabilities, or participants with unique characteristics living in small geographic regions through deductive disclosure. Because of these sensitive issues, and the importance of data-sharing, investigators would benefit from guidance regarding the conditions that might preclude data-sharing, if any, and investment in the development of infrastructure, mechanisms, and procedures to share data for specialized populations.

De-identification is a specific issue that will arise for NIH funded research due to the frequency of human subject data. In order to meet these increased needs so that research data may be appropriately handled and shared, we encourage specific funding to identify improved de-

identification mechanisms and for training researchers and students in use of them in order to properly protect human subjects as data sets become more transparently available.

We also need clear policies and enforcement mechanisms from the NIH about de-identified datasets stored in repositories. Many institutions engage in research of vulnerable populations for whom there is a high risk of re-identification with potential harmful impacts. With de-identified datasets, there is no present federal law governing the use or misuse of this data, which further increases the possible loss of trust with minority and vulnerable populations.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The current draft recommends that the data management plan will not be required for grant applications until the 'Just in Time' period. Among the recommendations for this timing is to prevent required efforts by the program officers in reviewing the programs before the scientific merit has been established. However, this may create several issues.

Most significant among these is the financial impacts of data management may not be fully realized until writing a DMP. The consequences of this is that researchers will not have appropriately addressed the financial and personnel costs required. NIH guidance on allowable budget adjustments at JIT will ensure that sufficient resources for data management have been allocated in order to comply with requirements. We would encourage having NIH program staff review the DMPs to ensure adequate budging for data preservation and reuse and the need for clinical models.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

UIC Comments DMP 2020 01 10.pdf

Description:

UIC Comments 2020 01 10



Comments and Response from the University of Illinois at Chicago

Over the past 15 years, the importance of data sharing, curation, and preservation has grown more explicit. Datasets are increasingly recognized as independent scholarly objects that can be documented in reports to funding agencies.

We are pleased to see the DRAFT NIH Policy for Data Management and Sharing which documents the growing recognition of the need for data curation, preservation, and sharing to enhance biomedical research veracity, impact, and return on tax payer investment. We concur with the NIH's expansion of the data sharing requirement beyond those grants requesting SOOK of direct costs each year and the inclusion of subcontracts. We support the increasing requirements for documentation of the responsibility needed for data management activities and the identification of the individuals who will be undertaking those duties.

We were pleased to see inclusion of archival repository storage and ongoing preservation costs, which has proved a significant barrier for researchers and their institutions. Allowing for prepayment of longterm storage will be a benefit to allowing researchers to create data sets which can have ongoing support rather than preservation only in fixed media or hard drives which may encounter failures.

We appreciate the recognition of the interest of the public in the research and scholarship funded by the NIH and their desire to better engage with it. We too were pleased to see the explicit collaboration with the tribal communities and the recognition of the need for protecting research from vulnerable populations.

There are several points of the suggested draft which may have deleterious impact on the efficacy of the data management plan

- The current draft recommends that the data management plan will not be required for grant applications until the 'Just in Time' period. Among the recommendations for this timing is to prevent required efforts by the program officers in reviewing the programs before the scientific merit has been established. However, this may create several issues.
 - o Most significant among these is the financial impacts of data management may not be fully realized until writing a DMP. The consequences of this is that researchers will not have appropriately addressed the financial and personnel costs required. NIH guidance on allowable budget adjustments at JIT will ensure that sufficient resources for data management have been allocated in order to comply with requirements. We would encourage having NIH program staff review the DMPs to ensure adequate budgeting for data preservation and reuse and the need for clinical models.

- One particularly problematic phrase is the allowance for researchers to insert "to be determined" into their plans. This goes counter to what is earlier in the document, which explicitly states that ICOs have the authority to ask for more specific details, and also does not reflect the annual regular review that the NIH program officers will conduct. This phrase may provide the opportunity for researchers to underestimate the likely costs of data management and preservation, which will impede their ability to create an accurate budget and appropriately account for the labor and costs of providing long term access for reuse and replicability. Researchers should be required to document their stated intentions in the data management plan at the beginning of the grant, knowing that during the annual regular review cycles they are likely to need to make adjustments.
- NIH does not presently identify requirements for the repositories where NIH funded and conducted data will be stored and we recommend the implementation of minimal standards for these repositories to encourage reusability. Among the standards should be ways for researchers to document or add new versions; techniques for the original researchers to obtain a copy of their data in the future; minimum metadata standards; fixed URLs (handles) or DOIs; mechanisms for verification of the protection of sensitive data; policies for when data might be deleted; or when more detailed data than what can be stored in a national repositories is required and when can others use.
- One mechanism described for data sharing is through approval of the requester by the original researcher, This will be harmful when coming to data sharing. As documented by Wallis et al, researchers approach data from a gift culture perspective - they were willing to share with colleagues they know and trust (Wallis 2013). This combined with other research which has affirmed that researchers are prone to discriminate in their communications, collaborations, and citations based on perceived race, gender, ethnicity introduces the issue that data may only be shared within small privileged circles. We support the use of repositories, including biorepositories, to provide an intermediate to handle the administrative burden, regulatory compliance, and navigation of access. We additionally recommend language to require sharing *unless* rather than only sharing *if*.
- In terms of noncompliance, we would like to see more clarification about data ownership and the obligations of ownership; as ownership is usually retained by the institution receiving the grant. Clarification on what is an acceptable length for embargoing data is also needed so that an embargo is not misused to prevent sharing and advancement of research solely for individual profit.
- Within the description of the DMP, NIH acknowledges that legal, ethical, and technical issues might limit the ability of Investigators to commit to data sharing. For instance, the likelihood of a breach of confidentiality through data sharing is higher for research that includes individuals from specialized populations such as those with rare diseases, uncommon disabilities, or participants with unique characteristics living in small geographic regions through deductive disclosure. Because of these sensitive issues, and the importance of data-sharing, investigators would benefit from guidance regarding the conditions that might preclude data-sharing, if any, and investment in the



development of infrastructure, mechanisms, and procedures to share **data** for specialized populations.

- De-identification is a specific issue that will arise for NIH funded research due to the frequency of human subject data. In order to meet these Increased needs so that research data may be appropriately handled and shared, we encourage specific funding to identify improved de-identification mechanisms and for training researchers and students in use of them in order to properly protect human subjects as data sets become more transparently available.
- We also need clear policies and enforcement mechanisms from the NIH about de-identified datasets stored in repositories. Many institutions engage in research of vulnerable populations for whom there is a high risk of re-identification with potential harmful impacts. With de-identified datasets, there is no present federal law governing the use or misuse of this data, which further increases the possible loss of trust with minority and vulnerable populations.
- While the document calls for the Identification of individuals who will have responsibility for data management; this may incorrectly be defaulted to the primary investigator and may not appropriately reflect those who are responsible for the granular data management requirements. In addition to promoting greater transparency of the roles for data management and a clarification of what the definition of "responsible" means, we encourage a move towards standards of expertise and training to ensure that NIH funded researchers have appropriate data management support. Further, there continues to be a need for more systematic data management education for researchers at all levels.
- Throughout the document are phrases which are likely to be misused or used as an excuse for noncompliance, such as "as soon as practicable" / "timely" / "as long as it is deemed useful" / "reasonable efforts". It is unclear how the NIH is planning to vet and promote best practices in data sharing, retention, and destruction. We encourage the ICOs to create guidelines which clarify appropriate disciplinary timelines for the datasets generated under their purview.
- The draft document points to the FAIR Standards. While the premise of these standards is sound, the current standard is hosted on a community organization website as opposed to a recognized standards organization and could be subject to change in the future before there is a revision of the data management plan requirement document. This could mislead researchers in the future. Instead, we recommend that NIH work with standards organizations to define which principles specifically concern data management reusability rather than relying a potentially changing document.

As the NIH continues their efforts in promoting data reuse, sharing, and preservation, there continues to be a need for infrastructure to meet these demands at the agency and national level in order to prevent the ability to comply with data management and sharing requirements only to



those most financially rich institutions who are able to offset the costs of infrastructure. The NIH is not alone in this, as documented in late 2018 by the NSF Bridging the Gap report, which detailed the need for this sort of infrastructure. The NIH has a clearly demonstrated history and current practice of developing infrastructure and mechanisms for specific types of data sharing. We were encouraged to see the pilot collaboration with figshare and the opportunity for using Amazon Web Services for certain grants. However, there is ongoing need for a centralized repository of biomedical research, particularly that which contains human subject data and other sensitive data that may fall under HIPAA, Tribal or other privacy laws.

Further, as NIH enhances their efforts towards data preservation, reuse, and sharing, one specific charge made in the draft document is that scientific data should be findable. We agree and encourage the development of a supplemental tab to NIH Reporter or a central database which points to the final homes for datasets to enhance and improve discovery. This will allow for enhanced record linking to the National Library of Medicine's current discovery tools, particularly PubMed.

Beyond the discovery, however, is the need to protect the access to this taxpayer funded research data. Already, scientific research is frequently inaccessible to other researchers and the public whose funds supported it due to publisher paywalls. The NIH's Open Access Policy and the enforcement through PubMedCentral was critical in recognizing the inherent inequity of the current publishing standards and moving towards appropriate access. Research data runs similar or even increased potential for paywalls which will exploit access. While there will need to be restricted access to sensitive data and there should be recognition of the ongoing costs of maintaining and providing access to larger datasets, there should also be further guidance and policies from the NIH which will ensure the future accessibility of NIH funded data.

References:

- 1) NSF Bridging the Gap: <https://www.nsf.gov/nsb/publications/2018/NSB-2018-40-Midscale-Research-Infrastructure-Report-to-Congress-Oct2018.pdf>
- 2) Wallis, If We Share Data; Will Anyone Use Them? Data Sharing and Reuse In the Long Tail of Science and Technology
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067332#pone-0067332-g001>

Submission ID: 1378

Date: 1/10/2020

Name: Patrick Dunn, Emma Afferton, John Campbell, Henry Schaefer, Elizabeth Thomson

Name of Organization: ImmPort (www.immport.org)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Immunologic data: genomic, transcriptomic, proteomic, metabolomic, imaging, clinical lab tests

Type of Organization: Other

Type of Organization - Other: NIAID contractor

Role: Other

Role - Other: Data curator

Domain of Research Most Important to You or Your Organization:

Immunology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Overall, the NIAID DAIT funded ImmPort team supports the development of the NIH Policy for Data Management and Sharing. From a repository perspective, an effective level of communication and curation efforts are needed between repositories and researchers in order to publicly share research data in a FAIR manner. Requiring researchers to provide Data Management and Sharing Plans prior to conducting research will improve the timeliness of transitioning research data to repositories.

Is NIH considering a timeline to move from a position of encouraging data sharing to expecting robust data sharing plans?

Section II: Definitions:

The definitions for Data Management and Sharing Plan, Data Management, Data Sharing, Metadata do not include a mention of data repository to share the data. The choice of data repository may affect all of the other elements mentioned. Repositories usually have standards for acquiring, curating and sharing data sets and will affect the FAIRness of a data set. This is described in greater detail in the "Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan (Plan)" comments.

The comment in sub-section Scientific Data "Scientific data do not include ...completed case report form" could be clarified to something like "Scientific data do not include ...case report form completed for a research subject".

Section III: Scope:

The scope of the draft policy is sufficiently broad and comprehensive. This suggests the level of effort and time that is needed to implement this policy.

Section IV: Effective Date(s):

There are current ICO program announcements that specify data sharing goals and methods. Initial focus on larger grants to foster the change in community expectations and develop the skills to move the biomedical research enterprise towards integrating data sharing into common practice is an incremental approach and tractable, as demonstrated by NIAID FOAs that highlight data sharing policies.

Section V: Requirements:

Please consider rewording for clarity "Compliance with the NIH ICO-approved Plan" to "Compliance with the Plan once NIH ICO-approved".

Section VI: Data Management and Sharing Plans:

The policy mentions that "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." What metrics or considerations will be used to assess the value of data to the community?

Plan Assessment

What provisions are NIH considering for making available the evaluations of extramural awards, contracts and intramural research projects? Is NIH considering the reviewing of compliance by intra-ICO extramural awards and contracts by intramural panels and intramural projects to be reviewed by extramural teams? Is there an intention to evaluate inter-ICO compliance guidelines and implementations?

Section VII: Compliance and Enforcement:

The elements of evaluating compliance with a data sharing plan should be described in supplemental NIH ICO guidelines. Is NIH considering review of ICO specific compliance guidelines to encourage adoption of best practices across NIH?

The time and effort needed to adequately describe a data set should be included in the evaluation of how effectively a data sharing plan was implemented. This should include the

criteria for choosing data repository(ies), data types to share, metadata standards to describe results, and software and tools used to generate and analyze data sets.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The discussion of "Curating data and developing supporting documentation" is a helpful outline of the curatorial process and highlights the need for allocating resources that may otherwise be committed to data generation, analysis or publication.

"Local data management considerations" might be considered as amortizable costs since robust data management systems can be used over time to facilitate data analysis and sharing requirements.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Since the selection of a data repository(ies) has a significant impact on the types, description, analysis software, accessibility, findability and preservation standards of deposited data, it is recommended that the choice of repositories be a primary factor when designing Data Sharing Plans. We recommend several considerations to be made when choosing a repository. NIH genomic data sharing policy should be considered where appropriate. The ICOs' program announcements may indicate recommended repositories for data sharing (e.g. <https://grants.nih.gov/grants/guide/rfa-files/RFA-AI-17-040.html>). The editorial policies of journals may affect choosing a repository (ex. NAR). Does each repository's data description, findability and access standards meet the expectations of NIH and researchers?

Where research projects generate multiple data modalities, more than one repository may be indicated. For these projects, the ease with which data can be deposited and linked across multiple repositories, the consistency of data access and description standards should be evaluated. As an example, NIAID's ImmPort repository supports the deposition of immunology data modalities, recommends consideration of NIH data sharing policies, and supports links to other repositories indicated by investigators.

It would be helpful if NIH noted that the sharing of negative scientific results that were the result of robust experimental methods are examples of scientific data that are encouraged to be shared even if they are not part of a publication.

The "Data Types" sub-section contains very useful descriptions providing details on what and why data will be shared.

If NIH is encouraging the use of software repositories (e.g. Github), it would be helpful to mention that in the "Related Tools, Software and/or Code" sub-section.

The "Standards" sub-section would benefit by focusing on NIH funded standards projects and the inclusion of ontology examples. The CDE Resource Portal was notable for its absence of mentioning sources of biomedically relevant ontologies.

"DataSharing Agreements, Licenses, and Other Use Limitations" makes the comment "sharing are consistent with community expectations". The "community" is left relatively undefined. Does it include the community of principal investigators who lead the research teams? Does it also include the community of new investigators who would benefit from findable data sets to generate hypotheses? Does it include NIH staff who oversee insightful research programs and are interested in seeing the data shared? The expectations of these communities may differ. To what extent is NIH focusing on the different communities?

"Oversight of Data Management" is a useful reminder that an effective Data Sharing Plan will note who's time will be allocated to data sharing tasks and how much time is planned.

Other Considerations Relevant to this DRAFT Policy Proposal:

The DRAFT Policy Proposal highlights that enhanced data sharing is a goal with evolving expectations. The sociological changes (e.g. attitudes and behavior) in NIH and its researchers that are commensurate and obligatory for effective execution of enhanced data sharing policies are worthy of study in its own right.

Attachment:

Description:

Submission ID: 1379

Date: 1/10/2020

Name: Elisa Hurley

Name of Organization: Public Responsibility in Medicine and Research (PRIM&R)

Type of Data of Primary Interest:

Type of Data of Primary Interest - Other:

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Please see attached comments.

Section II: Definitions:

Please see attached comments.

Section III: Scope:

Please see attached comments.

Section IV: Effective Date(s):

Section V: Requirements:

Please see attached comments.

Section VI: Data Management and Sharing Plans:

Please see attached comments.

Section VII: Compliance and Enforcement:

Please see attached comments.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Please see attached comments.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Please see attached comments.

Other Considerations Relevant to this DRAFT Policy Proposal:

Please see attached comments.

Attachment:

PRIM&R's Comments_January 10_final.pdf

Description:

PRIM&R's comments

Chair

Natalie L. Mays,
BA, LATG, CPIA

Vice Chair

Suzanne Rivera, PhD, MSW

Secretary

Martha Jones, MA, CIP

Treasurer

Owen Garrick, MD, MBA

Board of Directors

Albert J. Allen, MD, PhD

Elizabeth A. Buchanan, PhD

Holly Fernandez Lynch, JD, MBE

Bruce Gordon, MD

Mary L. Gray, PhD

F. Claire Hankenson,
DVM, MS, DACLAM

Karen M. Hansen

Megan Kasimatis Singleton,
JD, MBE, CIP

Jori Leszczynski, DVM, DACLAM

Vickie M. Mays, PhD, MSPH

Gianna McMillan, DBe

Robert Nobles, DrPH, MPH, CIP

Stephen Rosenfeld, MD, MBA

Ex Officio

Elisa A. Hurley, PhD
Executive Director

Submitted electronically at <https://osp.od.nih.gov/draft-data-sharing-and-management/>

Francis S. Collins, MD, PhD
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

RE: Draft NIH Policy for Data Management and Sharing and Supplemental Draft Guidance

Dear Dr. Collins:

Public Responsibility in Medicine and Research (PRIM&R) appreciates the opportunity to comment on the National Institutes of Health (NIH)'s Draft Policy for Data Management and Sharing and Supplemental Draft Guidance, published November 8, 2019.

PRIM&R is a nonprofit organization dedicated to advancing the highest ethical standards in the conduct of research. Since 1974, PRIM&R has served as a professional home and trusted thought leader for the research protections community, including members and staff of human research protection programs and institutional review boards (IRBs), investigators, and their institutions. Through educational programming, professional development opportunities, and public policy initiatives, PRIM&R seeks to ensure that all stakeholders in the research enterprise understand the central importance of ethics to the advancement of science.

PRIM&R strongly agrees with the NIH that sharing data resulting from taxpayer funded research enhances the value of that research, advances the pace of scientific discovery, and, in the case of human subjects research (which will be our focus), maximizes the contributions of human research subjects. We therefore appreciate the NIH's proposal to require NIH-funded researchers to provide a comprehensive plan describing how scientific data will be managed and shared before the launch of a study. **However, we note that NIH's draft Policy for Data Management and Sharing does not articulate a mandate to share such data. We strongly urge the NIH in the final policy to make a clear statement requiring researchers in**

both pre-clinical and clinical research to share their data, unless the agency determines that there is a compelling scientific, ethical, and/or logistical reason to not do so.

Evidence suggests that research subjects are eager to see their data shared and their contributions put to the best use.¹ Even individuals with rare diseases believe their research data should be made available to outside researchers,² despite the heightened privacy risks associated with being part of a smaller or more easily identified population. People participate in research in large part because they believe their contributions will advance science, which is more likely when more researchers are able to access and analyze their data. While there are of course ethical reasons not to share data in some cases—we explore some of these human subject research concerns below—**we believe there should be a rebuttable presumption that data will be shared**. The fact that data will be shared should, in turn, be disclosed in the informed consent process.

1. Review of data sharing plans for privacy and security issues

The NIH has an obligation to facilitate the ethical sharing of data. While we believe the NIH should require that data be shared, we also believe the agency has a simultaneous responsibility to **continue to revisit its practices and policies, in order to set appropriate expectations for the protection of research subjects' data by its grantees**. This should include vetting grantees' proposed data repositories and sharing platforms to ensure they support the secure and ethical sharing of data.

Deidentification is one privacy risk mitigation strategy currently discussed in the supplemental draft guidance. However, it is dangerous to think that deidentification will sufficiently protect research subjects' privacy interests, given that it is no longer possible to guarantee that data will remain permanently deidentified. At the very least, this fact should be appropriately communicated to grantees, oversight bodies, and other relevant stakeholders in both the final policy itself as well as any supplemental draft guidance the NIH develops. We also encourage the NIH to **think creatively about what additional risk mitigation strategies it might suggest**.

According to the current NIH proposal, data management and sharing plans would be required only once an application has gone through peer review and received a “fundable score.” Review of submitted data management and sharing plans will, then, be done by individual program officers throughout the year, following the NIH grant cycle. **We urge the NIH to take additional steps to supplement and support this review process**. One option would be to convene a technical review group that includes individuals who are independent from the NIH and its grant recipients, which could more fully assess and

¹ [Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing](#). Mello, M., Lieou, V. & Goodman, S. (2018). *New England Journal of Medicine*.

² [Share and Protect Our Health Data: An Evidence Based Approach to Rare Disease Patients' Perspectives on Data Sharing and Data Protection - Quantitative Survey and Recommendations](#). Courbier, S., Dimond, R., & Bros-Facer, V. (2019). *Orphanet Journal of Rare Diseases*.

address data security and privacy issues.³ This technical review group could draft guidance documents that program officers could then use to review individual plans, thus standardizing reviews of data security and privacy issues across projects.

Alternatively, **the NIH could consider using such a technical review group as a centralized review entity that would weigh in on the merits of individual data management and sharing plans.** Such a group would be better equipped than individual program officers, or institutions' IRBs, to ensure that research subjects' privacy and security interests are protected. We acknowledge that the proposed policy and timing of review does not currently provide an opportunity for such a robust review process, but believe this approach would be of great benefit to both investigators who create data management and sharing plans, as well as the IRBs who review them.

The aforementioned guidance documents could also be made public and shared with the research oversight community which is struggling with the complexities of this new domain. While IRBs are increasingly aware of the privacy and security risks associated with the sharing, storage, and aggregation of scientific data, most do not have access to privacy and security experts who can advise them on the full range of issues or, most importantly, their mitigation. Ideally IRBs might work with computer scientists and engineers at their respective institutions to identify and respond to basic privacy and security trends; however, the current lack of funding support for such interdisciplinary collaboration makes this approach unlikely.⁴ Until there is a shift in funding incentives, or the field of experts in differential privacy grows, we encourage the to NIH lead the way by creating robust mechanisms for reviewing data management and sharing plans for security and privacy concerns.

2. Areas for further guidance

In response to the NIH's request for areas in which further guidance is needed, PRIM&R suggests the agency offer **specific guidance on the ethical issues involved in data sharing for the research oversight community, including IRBs.** Such guidance should help IRBs ensure that participants are adequately informed of the limits of deidentification and include clear recommendations for how both the facts about data sharing and its inherent risks should be conveyed during the informed consent process. The Common Rule now requires informed consent to include a statement when data collected during a research project will be deidentified for subsequent research use, including that further consent will not be sought for such use. We believe this statement is likely inadequate, given the limits of deidentification when data sets can be aggregated, and encourage the

³ Given the many shortcomings of deidentification, the NIH should consider the merits of differential privacy, which is currently the best option for eliminating any identifying features in a dataset, and consider incorporating differential privacy expertise. Because the number of differential privacy experts is small, we urge the agency to capitalize on their stature in the research field and contract with the few number of experts in this space accordingly.

⁴ [Credit Data Generators for Data Reuse](#), Pierce, H., Dev, A., Statham, E., & Bierer, B. (2019). *Nature*.

NIH's future guidance to better address how to communicate with prospective subjects about the realities and risks of data sharing.

Given the growing number and complexity of issues that institutional research oversight bodies will need to understand and monitor as data sharing efforts expand, we also urge the NIH to revisit the Supplemental Draft Guidance on Allowable Costs for Data Management and Sharing's language on facilities and administrative costs. Currently, the draft guidance states, "Budget estimates should not include infrastructure costs typically included in institutional overhead (e.g., Facilities and Administrative costs)." However, institutional overhead costs may rise with increased efforts to share data; as such, **current Facilities and Administrative allowances may be insufficient to cover increased institutional overhead costs.**

Relatedly, although the agency proposes that NIH budget requests may include costs tied to data curation, making data available in repositories, and local data management considerations, we note that many institutions and research investigators do not have the expertise needed for such efforts. At a minimum, the **NIH needs to provide potential grantees with examples of what kinds of costs they should make requests for**, e.g. what kinds of technology might need to be in place to ensure such efforts are successful. We are concerned that without an explicit, and more descriptive, acknowledgement of what these costs might look like, grantees might not make the appropriate requests for the funding needed for important privacy and security measures.

We also request more **clarification on whether grant funds may be requested in budgets or used for the costs associated with the continued storage and sharing of the data after the research has concluded.** It is presently unclear how researchers would be able to cover the annual costs of a data repository or the costs for deidentifying or processing data for sharing years after the grant is over. Relatedly, the **NIH should issue more guidance about how long they expect data to be available after the grant funding ends.**

3. Other issues

Given the number of complex matters we detail above, we again suggest that the proposed **two-page cap for data management and sharing plans is likely to be impracticable.**

Finally, it would be helpful to get some **clarification from the NIH about the relationship between this NIH-wide policy for Data Management and Sharing Policy and the policies that may be promulgated by specific NIH institutes, centers, and offices.** How much discretion will the separate institutes and centers have to create their own requirements, and how much can those requirements go beyond the NIH-wide policy? While there are no doubt good reasons to allow individual institutes to put in place additional rules, for instance, to protect special populations or particularly sensitive data, **we hope the NIH will consider the logistical difficulties and potential burdens of a**

dataset being subject to a number of different jurisdictions, and encourage harmonization of policies across the NIH as much as possible.

Thank you again for the opportunity to comment and for the NIH's continued work on this important issue. We greatly appreciate that the draft policy includes more language regarding the need to protect the rights and interests of research subjects than the 2018 RFI on the topic, and we hope our comments on the current draft policy will be useful in your next stage of policymaking in this area. PRIM&R stands ready to provide any further assistance or input that might be useful. Please feel free to contact me at 617.303.1872 or ehurley@primr.org.

Respectfully submitted,

A handwritten signature in black ink that reads "Elisa A. Hurley". The signature is written in a cursive style with a large, sweeping initial 'E'.

Elisa A. Hurley, PhD
Executive Director

cc: PRIM&R Public Policy Committee, PRIM&R Board of Directors

Submission ID: 1380

Date: 1/10/2020

Name: Ibraheem Ali

Name of Organization: University of California, Los Angeles

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Sciences Data Librarian

Domain of Research Most Important to You or Your Organization:

Information Management and Scholarly Communication

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

"Data sharing enables researchers to rigorously test the validity of research findings, strengthen analyses through combined datasets, reuse hard-to-generate data, and explore new frontiers of discovery"

We agree with these statements. We would also add that data sharing enables researchers to reuse hard-to-generate research methodologies and protocols.

"Shared data should be made accessible in a timely manner for use by the research community and broader public."

It is unclear what is a "timely manner" when it comes to making the research data available as this may vary across disciplines. Standards set by journals imply that sharing research data upon publication is "timely." Considering there are no clear standards for what is "timely," this may provide justification for researchers to significantly delay publication of their data. We recommend setting clear guidelines as to what timely means.

We also recommend including describing the benefits of sharing research metadata using language such as this:

"Shared metadata helps researchers be aware of what topics are being investigated in the research space. This enables collaboration and helps protect NIH funded researchers from redundant efforts or 'scooping'."

Section II: Definitions:

Under "Metadata" we recommend including "methodological descriptions" as one of the examples. This would be consistent with Supplemental DRAFT Guidance: Elements of a NIH DATA Management and Sharing Plan. This would also highlight the importance of transparent methodologies associated with NIH funded research.

Section III: Scope:

We think the scope is appropriate for this policy.

Section IV: Effective Date(s):

We recommend implementation dates for NIH Intramural research begin 12 months in advance of Extramural research, contracts and other funding agreements and that the NIH uses this period to establish best practices for review and efficacy of Data Management and Sharing Plans for NIH projects and grants.

Section V: Requirements:

These requirements seem acceptable.

It is important to note that some literature indicates that the submission of data management plans are not actually effective in improving data management and sharing (Smale, et al. 2018. BioRxiv. "The History, Advocacy and Efficacy of Data Management Plans"). Notable reasons for the failure of DMPs in improving data management are (1) due to lack of monitoring and remediation of poorly described plans and (2) creation of DMPs from templates results in minimal engagement and a lack of substantive data literacy training for researchers.

Section VI: Data Management and Sharing Plans:

In paragraph 1 the proposed policy states "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." It is unclear how this stipulation is set to be defined as there will be variability between fields. The policy should point some examples that define what it means for research to be "useful to the research community or the public."

Section VII: Compliance and Enforcement:

The NIH should formalize some language stipulating how NIH ICOs will evaluate RPPRs and Data Management and Sharing Plans. We worry that if the committees evaluating grants for a particular award do not take the policy seriously or find Data Management and Sharing Plans onerous to evaluate there will be no substantive enforcement of the policy. This will result in a lack of compliance similar to what happened with the NIH Public Access Policy of 2008.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

"Local data management considerations, such as unique and specialized information infrastructure necessary to provide local management, preservation and access to data, (e.g., before deposit into an established repository."

We recommend that this stipulation explicitly states the funding can cover cloud storage for research information that is not otherwise covered by institutional overhead costs. This will enable researchers with less institutional support to engage in good data storage practices using NIH funds.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

"If certain elements of a Plan have not been determined at the time of submission, an entry of "to be determined" may be acceptable if a justification is provided along with a timeline or appropriate milestone at which a determination will be made."

This stipulation is problematic because it provides applicants a way of opting out of effective data management practices in the event of lack of monitoring for compliance. Some minimum requirements of what what should be stated in a data management plan should be explicitly stated in the policy. Required aspects of the DMP should be drawn from sections 1-6 of the DRAFT guidance. Under required aspects of the Plan we recommend researchers include a plan for managing metadata (protocols and summaries of research questions), data types, related tools, standards (if they exist), modes for short and long-term storage, and mechanisms for oversight.

Other Considerations Relevant to this DRAFT Policy Proposal:

"If certain elements of a Plan have not been determined at the time of submission, an entry of "to be determined" may be acceptable if a justification is provided along with a timeline or appropriate milestone at which a determination will be made."

This stipulation is problematic because it provides applicants a way of opting out of effective data management practices in the event of lack of monitoring for compliance. Some minimum requirements of what what should be stated in a data management plan should be explicitly stated in the policy. Required aspects of the DMP should be drawn from sections 1-6 of the DRAFT guidance. Under required aspects of the Plan we recommend that researchers include a plan for managing metadata (protocols and summaries of research questions), data types, related tools, standards (if they exist), modes for short and long-term storage, and mechanisms for oversight.

Attachment:

Description:

Submission ID: 1381

Date: 1/10/2020

Name: Research Triangle Institute

Name of Organization: Research Triangle Insititute

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: RTI holds data from genomic to basic research to qualitative data based on the breatdt of our scope of research.

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

We address the world's most critical problems with multi-disciplenary science-based solutions to improve the human condition by turning knowledge into practice.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

HHS and NIH have increasingly recognized the vital value of data in advancing biomedical research (ASPE, 2019; NIH, 2018). Our experience in performing scientific research across the gamut of funding mechanisms with the goal of improving human health puts us in strong alignment with the stated purpose of requiring more of the data sharing to increase assessment of research soundness, opportunities for re-analysis or methodological improvement, investigation of new ideas, and maximization of research investment. To reach this vision, data management across the lifecycle must be baked into any project from the outset. The purpose describes NIH's intent to require a Data Management and Sharing Plan (Plan) for all NIH-funded scientific research. We applaud this step but would support NIH taking a stronger position to advance the transformative potential of data management and sharing.

First, it is unclear how this new Plan requirement differs in intent from the Resource Sharing Plan that is currently included in the Grant Applications Plan. In general, the implications of these Plans depend largely on requirements for plan acceptability. It might be a meaningful extension, for example, for NIH to require that the Data Sharing Plans define metadata needed to support validation and replication. Second, NIH describes that the data sharing should be

completed in a "timely manner." We would encourage NIH to set some further guidelines around what would be expected to make an action "timely." Third, we assume that expansion of the Plan requirement to include all mechanisms of scientific funding represents a first step along a path toward increased management and sharing, yet a mission and vision for such a path is not described. We assume that this policy is a component of the broader NIH Strategic Plan for Data Science, but the relationship between the two is not explicit in the draft policy.

Section II: Definitions:

The current definitions are very useful in clarifying the Request for Information language. We would like to suggest adding definitions for the following:

- Scientific reproducibility, validity, or rigor
- Data standards
- Data harmonization

Section III: Scope:

NIH provides a clear description of who is expected to submit a Plan. What is less clear is the scope of the Plan—particularly, if steps should be included between data collection and finalized datasets (e.g., data cleaning or data normalization). Reuse or reproducibility efforts often require additional insight into data preparation steps that are frequently excluded in metadata descriptions or data sharing. We encourage NIH to consider highlighting the importance of addressing these steps, where applicable.

Section IV: Effective Date(s):

The language in the draft indicates that NIH intends the policy to apply prospectively only. This is the only reasonable approach; however, linking a potential timeline for Policy implementation to a time-bounded vision of NIH data management and sharing would help convey a general understanding of how NIH may proceed in the coming years. A timeline that includes a rough indication of other related activities that NIH anticipates, such as Policy review and assessment would be a welcome addition and assist in understanding how the Policy fits within NIH's vision.

Section V: Requirements:

We encourage NIH to consider setting specific expectations for Plans, which is likely to differ depending on the type of research and associated data. These expectations would inform the acceptability of a Plan. For example, the Irish Health Research Board has a policy of data sharing as the default, unless specific reasons (such as privacy and consent issues) make the research data exempt (RDA, 2019). It is important that the NIH expectations include more detailed description of the scope of sharing within a study. The experience of seeking to reuse data in the database of Genotypes and Phenotypes (dbGaP) shows that studies vary widely in the

phenotype or exposure variables that study investigators have elected to share. A restricted variable set hugely limits the utility of the sharing effort. Further direction from NIH on the expectation for "timely" sharing as well as repository or sharing solutions should be included to support meaningful interpretation of the policy.

The importance of recognizing the potential impact for the data management and sharing activities on a project's budget is critical. It is our current understanding of the Policy that the cost ceiling of grants will not be increased to accommodate any additional budget needed to support data management and sharing activities. If this is the case, the concern is that quality Plans and implementation of those Plans will come at the cost of research activities. This creates a disturbing incentive for a Plan to be just good enough to pass as "acceptable" in order to reserve funding for scientific investigation. Furthermore, in contracts and other funding agreements, we assume that budgets may increase to include data management and sharing activities; given the cost competition of these types of awards, however, we are concerned that submitters may again choose the minimum acceptable solutions. While RTI is a proponent for efficient spending in any award, we ask NIH to consider the impact of additional required activities within the current funding ceiling and cost competition for minimum acceptability of both grants and contracts, respectively. One possible solution is for NIH to again consider setting expectations for cost to support "acceptable" Plans for different types of research. For example, 10% of an R01 might set a reasonable cost signal to the community.

Section VI: Data Management and Sharing Plans:

In grant applications, RTI's general experience with NIH Genomic Data Sharing Plans is that these Plans are often thin and lack the robustness to truly ensure that data are meaningfully shared. In our view, meaningful sharing requires careful planning from the outset, documentation of provenance, use of clear and descriptive metadata, and responsible submission times. Excluding Plans from the technical evaluation of the grant suggests that NIH considers data stewardship to be of secondary importance to the scientific and technical merits of the proposal. Furthermore, the assessment of Plans during the technical evaluation of contracts seems uneven as compared to the consideration of Plans for grants.

The assessment approach described in the Policy raises the potential risk of varying interpretation and enforcement by individual NIH ICOs. While flexibility for the Plans to fit the research endeavor is critical, dramatic differences in expectations or enforcement on the NIH side could impede data management and sharing behaviors by the extramural community. Pragmatically, we suggest that NIH considers requiring Plans to be submitted, as the Policy describes, as a part of the Just-in-Time, as an interim step toward eventual inclusion into the technical evaluation. We similarly suggest that Plan activities be included in the NIH ICO progress reports.

There is little in the Policy to describe the assessment for intramural opportunities or those funded as Other Transactions (OTs). The NIH intramural community has traditionally developed some seminal cohorts and data resources that have not been shared and have not been replicated extramurally. The importance of bringing some of these high-value resources to the public cannot be overstated. As compared to the intramural program, scientific research data funded by OT are relatively new. The use of OT mechanisms was called out in the NIH Strategic Plan for Data Science as supporting partnerships for deposition, storage, and access of high-value data. Even if OTs are not generally responsible for data generation, the management and sharing responsibilities in projects like the NIH Data Commons Pilot Phase or the National Heart, Lung, and Blood Institute BioData Catalyst are critical. We would advocate that NIH consider taking an approach similar to that of contracts.

Section VII: Compliance and Enforcement:

While we understand that compliance and enforcement are likely to be decided on a case-by-case basis, the language in the Policy does not provide much insight into how NIH is likely to address any issues. Enforcement of the Genomic Data Sharing Plans has generally been sparse. Furthermore, most data sharing necessarily happens after the project is complete, at which time NIH has limited recourse in the current scheme. One option might be for NIH to consider reward or recognition mechanisms in addition to punitive action; such incentives could encourage the desired behavior, particularly after an award has ended.

There are circumstances where researchers may encounter obstacles when attempting to share their data in good faith. Some communities, for example, lack institutional repositories or domain repositories, or may have nascent repositories where the data provider may either need to bear the cost of hosting the data and ancillary information themselves, or may have a degree of uncertainty in the sustainability of the available repository. These scenarios require consideration of funding—either to the institution, researcher, or repository—to ensure long-term access. Additionally, particular domains produce data that are particularly costly to host (e.g., imaging). The burden to share these data could potentially disadvantage researchers with constrained budgets or limited institutional support. Given NIH's view that data sharing can have a democratizing effect on scientific contributions (Brennan, 2019), we encourage consideration for allowing supplemental support for data management and sharing activities, particularly in grant-funded projects where investigators may genuinely struggle to comply with domain best practices, so as not to further contribute to research disparities.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We suggest that NIH consider NIH grant support for data access fees, the allowability of which is unclear. It is also unclear if data management and sharing support costs should be

categorized as direct or indirect costs. This designation will inform the budget available for the grant (i.e., direct veiling vs. total ceiling). Additionally, we recommend consideration of an information-gathering exercise across diverse domains to benchmark the cost of appropriate and responsible data management and sharing.

Per our comments on Section 7, we would advocate that NIH consider the possibility of supplementary funds, particularly for grant-funded work, to support specific, costly data management or sharing needs such as de-identification or uploading to the cloud or an alternative data host.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

N/A

Other Considerations Relevant to this DRAFT Policy Proposal:

RTI International is pleased to provide input on the Draft Policy for Data Management and Sharing in support of the agency's long-standing goal to make scientific research results publicly available. RTI is an independent, nonprofit research institute dedicated to improving the human condition. RTI has almost 5,000 staff worldwide and approximately 3,000 ongoing projects across a range of clients; the largest being the U.S. Department of Health and Human Services (HHS). We answer questions for our clients that demand an objective and multidisciplinary approach—one that integrates expertise across the biomedical and laboratory sciences, engineering, statistics, epidemiology, and social sciences. We have been conducting federal scientific research successfully for over 60 years. Notable scientific achievements include the development of Taxol[®] and camptothecin[™], speech processing developments for cochlear implants, and novel survey technologies, along with advancements in novel survey technologies and in HIV and tuberculosis prevention and treatment programs.

RTI staff are well-versed in the requirements, obligations, and restrictions of government research through a variety of mechanisms, including contract, grants, and Other Transactional Agreements. Our researchers, engineers, and technologists bridge the gap between fundamental research and its application to complex, real-world challenges, with an inherent focus on treating data as first-class research objects. Our technical solutions blend expertise in research and information technology to enhance the collection, management, processing, and dissemination of scientific information.

Strong relationships with research universities, industry partners, and government stakeholders over RTI's 60-year history of domestic and international work afford RTI a unique perspective to provide insightful, experience-based input on the proposed National Institutes of Health (NIH)

Draft Policy for Data Management and Sharing (Policy). RTI International appreciates the opportunity to respond to the questions issued in this round of comments and looks forward to remaining engaged in the conversation as it progresses.

WORKS CITED

Assistant Secretary for Planning and Innovation (ASPE). (2019, January 30). Strategic plan FY 2018–2022. Washington, DC: U.S. Department of Health and Human Services. www.hhs.gov/about/strategic-plan/index.html. Accessed January 6, 2020.

Brennan, P. (2019, July 2). Democratizing information access. NLM Musings from the Mezzanine. Bethesda, MD: U.S. National Library of Medicine, National Institutes of Health. nlmdirector.nlm.nih.gov/2019/07/02/democratizing-information-access/. Accessed January 7, 2020.

National Institutes of Health (NIH). (2018, June 4). NIH Releases Strategic Plan for Data Science. Bethesda, MD: NIH Office of Communications and Public Liaison, U.S. Department of Health and Human Services. www.nih.gov/news-events/news-releases/nih-releases-strategic-plan-data-science. Accessed 6 January 6, 2020.

Research Data Alliance (RDA) (2019, July 12). Ireland’s National Framework on the Transition to an Open Research Environment Launched—RDA Ireland and NLI among contributing and endorsing organisations. Oxford, United Kingdom: RDA. www.rd-alliance.org/irelands-national-framework-transition-open-research-environment-launched-rda-ireland-and-nli-among. Accessed January 7, 2020.

Attachment:

RTI_Response_NIH_DataMgmtSharing.pdf

Description:

Response to the DRAFT NIH Policy for Data Management and Sharing

Submitted By

RTI International

P.O. Box 12194

Research Triangle Park, NC

27709-2194

<http://www.rti.org/>

Introduction

RTI International is pleased to provide input on the Draft Policy for Data Management and Sharing in support of the agency's long-standing goal to make scientific research results publicly available. RTI is an independent, nonprofit research institute dedicated to improving the human condition. RTI has almost 5,000 staff worldwide and approximately 3,000 ongoing projects across a range of clients; the largest being the U.S. Department of Health and Human Services (HHS). We answer questions for our clients that demand an objective and multidisciplinary approach—one that integrates expertise across the biomedical and laboratory sciences, engineering, statistics, epidemiology, and social sciences. We have been conducting federal scientific research successfully for over 60 years. Notable scientific achievements include the development of Taxol® and camptothecin™, speech processing developments for cochlear implants, and novel survey technologies, along with advancements in novel survey technologies and in HIV and tuberculosis prevention and treatment programs.

RTI staff are well-versed in the requirements, obligations, and restrictions of government research through a variety of mechanisms, including contract, grants, and Other Transactional Agreements. Our researchers, engineers, and technologists bridge the gap between fundamental research and its application to complex, real-world challenges, with an inherent focus on treating data as first-class research objects. Our technical solutions blend expertise in research and information technology to enhance the collection, management, processing, and dissemination of scientific information.

Strong relationships with research universities, industry partners, and government stakeholders over RTI's 60-year history of domestic and international work afford RTI a unique perspective to provide insightful, experience-based input on the proposed National Institutes of Health (NIH) Draft Policy for Data Management and Sharing (Policy). RTI International appreciates the opportunity to respond to the questions issued in this round of comments and looks forward to remaining engaged in the conversation as it progresses.

Response to Sections

Section 1. Purpose

HHS and NIH have increasingly recognized the vital value of data in advancing biomedical research (ASPE, 2019; NIH, 2018). Our experience in performing scientific research across the gamut of funding mechanisms with the goal of improving human health puts us in strong alignment with the stated purpose of requiring more of the data sharing to increase assessment of research soundness, opportunities for re-analysis or methodological improvement, investigation of new ideas, and maximization of research investment. To reach this vision, data management across the lifecycle must be baked into any project from the outset. The purpose describes NIH's intent to require a Data Management and Sharing Plan (Plan) for all NIH-funded scientific research. We applaud this step but would support NIH taking a stronger position to advance the transformative potential of data management and sharing.

First, it is unclear how this new Plan requirement differs in intent from the Resource Sharing Plan that is currently included in the Grant Applications Plan. In general, the implications of these Plans depend largely on requirements for plan acceptability. It might be a meaningful extension, for example, for NIH to require that the Data Sharing Plans define metadata needed to support validation and replication. Second, NIH describes that the data sharing should be completed in a "timely manner." We would encourage NIH to set some further guidelines around what would be expected to make an action "timely." Third, we assume that expansion of the Plan requirement to include all mechanisms of scientific funding represents a first step along a path toward increased management and sharing, yet a mission and vision for such a path is not described. We assume that this policy is a component of the broader NIH Strategic Plan for Data Science, but the relationship between the two is not explicit in the draft policy.

Section 2. Definitions

The current definitions are very useful in clarifying the Request for Information language. We would like to suggest adding definitions for the following:

- Scientific reproducibility, validity, or rigor
- Data standards
- Data harmonization

Section 3. Scope

NIH provides a clear description of who is expected to submit a Plan. What is less clear is the scope of the Plan—particularly, if steps should be included between data collection and finalized datasets (e.g., data cleaning or data normalization). Reuse or reproducibility efforts often require additional insight into data preparation steps that are frequently excluded in metadata descriptions or data sharing. We encourage NIH to consider highlighting the importance of addressing these steps, where applicable.

Section 4. Effective Date(s)

The language in the draft indicates that NIH intends the policy to apply prospectively only. This is the only reasonable approach; however, linking a potential timeline for Policy implementation to a time-bounded vision of NIH data management and sharing would help convey a general understanding of how NIH may proceed in the coming years. A timeline that includes a rough indication of other related activities that NIH anticipates, such as Policy review and assessment would be a welcome addition and assist in understanding how the Policy fits within NIH's vision.

Section 5. Requirements

We encourage NIH to consider setting specific expectations for Plans, which is likely to differ depending on the type of research and associated data. These expectations would inform the acceptability of a Plan. For example, the Irish Health Research Board has a policy of data sharing as the default, unless specific reasons (such as privacy and consent issues) make the research data exempt (RDA, 2019). It is important that the NIH expectations include more detailed description of the scope of sharing within a study. The experience of seeking to reuse data in the database of Genotypes and Phenotypes (dbGaP) shows that studies vary widely in the phenotype or exposure variables that study investigators have elected to share. A restricted variable set hugely limits the utility of the sharing effort. Further direction from NIH on the expectation for "timely" sharing as well as repository or sharing solutions should be included to support meaningful interpretation of the policy.

The importance of recognizing the potential impact for the data management and sharing activities on a project's budget is critical. It is our current understanding of the Policy that the cost ceiling of grants will not be increased to accommodate any additional budget needed to support data management and sharing activities. If this is the case, the concern is that quality Plans and implementation of those Plans will come at the cost of research activities. This creates a disturbing incentive for a Plan to be just good enough to pass as "acceptable" in order to reserve funding for scientific investigation. Furthermore, in contracts and other funding agreements, we assume that budgets *may* increase to include data management and sharing activities; given the cost competition of these types of awards, however, we are concerned that submitters may again choose the minimum acceptable solutions. While RTI is a proponent for efficient spending in any award, we ask NIH to consider the impact of additional required activities within the current funding ceiling and cost competition for minimum acceptability of both grants and contracts, respectively. One possible solution is for NIH to again consider setting expectations for cost to support "acceptable" Plans for different types of research. For example, 10% of an R01 might set a reasonable cost signal to the community.

Section 6. Data Management and Sharing Plans

In grant applications, RTI's general experience with NIH Genomic Data Sharing Plans is that these Plans are often thin and lack the robustness to truly ensure that data are meaningfully shared. In our view, meaningful sharing requires careful planning from the outset, documentation of provenance, use of clear

and descriptive metadata, and responsible submission times. Excluding Plans from the technical evaluation of the grant suggests that NIH considers data stewardship to be of secondary importance to the scientific and technical merits of the proposal. Furthermore, the assessment of Plans during the technical evaluation of contracts seems uneven as compared to the consideration of Plans for grants.

The assessment approach described in the Policy raises the potential risk of varying interpretation and enforcement by individual NIH ICOs. While flexibility for the Plans to fit the research endeavor is critical, dramatic differences in expectations or enforcement on the NIH side could impede data management and sharing behaviors by the extramural community. Pragmatically, we suggest that NIH considers requiring Plans to be submitted, as the Policy describes, as a part of the Just-in-Time, as an interim step toward eventual inclusion into the technical evaluation. We similarly suggest that Plan activities be included in the NIH ICO progress reports.

There is little in the Policy to describe the assessment for intramural opportunities or those funded as Other Transactions (OTs). The NIH intramural community has traditionally developed some seminal cohorts and data resources that have not been shared and have not been replicated extramurally. The importance of bringing some of these high-value resources to the public cannot be overstated. As compared to the intramural program, scientific research data funded by OT are relatively new. The use of OT mechanisms was called out in the NIH Strategic Plan for Data Science as supporting partnerships for deposition, storage, and access of high-value data. Even if OTs are not generally responsible for data generation, the management and sharing responsibilities in projects like the NIH Data Commons Pilot Phase or the National Heart, Lung, and Blood Institute BioData Catalyst are critical. We would advocate that NIH consider taking an approach similar to that of contracts.

Section 7. Compliance and Enforcement

While we understand that compliance and enforcement are likely to be decided on a case-by-case basis, the language in the Policy does not provide much insight into how NIH is likely to address any issues. Enforcement of the Genomic Data Sharing Plans has generally been sparse. Furthermore, most data sharing necessarily happens after the project is complete, at which time NIH has limited recourse in the current scheme. One option might be for NIH to consider reward or recognition mechanisms in addition to punitive action; such incentives could encourage the desired behavior, particularly after an award has ended.

There are circumstances where researchers may encounter obstacles when attempting to share their data in good faith. Some communities, for example, lack institutional repositories or domain repositories, or may have nascent repositories where the data provider may either need to bear the cost of hosting the data and ancillary information themselves, or may have a degree of uncertainty in the sustainability of the available repository. These scenarios require consideration of funding—either to the institution, researcher, or repository—to ensure long-term access. Additionally, particular domains produce data that are particularly costly to host (e.g., imaging). The burden to share these data *could* potentially

disadvantage researchers with constrained budgets or limited institutional support. Given NIH's view that data sharing can have a democratizing effect on scientific contributions (Brennan, 2019), we encourage consideration for allowing supplemental support for data management and sharing activities, particularly in grant-funded projects where investigators may genuinely struggle to comply with domain best practices, so as not to further contribute to research disparities.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

We suggest that NIH consider NIH grant support for data access fees, the allowability of which is unclear. It is also unclear if data management and sharing support costs should be categorized as direct or indirect costs. This designation will inform the budget available for the grant (i.e., direct veiling vs. total ceiling). Additionally, we recommend consideration of an information-gathering exercise across diverse domains to benchmark the cost of appropriate and responsible data management and sharing.

Per our comments on Section 7, we would advocate that NIH consider the possibility of supplementary funds, particularly for grant-funded work, to support specific, costly data management or sharing needs such as de-identification or uploading to the cloud or an alternative data host.

Works Cited

Assistant Secretary for Planning and Innovation (ASPE). (2019, January 30). *Strategic plan FY 2018–2022*. Washington, DC: U.S. Department of Health and Human Services. www.hhs.gov/about/strategic-plan/index.html. Accessed January 6, 2020.

Brennan, P. (2019, July 2). Democratizing information access. *NLM Musings from the Mezzanine*. Bethesda, MD: U.S. National Library of Medicine, National Institutes of Health. nlmdirector.nlm.nih.gov/2019/07/02/democratizing-information-access/. Accessed January 7, 2020.

National Institutes of Health (NIH). (2018, June 4). *NIH Releases Strategic Plan for Data Science*. Bethesda, MD: NIH Office of Communications and Public Liaison, U.S. Department of Health and Human Services. www.nih.gov/news-events/news-releases/nih-releases-strategic-plan-data-science. Accessed 6 January 6, 2020.

Research Data Alliance (RDA) (2019, July 12). Ireland's National Framework on the Transition to an Open Research Environment Launched—RDA Ireland and NLI among contributing and endorsing organisations. Oxford, United Kingdom: RDA. www.rd-alliance.org/irelands-national-framework-transition-open-research-environment-launched-rda-ireland-and-nli-among. Accessed January 7, 2020.

Submission ID: 1382

Date: 1/10/2020

Name: Jennifer A Doherty and Cornelia Ulrich

Name of Organization: Huntsman Cancer Institute

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Epidemiologic data

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Cancer research, population, clinical and basic science research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Rather than allowing variation across ICO's for data management and sharing policies, it makes sense to have a single policy to ensure consistency across the NIH. Consistency in the requirements and review process of data management and sharing plans is important.

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

1. It is important to provide guidance on how to handle studies in which some data may have been previously collected (either by an NIH grant or some other source of funding) but some new data may be generated as part of the award. What data have to be described in the data sharing plan and how can investigators address different expectations across varying funding sources.
2. The amount of information provided is very extensive and creates substantial non-science related work for the PIs who submit grants. Considering the low success rate of grant applications, especially at the NCI, perhaps a very brief data sharing plan should be part of the

application, and the full/detailed data sharing plan could be a second step, along with the JIT/Other Support documents.

Section VII: Compliance and Enforcement:

If there is non-compliance with data sharing after the funding/support period, this should not impact NIH consideration of funding for the recipient institution (as written). It will be not manageable and feasible for institutions to understand the details of all ongoing projects at their institutions and oversee appropriate data sharing. Data sharing and the related enforcement actions should be at the PI level.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1383

Date: 1/10/2020

Name: Christopher Carr

Name of Organization: RSNA

Type of Data of Primary Interest: Imaging

Type of Data of Primary Interest - Other:

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

Medical Imaging

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

RSNA concurs with the general purpose of the draft NIH Policy for Data Management and Sharing, particularly the intent to encourage data sharing in the health research community under consistent practices protecting the security and privacy of health information.

Section II: Definitions:

RSNA concurs with the recommended additions to the definitions section of the draft Policy submitted by AMIA, i.e., the addition of concepts for "Data Management," "Covered Data," "Covered Timeframe," and the refinement of definitions for "Metadata" and "Scientific Data." We also strongly support AMIA's proposed addition of "Scientific Software Artifacts" as a defined term and further recommend that "imaging acquisition or post-processing methods and algorithms" be included as examples of such artifacts.

Section III: Scope:

RSNA concurs that the Policy should apply to all research funded or conducted by NIH.

Section IV: Effective Date(s):

RSNA supports the recommendation submitted by AMIA that NIH consider a phased timeline for implementation of the requirements of the DMSP based on funding levels of research.

Section V: Requirements:

RSNA concurs with the essential requirements for submission of a Data Management and Sharing Plan (DMSP) and compliance with the NIH ICO-approved Plan. RSNA also supports the recommendation submitted by AMIA that the DMSP should be a peer-reviewed and scored element of funding requests in order to foster and reward adherence to best practices, as well as innovation in data sharing practices.

Section VI: Data Management and Sharing Plans:

RSNA generally concurs with the overview and outline of elements of DMSPs. We also strongly support AMIA's recommendation that, in addition to monitoring conformance to an approved DMSP, NIH should provide recognition to researchers, institutions and data repositories that share data according to FAIR principles and the NIH Data Science Strategic Plan. Recognized innovations should include making data sharing more cost efficient in medical imaging by implementing fluid data sharing models such as sharing managed links to image data, thus eliminating the need to copy vast stores of image data.

Section VII: Compliance and Enforcement:

RSNA generally supports the proposed provisions regarding Compliance and Enforcement. We also concur with the recommendation submitted by AMIA that a process should be implemented for updating and versioning the DMSP over the course of a funded research program.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

RSNA Comments on NIH Policy for Data Management and Sharing 2020-01-10.pdf

Description:

Formal Comment Letter

January 10, 2020

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
Office of Science Policy
National Institutes of Health

Submitted electronically at: <https://osp.od.nih.gov/draft-data-sharing-and-management/>

Re: DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Dear Dr. Wolinetz:

The Radiological Society of North America (RSNA[®]) is pleased to offer responses to National Institutes of Health's (NIH) request for comment on the Draft Policy for Data Management and Sharing. The RSNA is an international society of radiologists, medical physicists and other medical professionals with more than 54,000 members from 136 countries across the globe. The RSNA promotes excellence in patient care and health care delivery through education, research and technologic innovation.

RSNA strongly supports NIH efforts to develop standard policies and introduce incentives for research data management and sharing. Expanded access to quality research data will improve the rigor and transparency of scientific research and spur the pace of scientific discovery.

In response to the Draft Policy, we wish to affirm the vital role that medical imaging plays in a wide range of medical research and urge NIH to consider sharing of imaging data for research among its highest priorities. Aided by funding from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), RSNA established the Quantitative Imaging Biomarkers Alliance[®] (QIBA[®]) and the Quantitative Imaging Data Warehouse. This effort provides access to a growing and diverse body of high-quality image data sets in order to facilitate development, validation and implementation of quantitative imaging biomarkers. In another NIBIB-funded project, RSNA established the Image Share Network, which enables exchange of medical images and reports for clinical and research purposes. By providing incentives for researchers to share imaging data through such facilities, NIH could accelerate their growth and multiply their value for science and clinical care.

As to specific policy mechanisms to create incentives for data sharing, we would echo the response to this RFI provided by the American Medical Informatics Association (AMIA) with

whom we have consulted in preparing our own response. In particular, we endorse AMIA's recommendations that NIH should:

- 1) Make data sharing plans a "scoreable" element of grant applications subject to the existing policy,
- 2) Earmark support for data sharing as of part of applicable grants' direct costs and
- 3) Develop mechanisms that enable institutional rewards for researchers and institutions that create and/or contribute to public datasets and software that other researchers find useful.

We refer you to the AMIA response for detailed rationale and further elaboration of these recommendations. We provide our responses to specific elements of the draft policy below.

Thank you for the opportunity to share these views. We hope that this discussion and our collective efforts will bring a higher quality of care and better health outcomes to our patients.

Sincerely,



Bruce G. Haffty, MD
Chairman, RSNA Board of Directors



Curtis P. Langlotz, MD, PhD
Liaison for Information Technology and Annual Meeting, RSNA Board of Directors

AcademyHealth Comments on the NIH Draft Data Management and Sharing Policy and Supplemental Guidance

Submitted by Dr. Lisa Simpson, President and CEO

January 10, 2020

AcademyHealth represents 4,000 individuals and organizations in the research community using evidence and data to improve health and health care for all. Our organization recognizes the crucial role of data sharing in advancing scientific research, and ultimately improving clinical care and patient outcomes. We commend the National Institutes of Health for their efforts to optimize access to and use of shared data in research. Below, we provide our comments and suggestions on NIH's draft Policy for Data Management and Sharing. These comments intend to build on and reinforce [our feedback](#) provided in December 2018 on NIH's proposed provisions for this policy. At a high level, our feedback offers suggestions for enhancing clarity and specificity with respect to terminology, requirements, and components of a data management and sharing plan (Plan), and perspective on the use of resolute language to effectively communicate NIH's expectations, and commitment to upholding Plans to rigorous scientific standards.

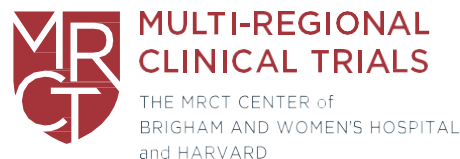
Section I: Purpose

AcademyHealth believes it is crucial to pay careful attention to context and rationale provided for the purpose of data sharing. The argument to support enhanced data sharing must be compelling and it needs to resonate with constituents to garner support. The first sentence of the policy describes NIH's commitment to "making results and outputs of the research it funds and conducts available to the public." We would argue that the scope of data management and sharing should be broadened to encompass data creation and collection, and not solely focus on results and research outputs. At the outset, the policy should state not only the role of shared data in optimizing research results, but also underscore broader implications for the field, including enhanced collaboration, transparency and accountability.

In terms of setting context, how the benefits of data sharing are communicated, including the order in which they are described, are important. Currently, "to test the validity of research findings" is listed as the first (and presumably most important) benefit of data sharing. We would argue the benefits may be better expressed tied to the goals of NIH, and the value of reusing data for deeper or broader discovery should be the emphasis.

Further, while FAIR principles are valuable for considering data sharing, it is unclear whether investigators subject to the proposed policy for data management and sharing are responsible for each of these principles. We believe that the focus of sharing should be to make the data accessible and interoperable, and the role of the NIH is to make data findable and usable. As currently written, investigators sharing the data take on the full burden, which may be counter-productive. We suggest revising the plan such that it diminishes investigator burden through technical, infrastructure and cost sharing across the spectrum of stakeholders in the publication and sharing of data.

Section II: Definitions



January 7, 2020

Francis S. Collins, MD, PhD
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892
Submitted electronically: <https://osp.od.nih.gov/draft-data-sharing-and-management/>

RE: DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Dear Dr. Collins:

The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard (MRCT Center) appreciates the opportunity to comment on the National Institutes of Health (NIH) draft NIH "Policy for Data Management and Sharing and Supplemental DRAFT Guidance" (hereinafter the "Policy"), published in the Federal Register Vol. 84, No. 217 on November 8, 2019.

The MRCT Center is a research and policy center that addresses the ethics, conduct, oversight, and regulatory environment of international, multi-site clinical trials. Founded in 2009, it functions as a neutral convener to engage diverse stakeholders from industry, academia, patients and patient advocacy groups, non-profit organizations, and global regulatory agencies. The MRCT Center focuses on pre-competitive issues, to identify challenges and to deliver ethical, actionable, and practical solutions for the global clinical trial enterprise. Over the last five years, the MRCT Center has been intimately involved in data sharing, including (1) developing guidance for sharing aggregate plain language summaries for participants and the public, (2) developing guidance for sharing individual results with participants, (3) promoting principles of individual participant data (IPD) sharing including protections of patient/participant confidentiality and privacy and of confidential commercial information, (4) developing template data use agreements and data contributor agreements for IPD and other data sharing, (5) crafting informed consent language to promote participant understanding of the implications of sharing de-identified data, (6) launching Vivli, a platform for global data sharing of IPD data, and (7) furthering the establishment of credit for data sharing for those individuals who choose to share their data, among other efforts. Of note, the responsibility for the content of this document rests

with the leadership of the MRCT Center, not with the its collaborators, nor with the institutions affiliated with the authors.¹

The MRCT Center strongly endorses the NIH draft policy and the importance that it places on data management and data sharing. This draft policy demonstrates an ongoing appreciation by the NIH of the utility and value of previously collected data and metadata not only for replication but for new discoveries. Further, proper stewardship of data is important, and the requirement for the submission of data management and data sharing plans prior to initiation of the research will be helpful in that regard. We are enthusiastic that NIH has taken this further step to include all scientific data (and metadata) as defined, of all data types and all sizes, and for all research funded by the NIH. We also understand that the NIH has outlined only the minimum expectations for NIH-wide Plans, and that the NIH ICOs may add additional requirements or expectations. We believe, however, that the NIH policy should be stronger, while nevertheless still permitting some flexibility.

We feel strongly that the **NIH should require data sharing, unless there is an ethical, scientific, or other defensible reason not to do so.** There should be a rebuttable presumption to share data; the burden should be on the investigator to provide cogent reasons that the data should not or cannot be shared. Subjective evaluations by investigators of potential data utility to the research community or the public should not be considered a sufficient reason not to share data.

There are risks to data sharing, including that of participant and patient privacy for studies that involve human participants and their data or biospecimens. Not all data need be downloadable and freely accessible: **measures to protect privacy and confidentiality** should be required. Those measures include de-identification, as mentioned in the draft policy, but also include other risk mitigation strategies: physical and technical security measures (e.g. data maintained in a repository, in a fit-for-purpose compute environment and not downloadable), controlled data access by qualified users, and other more novel methods (e.g. differential privacy, block chain technologies, etc.). We encourage the NIH to invest in the development and dissemination of these technologies to promote data sharing of sensitive data, and to issue appropriate guidance for their use. We further encourage the NIH to require disclosure of—and explanation of—data sharing plans to research participants during the informed consent process.

We encourage the NIH to provide **minimum expectations** for data management and scientific data, either within the policy or as additional guidance. The breadth of research and data acquisition supported by the NIH is expansive, covering different disciplines and including the spectrum of basic, translational, and clinical research. Guidance is needed to

¹ Brigham and Women's Hospital, Rope & Gray LLP, Harvard Medical School, Harvard University, and Yale Law School.

assist investigators and institutions, many unfamiliar with optimal data management and data sharing approaches.

Specific, required elements of the Plan should be developed, and an approximate (or “not to exceed”) **time frame** regarding when the data will be made available should be stated. The completeness and sufficiency of the Plan will only be encouraged by written detail.

We appreciate the development of the Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (Plan). While the descriptions of the specific data elements provide the reader with guidance on the development of a Plan, we encourage NIH to further complement this guidance with examples of (potential) comprehensive data sharing plans for different data types.

We also encourage NIH to provide **minimum expectations for data repositories and data sharing platforms** that meet requirements of the policy. We encourage NIH to develop and **maintain a database** that recognizes those repositories and platforms.

The policy states that “NIH may make Plans publicly available.” We believe that the **NIH should affirm its commitment to make available to the public the Plans** of funded research proposals and contracts. Public visibility of the Plans will be informative and educational, permit tracking, and encourage compliance. ClinicalTrials.gov should be used to disseminate the Plans for registered clinical trials, and the Plans should be posted prior to study initiation. Additional repositories can be used for other types of research, or the NIH can simply publish the Plan as an additional field linked to or hosted on the NIH RePORTER.

Data holders and data contributors should be encouraged to apply **data tags (i.e. metadata) that describe how the data can be used**—and applicable restrictions to its use—to reflect any contractual terms (e.g. licensing, copyright), informed consent parameters, and institutional, state, and federal policies. Metadata that describe the terms of use will help ensure the appropriate and compliant use of the data in the future. Further, NIH should invest in developing a universal language or library for such data tags and tools to render such metadata machine-readable.

The burden of managing and sharing data does not rest solely on the data contributor but equally on the data scientists and researchers who have access to the data. **Strict policies with enforcement provisions should be communicated to those who access the data**, and data use agreements employed as appropriate. Data tagging as described above will make compliance both easier for the user and auditable if necessary.

The data management and sharing plan should be an important and determinative part of any NIH proposal, and the **Plan should be reviewed and scored by the study section** (or contracting entity). The Plan should not be relegated to a “Just-In-Time” submission but should affect whether a proposal is prioritized for funding. Consideration of data

management, integrity, and stewardship (and, later, sharing) is an integral part of study design and quality.

We believe further that **no two-page limitation** should be imposed on the Plan. The prospective description of data management and sharing of data and metadata should be as long as necessary to describe all important details. To support its significance, the Plan should not “count” against the page limitations of the proposed science.

Finally, given that a principal goal of the NIH policy is to “serve the public,” we believe strongly that this is a time when the **NIH should require return of aggregate study results** to participants, at least for the results of clinical research, and in plain language understandable to an individual. Absent a cogent reason, these aggregate results should be available to the public. While there are many issues with return of individual results to a participant that require consideration and analysis, summary results of clinical trials and clinical research should be widely available and understandable—and may help to promote public engagement and public trust in the research and scientific enterprise.

Thank you again for the opportunity to comment on this important issue. We believe that the NIH is in a unique position to harness the power of data sharing for the public good, but only if it uses this opportunity to advance the culture of, and infrastructure to support, data sharing.

We are available to discuss our comments with you if that would be helpful and would be happy to work with you on any of the aforementioned items. Please feel free to contact the MRCT Center at bbierer@bwh.harvard.edu, sawwhite@bwh.harvard.edu, and mark.barnes@ropesgray.com.

Respectfully submitted,

Barbara E Bierer, MD
Sarah A White, MPH
Mark Barnes, JD, LLM

Data Management and Sharing Plan (Plan): The definition of the Plan in the proposed draft should be considered more broadly. Rather than just defining how the data will be “managed, preserved and shared,” it is important that the Plan also identify at least one target “other” who would use the data, and describe how that user would actually use it. Otherwise, the Plan will only describe how to make it possible, rather than to enable it.

The definition also does not address how the scientific data will be collected or described in the shared data set, which are important components of the Plan.

Data Management: No comment

Data Sharing: The proposed definition for data sharing is limited, and should more explicitly state that sharing is more than increasing access. Sharing data must involve enabling the access and reuse of data to facilitate and optimize research. In addition, because it is important to understand how results were determined, and considering that current data analytics provide unique methods of data analysis and interpretation, sharing the associated code that was used to determine the accuracy of analyses should be required.

Metadata: No comment

Scientific Data: The proposed definition of scientific data explicitly excludes “preliminary analyses” as eligible material. However, many data will be rendered futile for the purposes defined in data sharing without a clear understanding of preliminary analyses, and in some cases access to that data. At a minimum, the definition could state the scientific data “may exclude” preliminary analyses, rather than an unequivocal exclusion.

Section III: Scope

We believe an important opportunity inherent in this policy is to expand access to data extracted from electronic health records for secondary analyses, which represents a growing area of important research. It is possible that many studies using secondary data analysis will not explicitly de-identify the data if not legally required (e.g., if the data do not leave the institution). To mitigate this potential limitation, the scope should be broadened to recommend the creation of shareable data sets to support these analyses.

Section IV: Effective Date(s)

No comments.

Section V: Requirements

NIH guidance on the submission of the Plan should be as specific and prescriptive as possible. For example, the “Requirements” section notes that the submission of the Plan should outline how scientific data will be managed and shared, “taking into account any potential restrictions or limitations.” Instead of taking into account potential restrictions or limitations, the Plan should clearly describe what those restrictions and limitations are. For example, if the data contains proprietary information that imposes restrictions on sharing, this should be clearly described in the Plan prior to NIH funding decisions. The “Requirements” section should clearly state which elements of the Plan are required and which are optional.

We agree with the statement that “additional or specific information” may be requested to meet expectations for data management and sharing. NIH may consider elaborating on this statement and stating that the creation of shareable de-identified data sets may be requested (even if not required for observational studies). We encourage NIH to consider how to ensure sufficient resources (budget) exist to support that aspect of the Plan.

Section VI: Data Management and Sharing Plans

AcademyHealth members recommend several suggestions related to the Plan, which are organized around topic area below.

Data security and privacy. The draft policy suggests that investigators are responsible for ensuring data security and compliance with privacy protections throughout the life of the scientific data, even after it has been shared. This is a tremendous and daunting requirement. Further, with changing discoveries in security and privacy protections, this requirement may be beyond the capability of most researchers and institutions, and may limit sharing. One strategy to alleviate this burden would be the use of NIH repositories (see next subsection for additional thoughts and comments).

Use of repositories. The draft policy states that NIH “encourages the use of established repositories for preserving and sharing scientific data.” NIH should elaborate on the utility of repositories, why repositories are encouraged, and the criteria for establishing a repository. Given the statement that Plans should identify strategies or approaches to ensure data security and compliance with privacy protections through the life of the data, repository use could be incentivized (e.g. by making administrative supplements available, providing the investigator extra “points” for a track record of using repositories as part of subsequent grant reviews, etc...), not just encouraged. Use of established repositories or repositories within NIH would lift the burden of privacy and security protections from the researcher and be assumed by NIH, who can better represent the public need for the data.

Plan Elements. The proposed guidance suggests that investigators “consider” addressing specific Plan elements outlined in the supplemental guidance. We believe that requiring applicants to address all elements listed in the supplemental guidance would provide clarity to applicants on expectations for an adequate Plan as well as assist NIH with their review of the Plan.

Making plans public. The policy states that NIH may make Plans publicly available. Revising this statement to indicate that NIH *will* publish Plans for public consumption removes any uncertainty that this may or may not happen, and promotes transparency and accountability among the community of NIH researchers.

Peer review. As currently written, the proposed policy does not require that Plans be evaluated as part of the peer review process. This separation from Peer Review suggests that specific elements of the Plans are not subject to an acceptable level of scientific scrutiny, and therefore may not carry as much importance as the rest of the study plan. We believe that the Plan for making data accessible for evaluation and further research should be regarded with the same level of importance and integrity as the study plan, and note that doing so also requires the engagement of reviewers with appropriate knowledge and expertise to evaluate the Plan.

Section VII: Compliance and Enforcement

No comments.

Supplemental DRAFT Guidance: [Allowable Costs for Data Management and Sharing](#)

NIH's guidance on allowable costs for data management and sharing should include generating shareable datasets when legal privacy protections would otherwise restrict the sharing of data.

NIH should also collect information about the anticipated costs to researchers to access the study's scientific data in the intended repository. Data that will carry higher than average access costs become essentially inaccessible.

Supplemental DRAFT Guidance: [Elements of a NIH Data Management and Sharing Plan](#)

AcademyHealth members recommend additional suggestions related to elements of the Plan, which are organized around topic area below.

Introductory Language. The draft guidance states that NIH does not expect researchers to share all scientific data generated in a study. NIH should provide more clarity on what types of specific scientific data researchers are not expected to share. AcademyHealth feels strongly that scientific data created through NIH funding that will not be shared should include a justification for its exclusion.

Data Type. Descriptions of the modality, level of aggregation, and degree of data processing should be considered basic requirements of the Plan. These descriptions would form the minimally necessary metadata that will make the scientific data understandable to others. Without this information, other researchers would be burdened with producing this basic level of information, impeding and delaying further use and reuse, slowing the progress of science, and undermining the objective of data sharing.

Access to Data. With regard to access to restricted scientific data, the applicant should be required to explain their process for obtaining approval to the restricted data. NIH should be able to determine from the Plan how likely it is that others will also be able to access the restricted data.

Other Use Limitations. There should be specific consideration of limitations imposed by sharing protected health information and identifiable information. These are significant limitations and should be addressed directly, along with proposed strategies to mitigate these challenges.

Other Considerations Relevant to this DRAFT Policy Proposal

The draft policy alludes to the significance of the HIPAA Privacy Rule, but does not explicitly address the implications for data management and sharing. The HIPAA Privacy Rule identifies research as a public interest and benefit activity, and if such allowances are made, the Plan should include how to address protected health information. Otherwise, it may be easy to exclude research using protected health information from this policy.

An additional element and benefit of data sharing is the importance of open data to spark cross-disciplinary research. Members noted that the availability of valid data resources is a key strategy for bringing new investigators from other fields into health and health care research.

We believe the guidance should be prescriptive enough to provide funding panels clear insight into the likely success of the research project in terms of data sharing-- that is, the likelihood that the study's data could be used to validate or replicate study findings or be reused for further research. The NIH should make it clear to researchers that insufficient data management and sharing plans will affect the likelihood of receiving funding.

Finally, beyond and in addition to the whole of our comments above, we want to reemphasize the importance of providing centralized support – infrastructure, operational guidance, resources, logistical support and training – to the field in order to support meaningful implementation of this policy.

Conclusion

AcademyHealth appreciates the opportunity to provide additional input on the Draft Policy for Data Management and Sharing. We believe clarity and specificity will be paramount to a sound and effective policy – one that clearly communicates context, rationale, requirements, and expectations and advances progress toward the widespread and responsible sharing of data.

AcademyHealth consulted with a committee of members and thought leaders to offer a response to NIH's request for comments on their draft Policy for Data Management and Sharing. We thank and acknowledge Greg Downing for his valuable contributions and guidance as Chair of the Ad-Hoc Committee.

Detailed Recommendations and Comments to the Draft Policy

SECTION 1. Data Sharing Strategy Development

Section I: Purpose

RSNA concurs with the general purpose of the draft NIH Policy for Data Management and Sharing, particularly the intent to encourage data sharing in the health research community under consistent practices protecting the security and privacy of health information.

Section II: Definitions

RSNA concurs with the recommended additions to the definitions section of the draft Policy submitted by AMIA, i.e., the addition of concepts for “Data Management,” “Covered Data,” “Covered Timeframe,” and the refinement of definitions for “Metadata” and “Scientific Data.” We also strongly support AMIA's proposed addition of "Scientific Software Artifacts" as a defined term and further recommend that "imaging acquisition or post-processing methods and algorithms" be included as examples of such artifacts.

Section III: Scope

RSNA concurs that the Policy should apply to all research funded or conducted by NIH.

Section IV: Effective Date(s)

RSNA supports the recommendation submitted by AMIA that NIH consider a phased timeline for implementation of the requirements of the DMSP based on funding levels of research.

Section V: Requirements

RSNA concurs with the essential requirements for submission of a Data Management and Sharing Plan (DMSP) and compliance with the NIH ICO-approved Plan. RSNA also supports the recommendation submitted by AMIA that the DMSP should be a peer-reviewed and scored element of funding requests in order to foster and reward adherence to best practices, as well as innovation in data sharing practices.

Section VI: Data Management and Sharing Plans

RSNA generally concurs with the overview and outline of elements of DMSPs. We also strongly support AMIA's recommendation that, in addition to monitoring conformance to an approved DMSP, NIH should provide recognition to researchers, institutions and data repositories that share data according to FAIR principles and the NIH Data Science Strategic Plan. Recognized innovations should include making data sharing more cost efficient in medical imaging by implementing fluid data sharing models such as sharing managed links to image data, thus eliminating the need to copy vast stores of image data.

Section VII: Compliance and Enforcement

RSNA generally supports the proposed provisions regarding Compliance and Enforcement. We also concur with the recommendation submitted by AMIA that a process should be implemented for updating and versioning the DMSP over the course of a funded research program.

Submission ID: 1384

Date: 1/10/2020

Name: Salvatore La Rosa

Name of Organization: Children's Tumor Foundation

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All type of data equally

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Patient Advocate

Role - Other:

Domain of Research Most Important to You or Your Organization:

Neurofibromatosis

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

This section highlights well the benefits of sharing data and why NIH emphasizes the importance of good management practices. However, the section should sound more authoritative in requiring more than expecting all of the elements described in the policy in its entirety. This first paragraph has to state the expectation of this new policy clearly: "shared data 'must' be made accessible in a timely manner". Please avoid using the words 'should', 'encourage' or 'expect' and change with 'must' and 'require'.

The use of 'timely' in this section is probably ok, but it must be clearly defined in a specific section of the policy later. Options here are no later than the time of publication or, within the next 6 months after publication.

Section II: Definitions:

The data management and sharing plan definition is a good place to include details for how and when the resulting data will be shared, defining the 'timely' adjective used in the Purpose section above.

The Scientific data definition should make clear that NIH incentivize researchers to share data not only from data that is used in publications but also from projects that were not published or yielded negative or inconclusive results, as soon as the experimental details are valid and verifiable.

Scientific data should also include a description of protocols used in the data analysis, including proprietary software or any artifact used to manipulate or transform raw data.

Section III: Scope:

I appreciate the fact that NIH is broadening the scope of this policy to all research regardless of funding level of mechanism.

Section IV: Effective Date(s):

Ideally, the policy should state that implementation will be no later than 12 months after issuance of the final policy

Section V: Requirements:

This section should have an extra bullet point stating "NIH requires all researchers to share all scientific data generated in a study, with exceptions only when justified to a panel that included subject matter and data experts."

NIH should also specify that the repository for the data (and the software artifacts mentioned above) must comply with the FAIR principles.

Section VI: Data Management and Sharing Plans:

DMSPs need to be included in the regular submission of an application. Because effective DMSPs will increase the impact of the research, they need to be rewarded by increasing the overall score of an application. Reviewers receive guidance on how to score significance and approach and should also receive guidance on scoring effective DMSPs. The guidance should be based on how effectively the plan addresses the FAIR principles.

NIH (these can be ICO specific) can create and encourage the use of templates for DMSPs which can help to standardize the essential elements and layout of DMSPs. Researchers who do not use the standard template are not penalized as long as the essential elements are included, and the plan addresses the FAIR principles sufficiently.

This section states that "NIH may make Plans publicly available." The NIH must make plans publicly available. This will ensure transparency which can help to encourage compliance. The

NIH should publish DMSPs for funded awards (grants, contracts, fellowships, etc) with the abstracts in the NIH RePORTER. Knowing that these plans are available to the public will increase compliance among researchers. NIH can't police compliance 100%, though random audits would be valuable, along with the review of compliance during annual reports. Publishing plans in RePORTER also provides transparency and enables community scrutiny after the grant term has ended. It should also be made clear that all scientific data must be shared. Any exceptions to this must be justified and the funding conditioned on approval by an NIH advisory committee of data management experts.

A Board or Advisory council must also be responsible for oversight of DMSPs for intramural researchers. It is not prudent to give a single NIH official (such as the Scientific Director or Clinical Director) the ability to review and approve these plans.

Section VII: Compliance and Enforcement:

If the DMSPs are included as part of the application, there is a strong positive incentive for developing a robust plan for data management and sharing.

To help compliance during and after the award term, the DMSPs need to be included (in a machine-readable fashion) in NIH RePORTER. That record needs to also have contact information to request corrective action for violations of the Data Management and Sharing policy, or the published DMSP. The contact info needs to include PI's or project directors' email addresses, and contact information for the institutional official at the grantee institution. It is also important to include an NIH contact to whom queries should be sent. This gives enough information to handle the issue at the PI or grantee institution-level first but also enables escalation to the NIH when necessary. Similar information should be included for all mechanisms.

Researchers should be required to obtain and provide DOIs for data sets as an integral part of progress reporting. If these are not provided and no justification is provided than Remedies for Noncompliance should be put in place for the researchers and the recipient institution.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Guidance should specify that costs related to curating, developing supporting documentation and preserving data are allowed beyond the funding period (including personnel) and until the data will be made available. This point is essential and is very important in defining the time when the data has to be available (please see my comment to section 1 "Purpose").

It should also stipulate that only costs to deposit data into repositories that comply with the FAIR principles will be allowable.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The plan has to be submitted and scored at the time of the application submission. DMSP has to be considered for scoring an application and the lack of one should severely affect the quality of the application received. A DMSP is absolutely necessary at the time of submission and if there are elements that are still not clear or not determined within the plan they might be 'amended' later, but the overall plan and as much as possible of specific details need to be in place at the time of submission.

Without scoring this as part of the application the incentive to develop a robust plan is gone. In addition, to say that "if certain elements of a Plan have not been determined at the time of submission, an entry of "to be determined" may be acceptable if a justification is provided ..." completely defeats the purpose.

The sentence "NIH does not expect researchers to share all scientific data generated in a study." contradicts the whole philosophy of the policy and I strongly suggest to remove it.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1385

Date: 1/10/2020

Name: Houri K. Vorperian

Name of Organization: University of Wisconsin-Madison

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Anatomic-acoustic

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

speech communication

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The ending of the first paragraph (last sentence) is hard to read/understand. Linked with two 'and's .

It is not clear what is the grouping of the 'Digital Scientific Data from NIH ..' with FAIR ?

Section II: Definitions:

I think the definition of Data Sharing should explicitly state "through existing repositories". More specifically stating: "deposit data into existing repositories".

Providing some examples of repositories would be a good idea. An example from personal experience, I naively thought that I could deposit my data in a repository at the National Library of Medicine. However, I learned that they just host links to existing repositories i.e. they do not host the data. We identified multiple repositories some that charge annually, others that are restricted to types of data etc..... I finally re-contacted the Office of the Vice Chancellor and the Libraries at my institution (University of Wisconsin-Madison) demanding that they support their researchers and NIH supported generated data. They are now working on having a place for researchers to deposit federally funded data.

Section III: Scope:

It would not hurt to explicitly specify dollar amount (even if there are no restrictions). I was glad to see that the funds greater than 500K is no longer a requirement in the data sharing policy.

Section IV: Effective Date(s):

Date is a bit tricky when it comes to timing with funded period. Researchers would want to be in a position to use their data for publication prior to releasing it/sharing it. Yet, once the funding period is over - there is currently no mechanism to seek funds to aid in curating data.

Section V: Requirements:

If data sharing through repository is going to be a requirement, I suggest NIH make it also a requirement for Universities/Institutions to have 'established repositories' for federally funded data. A better option would be for NIH Institutes to have established repositories.

You had noted the need of examples, here is one from personal experience. Through NIH-NIDCD funds I have invaluable data (imaging and acoustic) that I want to make accessible to the scientific community. Repositories for interdisciplinary data are not available. I do not want to split my data into two separate repositories. If my Institution does not accommodate this need, I have nowhere to go. Also, I need to find a mechanism to fund the cost associated with sharing the data. It would be optimal if NIH provided an opportunity for supplemental funds for data sharing/curation. Data sharing efforts at the conclusion of a project are more likely to be thorough.

Section VI: Data Management and Sharing Plans:

Listing some examples of established repositories would be helpful.

Section VII: Compliance and Enforcement:

The section on Post Funding is most worrisome! Institutions should have a mechanism in place to support their researchers to meet compliance - the level of expertise needed for data curation is no joke and not a simple task as it requires a well thought out plan with input from knowledgeable experts.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Critical issues to address:

1. Long-term preservation and access should be defined. What is considered long-term? through perpetuity?
2. If a repository has an annual cost, what happens to the data when funding ends?

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

This seems a thorough section of the policy. Again, I think it would be helpful to give more examples particularly of standards that NIH approves.

Other Considerations Relevant to this DRAFT Policy Proposal:

I would like to highlight a few important consideration mentioned in my comments above:

- 'Depositing data into established repositories' is not common knowledge. Who hosts the data is critical to understand.
- How is long-term data sharing defined? and who absorbs the cost of maintaining the repository once funding period ends?
- Not all NIH institutes have established repositories. Who will support PIs with the meeting the expectations of this policy? If it is the role of the University, then NIH should make this expectation clear of them.
- Current set-up of repositories do not support multidisciplinary data. I think this is an important consideration.
- Establishing repositories with normative data is important have and trans-NIH efforts and guidance would be helpful for researchers/scientists to have.

Thank you for the opportunity to share some of my thoughts and concerns.

Attachment:

Description:

Submission ID: 1386

Date: 1/10/2020

Name: Sally Gore

Name of Organization: Lamar Soutter Library, University of Massachusetts Medical School

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Library Manager, Research & Scholarly Communications

Domain of Research Most Important to You or Your Organization:

Wide variety of biomedical

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We agree with the Policy's stated purposes and commend NIH on recognizing the value of research data towards advancing science and ensuring both validity and reproducibility. We do have questions and/or concerns regarding how the purposes will be assessed. This appears missing in the Policy at present and we believe it is vital in order to demonstrate the impact of public data sharing.

Section II: Definitions:

We believe the definitions provided are clear and correct.

Section III: Scope:

We believe the scope is clear and appropriate.

Section IV: Effective Date(s):

We feel that a 6-month time period between the time that the Policy is announced to when it goes into effect is fair. It should then apply to all grant submissions on or after this effective date. We view issues related to the release and implementation of the Policy through the historical lens of the NIH Public Access Mandate, meaning lessons learned from instituting the Mandate can well-inform the same for the Policy. A phased approach, while giving researchers and administrators time to understand the issues and create the necessary new steps within

their research process to share their data, also lacks the incentives necessary for adoption. As experienced in the Public Access Mandate, this type of approach resulted in confusion and delays in researchers' compliance. Expectations of researchers should be clearly stated and enforced from the beginning, rather than incremental steps along the way.

Section V: Requirements:

We believe the requirements are clear, though they still do not address concerns we expressed in the previous RFI. The Policy still lacks any mention of (1) how NIH will know if the data sharing requirements have been met and, related, (2) data citation. Will a system that generates something similar to a PMCID (PubMed Central Identifier) be created? Will an identifier, e.g. a digital object identifier, be assigned to data sets so that others can both locate the data and attribute the creators of it accordingly? We feel these remain significant gaps in the Policy.

Section VI: Data Management and Sharing Plans:

One of our greatest concerns in this section is in regard to the DMP being part of "Just-in-Time" requirements, rather than a section within the full grant proposal. As part of JIT, the plans will lack peer-review, give less time for researchers to seek out the support needed for writing a plan (particularly in the case of those developing one for the first time), and it creates a disconnect between budgeting for data management and preparing a budget for the initial application. There is also little to address the issue of qualifications for performing peer-review of a DMP. We feel there are a number of assumptions being made regarding researchers' knowledge and expertise in developing DMPs, as well as sharing and citing data. It is our experience, as academic medical librarians, that a knowledge gap indeed exists (much as it did when NIH introduce its Public Access Mandate for published literature), between an awareness of data management and the knowledge/skills needed to apply the principles in a manner to successfully meet the Policy's goals and purpose.

Section VII: Compliance and Enforcement:

Compliance and enforcement appear to be the weakest aspect of the Policy. The choice of language such as "NIH encourages..." and/or "researchers may..." fails to demonstrate the importance of the Policy, and in particular, its purpose. Similarly, language such as "deemed useful" is too ambiguous and inadvertently leaves a loophole in the Policy, allowing researchers to determine for themselves the value of their data. While we do not believe that the research community will, in mass, opt for such, however, as worded it provides an easy out for individuals lacking appreciation for the benefits of data sharing. This again undermines the purpose and value of the Policy, overall.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Our only concerns regarding allowable costs center on (1) knowledge gap among the research community regarding what might be needed, cost-wise, for DMPs and data sharing, and (2) separating this piece from the original budget, via putting it as JIT information.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We believe it would be of great value to create and provide a rubric for developing DMPs. This would be helpful in promoting best practices for managing and sharing data. We also suggest highlighting certain tools as part of the supplement; including providing metadata schema (e.g. DataCite), pointing to accepted standards for data sharing (e.g. fairsharing.org), and providing guidance, perhaps even links, to established data repositories. Again, we do not believe that the tools and/or processes around developing sound DMPs, nor the steps involved in making datasets both findable and accessible, are necessarily in the current purview of the majority of our biomedical and clinical researchers.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1387

Date: 1/10/2020

Name: Erin F. Hering

Name of Organization: Association for Research in Vision and Ophthalmology

Type of Data of Primary Interest:

Type of Data of Primary Interest - Other:

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: Manager, Science Communications

Domain of Research Most Important to You or Your Organization:

Eye and vision, ophthalmology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

a. Other funding agencies—including many agencies in countries other than the United States—do

not make data readily available to all researchers. Proposed changes to NIH's Data Management

and Sharing Policy could generate a disparate impact on/unintended consequences for our members in the US and non-US alike.

b. How will this policy be harmonized with existing NIH and international policies on patient privacy, intellectual property protections, etc.? Will the proposed policy supersede or be restricted by such existing policies?

c. Would this database also house data generated from non-NIH supported research? This is

especially important to consider when the data generated from large studies may be funded by non-NIH and NIH grants over different time periods. This may affect what aspects of the data can be shared on this proposed database.

Section IV: Effective Date(s):

Section V: Requirements:

a. There is no specific guidance in the draft policy as to what would constitute an acceptable vs. an

unacceptable "Data Management and Sharing Plan". For example, while the statement, "NIH does not expect researchers to share all scientific data generated in a study" will likely reduce the administrative burden for researchers, a classification of "required", "optional", "not required" is not included in the draft policy. Will NIH provide guidance on what data will be classified as "required", "optional" and "not required"?

b. What will the protocol be in cases in which the researcher does not own the data and is contractually obligated to not release data?

c. How will this policy accommodate fields that do not have a consensus or an agreed upon nomenclature on data formatting?

d. Many areas of biomedical research are lacking a consensus on the specifics of data sharing as compared to other fields where detailed standards have been developed, e.g. physics, geosciences, engineering, etc. How should data-which in our fields come in a variety of formats, complexities and sizes-be made available?

e. Will NIH generate one database for all data resulting from NIH-supported research? If not, what

is being envisioned how grantees go about setup and maintenance of such databases? Several issues consequently arise:

- o Who will assess data quality prior to and after deposition?
- o Who will curate these databases?
- o How far back in time can data still be uploaded, or will it be limitless if it meets the data standards?
- o How will the curation of these databases be adjusted as future demands on data availability (e.g., novel analytical methods) emerge?
- o How will databases be funded initially and long-term?

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

- a. How will this be funded initially and long-term?

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

☐ The CDC has requirements for a DMP, [https://www.cdc.gov/grants/additional-](https://www.cdc.gov/grants/additional-requirements/ar-)

25.html.

☐ All sharing policies should apply to NIH intramural and extramural research equally.

Attachment:

RFC NIH Data Management and Sharing Policy_1 10 2020.pdf

Description:

ARVO comments in response to NOT-OF-20-013



The Association for Research in Vision and Ophthalmology (ARVO) submitted comments in response to [NOT-OD-20-013](#), “[Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance](#)” on January 10, 2020.

[ARVO](#) is the largest and most respected eye and vision research organization in the world, with the mission to advance research worldwide into understanding the visual system and preventing, treating and curing its disorders. Our members include nearly 12,000 researchers from over 75 countries. These comments were initially drafted by ARVO’s Advocacy and Outreach Committee (AOC) and subsequently released to membership for feedback.

Section 1: Purpose

No comments at this time.

Section 2: Definitions

No comments at this time.

Section 3: Scope

- a. Other funding agencies—including many agencies in countries other than the United States—do not make data readily available to all researchers. Proposed changes to NIH’s Data Management and Sharing Policy could generate a disparate impact on/unintended consequences for our members in the US and non-US alike.
- b. How will this policy be harmonized with existing NIH and international policies on patient privacy, intellectual property protections, etc.? Will the proposed policy supersede or be restricted by such existing policies?
- c. Would this database also house data generated from non-NIH supported research? This is especially important to consider when the data generated from large studies may be funded by non-NIH and NIH grants over different time periods. This may affect what aspects of the data can be shared on this proposed database.

Section 4: Effective Dates

No comments at this time.

Section 5: Requirements

- a. There is no specific guidance in the draft policy as to what would constitute an acceptable vs. an unacceptable "Data Management and Sharing Plan". For example, while the statement, "NIH does not expect researchers to share all scientific data generated in a study" will likely reduce the administrative burden for researchers, a classification of "required", "optional", "not required" is not included in the draft policy. Will NIH provide guidance on what data will be classified as "required", "optional" and "not required"?
- b. What will the protocol be in cases in which the researcher does not own the data and is contractually obligated to not release data?



- c. How will this policy accommodate fields that do not have a consensus or an agreed upon nomenclature on data formatting?
- d. Many areas of biomedical research are lacking a consensus on the specifics of data sharing as compared to other fields where detailed standards have been developed, e.g. physics, geosciences, engineering, etc. How should data-which in our fields come in a variety of formats, complexities and sizes-be made available?
- e. Will NIH generate one database for all data resulting from NIH-supported research? If not, what is being envisioned how grantees go about setup and maintenance of such databases? Several issues consequently arise:
 - Who will assess data quality prior to and after deposition?
 - Who will curate these databases?
 - How far back in time can data still be uploaded, or will it be limitless if it meets the data standards?
 - How will the curation of these databases be adjusted as future demands on data availability (e.g., novel analytical methods) emerge?
 - How will databases be funded initially and long-term?

Section 6: Data Management and Sharing Plans

No comments at this time.

Section 7: Compliance and Enforcement

- a. How will this be funded initially and long-term?

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

No comments at this time.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan

No comments at this time.

Other Considerations Relevant to this DRAFT Policy Proposal

- The CDC has requirements for a DMP, <https://www.cdc.gov/grants/additional-requirements/ar-25.html>.
- All sharing policies should apply to NIH intramural and extramural research equally.

Submission ID: 1388

Date: 1/10/2020

Name: Briana Ezray and Cynthia Hudson-Vitale

Name of Organization: The Pennsylvania State University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of the above

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Institutional Officials survey of Scientific Researchers

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

PSU_Researcher_Response_10Jan2020.pdf

Description:

Thank you for the opportunity to comment on the NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance: Guidance for Allowable Costs for Data Management and Sharing and Elements of a NIH Data Management and Sharing Plan. On behalf of the Pennsylvania State University (PSU), we, the undersigned, present the following results of a survey of PSU researchers regarding their perspectives related to the NIH draft policy for data management and sharing.

To gather PSU NIH-funded researcher perspectives, a Qualtrics survey was distributed via email to researchers who either had NIH awards ending within 2019 or 2020 or had NIH awards granted in 2019 (n=188). The survey ran from December 16, 2019 to January 9, 2020, with three reminder emails sent throughout this period.

In total, 60 people responded to the survey (response rate=32%). Three respondents were not included in the aggregated results as they did not meet the criteria. Thus, there were 57 respondents with usable data for some or all of the remaining survey questions. Respondents self-reported the NIH sub-agency from which they received an award. Overall, the awards received by responding PSU researchers were provided by 16 different sub-agencies (Table 1).

Table 1: Self-reported NIH sub-agency from which respondents have awards

Sub-Agency	n
National Institute of General Medical Sciences (NIGMS)	8
National Institute on Aging (NIA)	6
Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)	6
National Heart, Lung, and Blood Institute (NHLBI)	5
National Institute of Allergy and Infectious Diseases (NIAID)	5
National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)	5
National Institute of Mental Health (NIMH)	5
National Institute of Neurological Disorders and Stroke (NINDS)	4
National Cancer Institute (NCI)	3
National Institute on Drug Abuse (NIDA)	3
National Eye Institute (NEI)	2
National Institute on Alcohol Abuse and Alcoholism (NIAAA)	2
National Institute of Biomedical Imaging and Bioengineering (NIBIB)	2
National Institute of Dental and Craniofacial Research (NIDCR)	2
National Institute of Environmental Health Science (NIEHS)	2
National Human Genome Research Institute (NHGRI)	1

Purpose

Overall survey results indicate that PSU researchers understand the importance of data sharing and the management of research data throughout the project lifecycle. Results indicate that 84% of the respondents found data sharing extremely important or very important (Figure 1).

Yet, there was a general concern that the changing policy could lead to increased administrative burden.

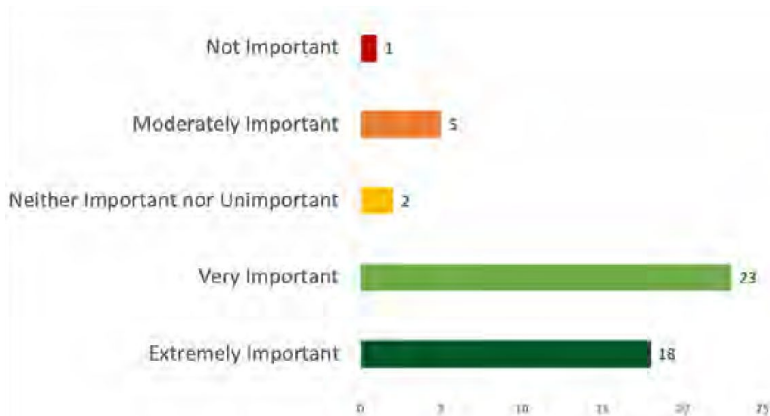


Figure 1: Counts of the responses of PSU researchers assessing how important they felt data sharing is.

Definitions

Numerous respondents suggested the NIH provide additional guidelines regarding which types of data are expected to be managed and shared. One respondent stated, *“We need sensible and readily interpretable guidelines that indicate which type of data should be shared and which are excluded from this mandate.”* In particular, PSU researchers indicated it would be helpful if the policy was revised to be more explicit and detail if data such as images or data resulting from simulations, various levels of analyses, etc., were covered under the scope of this policy. Investigators would also appreciate a clear definition of the types of data excluded from this policy for reasons other than legal or ethical factors. We recommend that the NIH create a matrix of data types and formats indicating expectations for sharing.

Finally, there was also the suggestion that the types of data included in the policy should be considered more broadly. For instance, while it was recognized the definition of data does not include laboratory notebooks, some researchers felt strongly that devising some means of disseminating information from wet-lab experiments could lead to important discoveries and aid future scientists.

Scope

Some PSU researchers suggested that they felt the NIH draft policy significantly increases researcher/research team administrative burden, at times to the detriment of their research. Other responses indicated that the policy requires more elements than is actually useful to researchers.

Requirements

Overall, the majority of PSU researchers 82% (42/51) felt comfortable with the requirement that all NIH proposals include a data management and sharing plan. Regarding the requirement that non-compliance with the NIH ICO-approved plan may be taken into account by the funding NIH ICO for future funding decisions for the recipient institution, 35% (18/51) were strongly or somewhat comfortable with this requirement, 35% (18/51) of respondents indicated they were somewhat or extremely uncomfortable with this requirement, and 29% (15/51) were neutral.

Additionally, a number of researchers indicated training and education would be a critical component of supporting the plan. A PSU researcher stated, *“This is going to require a great deal of education at the local level. Most of us: 1) strongly agree this is the way to go, while 2) having no background in doing this the right way.”*

Data Management and Sharing Plans

Overall, the majority of researchers somewhat or strongly agreed that they were comfortable (86%, 44/51) with the practice that Plans may be updated (with appropriate NIH ICO approval) during regular reporting intervals if changes are necessary or at the request of NIH ICO to reflect changes in the previously documented approach to data management and data sharing throughout the research project, as appropriate (Figure 2).

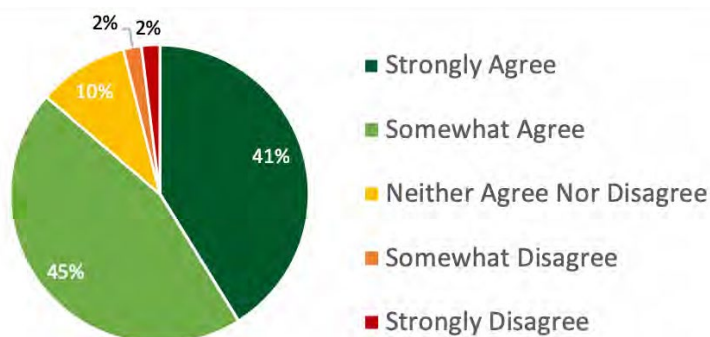


Figure 2: Percent of PSU researchers who agree/disagree that they are comfortable that Plans may be updated during regular reporting intervals.

Additionally, approximately 69% (35/51) of PSU researchers somewhat or strongly agreed that they were comfortable with NIH possibly making Plans publicly available (Figure 3).

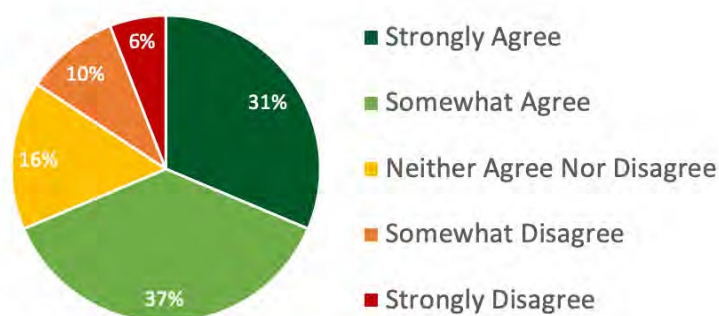


Figure 3: Percent of PSU Researchers who agree/disagree that they are comfortable that NIH may make Plans publicly available.

With regards to plan assessment, PSU researchers expressed a need for education and guidance on developing quality data management and sharing plans. More specifically, there was a desire for guidelines and examples of data management plans for data that cannot be shared because of privacy or human subject concerns.

Please see Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan section for review of researcher perspectives on the elements of a NIH Data Management and Sharing Plan.

Compliance and Enforcement

Overall, feedback from PSU researchers indicates the impact of compliance and enforcement mechanisms on faculty burden and administrative costs is of considerable concern.

For example, a PSU researcher stated, *“It definitely will add effort and take budget away from the project that would otherwise be used to run the project. With budgets as tight as they are this does create additional burden. However, the sharing of data is important despite the costs.”*

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

Overall, PSU researchers identified having funds to share data as the data management activity they anticipate being the most difficult. A large majority (78%, 39/50) of respondents found this task to be somewhat or extremely difficult (Figure 4).

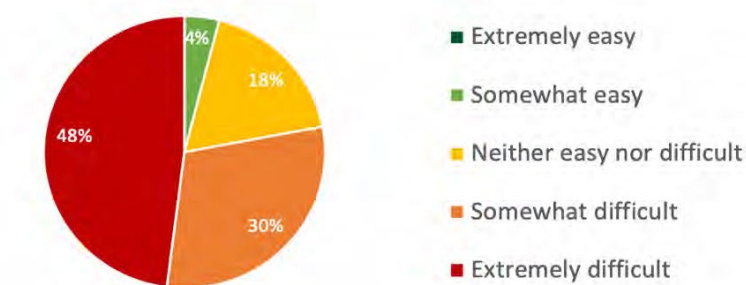


Figure 4: Percent of PSU Researchers who identified having funds to share data as easy/difficult.

Furthermore, 56% (29/52) of PSU researchers suggested they were likely to use some of their grant budget for (1) curating data and developing supporting documentation, (2) preserving and sharing data through established repositories, and (3) local data management considerations, whereas 35% (18/52) of PSU researchers suggested they were unlikely to use some of their grant budget for these purposes (Figure 5).

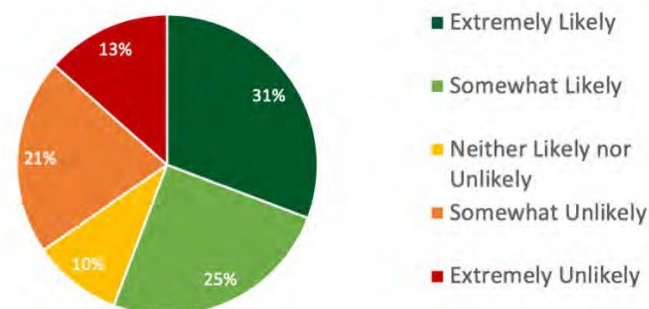


Figure 5: Percent of PSU Researchers who would likely include data management and sharing costs in their grant budgets.

PSU researchers who were unlikely to use the grant budget for data management costs pointed to the following reasons: (1) increased administrative burden (e.g., extra internal paperwork) to manage such an expense both at PSU and at the NIH, (2) data management and sharing will occur after the grant period has ended, and (3) the budgets are already limited given the scope of many research projects and there is no room in the budget for additional data management and sharing costs such as personnel time. PSU researchers did suggest that they would like to see the following expenses as allowable costs associated with data management and sharing: long term storage costs for large datasets, costs to cover staff tasked with data management tasks, IT support, database development, high-performance computing costs, data access fees, consulting fees, association fees, and systems to monitor how data is being used to prevent duplication or conflict.

Overall, PSU researchers felt that NIH needs to expand award budgets to allow for the inclusion of data management and sharing fees. As noted by an investigator, *"I generally struggle with*

keeping my budgets below existing thresholds and still ensure that the scope of the work is not undermined.”

Additionally, numerous PSU researchers felt that the policy needs to outline a means to have access to data management and sharing budgets post award. For instance, researchers stated:

“Some of the data management and sharing activities would need to take place and be paid for after the award period is over”

“Investigators conducting the usual R01 often require additional time to create the necessary documentation, develop a repository and a host of other activities as part of data sharing, and then address their specific aims. These tasks are often times not possible until toward the end of the award period and frequently continue on. NIH should consider additional funds or the development of systems that would support this work independent of awards, if they plan to require this level of ‘extra effort’.”

Another researcher asked whether there would be *“a post-award reimbursement once completed”*?

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan

Regarding the DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan, PSU researchers had minimal feedback on additional needed elements of the plan. Overall, fourteen respondents indicated that no additional elements were needed, and the current elements were sufficient. Additional suggested elements included: metadata accuracy and data maintenance.

When asked the expected level of difficulty for completing a variety of data management tasks, PSU researchers indicated they expected gathering and organizing supplementary files associated with the data (66%, 33/50) and preparing data for deposit in a repository (71%, 35/49) to be somewhat to extremely difficult. Survey results also indicated a number of data management tasks and activities were found to be neither easy nor difficult.

PSU researchers were strongly or somewhat uncomfortable (59%, 30/51) with the proposition that scientific data should be made available as soon as practicable, independent of award period and publication schedule. Overall, the majority of researchers suggested that data sharing in a “timely manner” constitutes following publication and/or submission for peer review (Figure 6). Researchers suggested that it is important that *“when the data has been prepared for publication that raw data should be made available along with all needed code to replicate published findings”* and that *“researchers must be given the opportunity to analyze and report on the data they collect before the data is shared.”*



Figure 6: PSU researcher responses to what they believe constitutes sharing data in a “timely manner.”

Finally, from this survey it is apparent that data sharing often coincides with the requirement to share data for journal article submissions and publications. It is vital that the NIH be more explicit about what data are shared. For instance, will the NIH be limiting the requirement to data shared in a publication? PSU researchers recommend that if the NIH is suggesting that the whole data package be shared, a longer duration be permitted to allow the researcher to fully mine the data before sharing.

Submitted on behalf of Penn State University:

Sarah Damaske, Penn State Population Research Institute

Esther Dell, Penn State College of Medicine

Briana Ezray, Penn State University Libraries

Wayne Figurelle, Penn State Institute for Computational and Data Sciences

Laura Heath, Penn State College of Medicine

Cynthia Hudson Vitale, Penn State University Libraries

Maurie Kelly, Penn State Data Commons

Greg Madden, Penn State Office of the Senior Vice President for Research and Office of the Vice President for Information Technology and Chief Information Officer

Leslie Parent, Penn State College of Medicine

Connie Rogers, Penn State Huck Institutes for the Life Sciences

Submission ID: 1389

Date: 1/10/2020

Name: Holly J. Falk-Krzesinski, PhD

Name of Organization: Elsevier

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of the above

Type of Organization: Other

Type of Organization - Other: Research Information Analytics and Publisher

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

All types of research data

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Sharing research data has the potential to make research more reproducible and efficient. Scientific research is a complex process and it is crucial that, at the different stages of this process, researchers handle data in a way that will allow sharing and reuse. Creating a good data ecosystem that supports each of these data needs requires collaborations between all parties that are involved in the generation, storage, retrieval and use of data: researchers, librarians, institutions, government offices, funders, data providers and publishers. Both in its publishing program and through its Research Data Management Services, Elsevier aims to develop tools, processes, and standards to support effective, rigorous, and open research data management practices.

Our team of developers, publishers and data thought leaders works closely with academic, government, and industry partners to develop a range of industry standards that enable institutions and their researchers to unlock the full potential of research data. At the heart of our RDM Services is Elsevier's data repository, Mendeley Data, a trusted research data repository with CoreTrustSeal certification (see <https://www.coretrustseal.org/>), which is free to researchers. Both through the Mendeley Data team and through our publishing, research collaboration and strategic alliances groups, Elsevier representatives are involved in numerous cross-sector community initiatives committed to advancing open science through research data

sharing. In collaboration with the Open Science Foundation, among others, Elsevier has developed journal data guidelines that align with the Transparency and Openness Promotion (TOP) Data Standards and has implemented these across the vast majority of our 2,000-journal portfolio, integrating them within our submission system to ensure authors can easily share and/or link to their data (refer to <https://www.elsevier.com/connect/elsevier-supports-top-guidelines-in-ongoing-efforts-to-ensure-research-quality-and-transparency> for additional information). As co-leads of the Publisher's Action Team, Elsevier has helped draft and implement the American Geological Union (AGU)-lead effort 'Enabling FAIR Data' and drafted some of the most forward-looking guidelines to support transparency and reproducibility across the Earth, Space and Environmental Sciences (refer to <http://www.copdess.org/enabling-fair-data-project/>). Other efforts include:

- Force11: Co-founder; co-authors FAIR Data principles; leading implementation data citations principles for publishers
- ICSU: Active Member
- ORCID: Co-founder
- Pistoia Alliance: Active member
- Scholix: Co-founder
- Research Data Alliance: Active member and co-chair for a number of working and interest groups
- Research Elements: Market leader in data journals (in English)
- STM: Supporting Brussels open data declaration

It's through Elsevier's experience and community engagement that we present our response to the Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance (NOT-OD-20-013) below. In addition to this current Request for Public Comments response, Elsevier has previously submitted responses to all NIH research data-related Requests for Information (RFI) over the last five years:

- NOT-OD-19-014, Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research
- NOT-OD-17-015, Strategies for NIH Data Management, Sharing, and Citation
- NOT-OD-16-133, Metrics to Assess Value of Biomedical Digital Repositories
- NOT-ES-15-011, Input on Sustaining Biomedical Data Repositories

- NOT-AI-15-045, Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services
- NOT-OD-15-067, Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the National Library of Medicine (NLM), Comment 5

Elsevier supports the NIH's effort to update its policy for data management and sharing toward making research data more effective. Data sharing enables researchers to reuse the results of experiments and supports the creation of new science that is built upon previous findings, making the research process more efficient. Data sharing also supports transparency and reproducibility, building trust in science. Applying FAIR data principles within the policy supports researchers to store, share, discover, and reuse research data.

Section II: Definitions:

The definitions for Data Management and Sharing Plan (Plan); Data Management; Data Sharing; and Metadata are sufficient and clear, no recommended changes.

The proposed definition for Scientific Data is consistent with those from the OSTP Public Access Memo

(https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf) and OMB circular A-110 (<https://georgewbush-whitehouse.archives.gov/omb/circulars/a110/a110.html>) and it is sufficiently flexible to allow for discipline-specific data standards setting. Having the benefit of being able to build on earlier definitions, we propose a slightly amended definition for greater clarity to the research community:

"Scientific Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, preprints, accepted manuscripts, final published journal articles, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens."

The addition of "preprints," "accepted manuscripts," and "final published journal articles," to the proposed revised definition provides further explicit clarity that research data are distinct from the text and visualizations within preprints, accepted manuscripts, or final published articles

(which also include affiliated supplementary materials), an important distinction between research data versus creative arrangements, interpretations, or presentations of research data.

We recommend adding a definition for "Research Data Repository" to signal to the research community that sharing, storage, and long-term preservation of research data necessitates the use of purpose-built infrastructure and research data repositories are referred to elsewhere within the policy:

Research Data Repository: A digital platform where research data is stored for the purposes of publishing, sharing, re-use, linking, and preservation.

Section III: Scope:

For the sake of clarity, we ask that a minor adjustment to second sentence be made:

"This includes research funded or conducted by extramural grants, contracts, intramural research projects, or other NIH funding agreements regardless of funding level or funding mechanism."

Section IV: Effective Date(s):

We request a similar minor adjustment, for clarification, to the last bullet point:

"Other NIH funding agreements (e.g., Other Transactions) that are executed on or after a future date (date yet to be determined), unless otherwise stipulated by NIH."

Section V: Requirements:

We commend the NIH for stating explicitly that, "Costs associated with data management and data sharing may be allowable under the budget for the proposed project," as that effectively signals to researchers that there may be costs associated with sharing research data in compliance with this policy and that they should consider costs during the project planning stages.

Section VI: Data Management and Sharing Plans:

Given that some projects are data-intensive, and some are complex research programs and/or multi-institutional proposals, the two-page limit may be too constraining and not allow for researchers to provide sufficient detail necessary for review by peer reviewers and program staff. We recommend raising the Plan page limit to five pages.

In this section of the policy, it is noted that, "NIH encourages the use of established repositories for preserving and sharing scientific data." We recommend that the policy further include basic guidance to researchers on criteria that constitute an "established" or trustworthy research data repository, and require Plans include a description of the research data repositories researchers will use to deposit and share data. Some very useful resources about trustworthy repositories include:

- CoreTrustSeal (<https://www.coretrustseal.org/>): Offers certification based on the Core Trustworthy Data Repositories Requirements catalogue and procedures. This universal catalogue of requirements reflects the core characteristics of trustworthy data repositories and is the culmination of a cooperative effort under the umbrella of the Research Data Alliance (RDA).
- Repository Finder (<https://repositoryfinder.datacite.org/about>; <https://www.re3data.org/>): A project of the Enabling FAIR Data Project in partnership with DataCite that queries the re3data registry of research data repositories.
- Scientific Data Recommended Data Repositories (<https://www.nature.com/sdata/policies/repositories#general>): The journal Scientific Data has compiled a comprehensive list of trusted discipline-specific, community-recognized, and generalist research data repositories.
- Recommended versus Certified Repositories (<http://doi.org/10.5334/dsj-2017-042>): A research article that examines both recommended and certified repository characteristics. Husen, S.E., de Wilde, Z.G., de Waard, A. and Cousijn, H., 2017. Recommended versus Certified Repositories: Mind the Gap. *Data Science Journal*, 16, p.42.

Section VII: Compliance and Enforcement:

Elsevier is committed to continuing to evolve and enhance our Research Data Management Services to support the implementation of and compliance with Plans put forward by NIH-supported researchers.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We commend the NIH for explicitly acknowledging within this policy that making data accessible and reusable for others may require costs above and beyond the routine costs of conducting research, and for proposing allowances for reasonable costs to be included in NIH budget requests when associated with: curating data and developing supporting documentation; preserving and sharing data through established repositories; and local data management considerations, such as unique and specialized information infrastructure. An indication of the approximate amount (in dollars or percentages) of the overall project budget that can be allocated toward these costs would be extremely useful for researchers. An example of explicit funding allocation for data curation is given in a recent NSF 'Dear Colleague'

letter on Effective Practices for Data (see <https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp>).

As noted above, we recommend the policy mandate Plans include a Cost section, when applicable, and that the costs included in this section are accounted for within the overall proposed project budget.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The list of elements is minimal. OSP may also wish to review the Elements of a Data Management Plan (https://www.lib.ncsu.edu/data-management/how_to_dmp) at North Carolina State University, an abbreviated compilation of data management plan elements from several sources. There are additional elements in that compilation that could be considered for inclusion in policy: Roles and Responsibilities; Data Formats and Metadata; and Costs.

OSP might consider working with the California Digital Library to add an NIH Plan template to the DMPTool (<https://dmptool.org/>), which is used by many research universities and institutions to prepare consistent, quality data management plans.

While it is premature to require researchers to develop machine-readable data management plans (machine-readable DMPs focus on assigning identifiers and machine-actionable components of a plan), it might be worthwhile for the policy to include language encouraging researchers to develop them when possible. These can be integrated with the funding application and tracked during the period of funding, enabling a better way to track compliance throughout the project.

For additional information on machine-readable tools and standards, we recommend the following resources, as well as others available from the Research Data Alliance (RDA) web site:

- NSF ‘Dear Colleague’ letter on Effective Practices for Data, <https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp>
- Miksa, T., Rauber, A., Ganguly, R., & Budroni, P. (2017). Information Integration for Machine Actionable Data Management Plans. *International Journal of Digital Curation*, 12(1), 22. <https://doi.org/10.2218/ijdc.v12i1.529>
- Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. *PLoS Comput Biol* 15(3): e1006750. <https://doi.org/10.1371/journal.pcbi.1006750>

- Research Data Alliance (RDA). (2017). DMP Common Standards WG | RDA. Retrieved from <https://www.rd-alliance.org/groups/dmp-common-standards-wg>

Other Considerations Relevant to this DRAFT Policy Proposal:

We recommend that OSP review and update this policy on a more regular cycle, perhaps every 3-5 years, continuing to seek input from the community broadly. A shorter review/revise cycle will allow OSP to be nimble and keep the policy up to date with advances in both technical and technology capabilities. A shorter cycle will also allow for timely revisions should unforeseen negative consequences result, or if previously unconsidered limitations are brought to light. Moreover, since the NLM has recently commissioned the National Academies of Sciences, Engineering, and Medicine (NASEM) to conduct a study on forecasting the long-term costs for preserving, archiving, and promoting access to biomedical data, it will be important to review this policy in consideration of the findings from that study.

We strongly encourage OSP to set a schedule for collecting data about research data sharing practices, evaluating the impact of sharing research data on both research and researchers, and work with RDA and other community partners to develop and establish research data sharing metrics—sharing the findings with the community. These efforts underpin an evidence-based approach to science policy consistent with the ‘science of science policy’ (see https://en.wikipedia.org/wiki/Science_of_science_policy and https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505610&org=SMA&from=home) and will provide data to inform future policy changes and revisions.

Attachment:

ELS Response.pdf

Description:

Elsevier Response_Nov 2019

Elsevier's Response to [NOT-OD-20-013](#)
Request for Public Comments on a
DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Name:

Holly J. Falk-Krzesinski, PhD

Name of Organization:

Elsevier

Type of Data of Primary Interest:

Other

Type of Data of Primary Interest - Other:

All of the above

Type of Organization:

Other

Type of Organization - Other:

Research Information Analytics and Publisher

Role:

Institutional Official

Research Area Most Important to You or Your Organization (e.g., clinical, genomics, neuroscience, infectious disease, epidemiology)

All types of research data

Introduction

Sharing research data has the potential to make research more reproducible and efficient. Scientific research is a complex process and it is crucial that, at the different stages of this process, researchers handle data in a way that will allow sharing and reuse. Creating a good data ecosystem that supports each of these data needs requires collaborations between all parties that are involved in the generation, storage, retrieval and use of data: researchers, librarians, institutions, government offices, funders, data providers and publishers. Both in its publishing program and through its Research Data Management Services, Elsevier aims to develop tools, processes, and standards to support effective, rigorous, and open research data management practices.

Our team of developers, publishers and data thought leaders works closely with academic, government, and industry partners to develop a range of industry standards that enable institutions and their researchers to unlock the full potential of research data. At the heart of our RDM Services is Elsevier's data repository, Mendeley Data, a trusted research data repository with CoreTrustSeal certification (see <https://www.coretrustseal.org/>), which is *free* to researchers. Both through the Mendeley Data team and through our publishing, research collaboration and strategic alliances groups, Elsevier representatives are involved in numerous cross-sector community initiatives committed to advancing open science through research data sharing. In collaboration with the Open Science Foundation, among others, Elsevier has developed journal data guidelines that align with the Transparency and Openness

Elsevier's Response to [NOT-OD-20-013](#)

Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Promotion (TOP) Data Standards and has implemented these across the vast majority of our 2,000-journal portfolio, integrating them within our submission system to ensure authors can easily share and/or link to their data (refer to <https://www.elsevier.com/connect/elsevier-supports-top-guidelines-in-ongoing-efforts-to-ensure-research-quality-and-transparency> for additional information). As co-leads of the Publisher's Action Team, Elsevier has helped draft and implement the American Geological Union (AGU)-lead effort 'Enabling FAIR Data' and drafted some of the most forward-looking guidelines to support transparency and reproducibility across the Earth, Space and Environmental Sciences (refer to <http://www.copdess.org/enabling-fair-data-project/>). Other efforts include:

- Force11: Co-founder; co-authors FAIR Data principles; leading implementation data citations principles for publishers
- ICSU: Active Member
- ORCID: Co-founder
- Pistoia Alliance: Active member
- Scholix: Co-founder
- Research Data Alliance: Active member and co-chair for a number of working and interest groups
- Research Elements: Market leader in data journals (in English)
- STM: Supporting Brussels open data declaration

It's through Elsevier's experience and community engagement that we present our response to the Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance (NOT-OD-20-013) below. In addition to this current Request for Public Comments response, Elsevier has previously submitted responses to all NIH research data-related Requests for Information (RFI) over the last five years:

- NOT-OD-19-014, Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research
- NOT-OD-17-015, Strategies for NIH Data Management, Sharing, and Citation
- NOT-OD-16-133, Metrics to Assess Value of Biomedical Digital Repositories
- NOT-ES-15-011, Input on Sustaining Biomedical Data Repositories
- NOT-AI-15-045, Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services
- NOT-OD-15-067, Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the National Library of Medicine (NLM), Comment 5

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose

Elsevier supports the NIH's effort to update its policy for data management and sharing toward making research data more effective. Data sharing enables researchers to reuse the results of experiments and supports the creation of new science that is built upon previous findings, making the research process more efficient. Data sharing also supports transparency and reproducibility, building trust in science. Applying FAIR data principles within the policy supports researchers to store, share, discover, and reuse

Elsevier's Response to [NOT-OD-20-013](#)

Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

research data.

Section II: Definitions

The definitions for Data Management and Sharing Plan (Plan); Data Management; Data Sharing; and Metadata are sufficient and clear, no recommended changes.

The proposed definition for Scientific Data is consistent with those from the [OSTP Public Access Memo \(https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf\)](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf) and [OMB circular A-110 \(https://georgewbush-whitehouse.archives.gov/omb/circulars/a110/a110.html\)](https://georgewbush-whitehouse.archives.gov/omb/circulars/a110/a110.html) and it is sufficiently flexible to allow for discipline-specific data standards setting. Having the benefit of being able to build on earlier definitions, we propose a slightly amended definition for greater clarity to the research community (additions noted in **bold** below):

“Scientific Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, **preprints, accepted manuscripts, final published journal articles**, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.”

The addition of “preprints,” “accepted manuscripts,” and “final published journal articles,” to the proposed revised definition provides further explicit clarity that research data are distinct from the text and visualizations within preprints, accepted manuscripts, or final published articles (which also include affiliated supplementary materials), an important distinction between research data versus creative arrangements, interpretations, or presentations of research data.

We recommend adding a definition for “Research Data Repository” to signal to the research community that sharing, storage, and long-term preservation of research data necessitates the use of purpose-built infrastructure and research data repositories are referred to elsewhere within the policy:

Research Data Repository: A digital platform where research data is stored for the purposes of publishing, sharing, re-use, linking, and preservation.

Section III: Scope

For the sake of clarity, we ask that a minor adjustment to second sentence be made, as noted in **bold** and strikethrough below:

“This includes research funded or conducted by extramural grants, contracts, intramural research projects, or other ~~NIH~~ **NIH** funding agreements regardless of ~~NIH~~ funding level or funding mechanism.”

Section IV: Effective Date(s)

We request a similar minor adjustment, for clarification, to the last bullet point, as noted in **bold** below:

“Other **NIH** funding agreements (e.g., Other Transactions) that are executed on or after a future date (date yet to be determined), unless otherwise stipulated by NIH.”

Elsevier's Response to [NOT-OD-20-013](#)
Request for Public Comments on a
DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Section V: Requirements

We commend the NIH for stating explicitly that, “Costs associated with data management and data sharing may be allowable under the budget for the proposed project,” as that effectively signals to researchers that there may be costs associated with sharing research data in compliance with this policy and that they should consider costs during the project planning stages.

Section VI: Data Management and Sharing Plans

Given that some projects are data-intensive, and some are complex research programs and/or multi-institutional proposals, the two-page limit may be too constraining and not allow for researchers to provide sufficient detail necessary for review by peer reviewers and program staff. We recommend raising the Plan page limit to five pages.

In this section of the policy, it is noted that, “NIH encourages the use of established repositories for preserving and sharing scientific data.” We recommend that the policy further include basic guidance to researchers on criteria that constitute an “established” or trustworthy research data repository, and require Plans include a description of the research data repositories researchers will use to deposit and share data. Some very useful resources about trustworthy repositories include:

- CoreTrustSeal (<https://www.coretrustseal.org/>): Offers certification based on the Core Trustworthy Data Repositories Requirements catalogue and procedures. This universal catalogue of requirements reflects the core characteristics of trustworthy data repositories and is the culmination of a cooperative effort under the umbrella of the Research Data Alliance (RDA).
- Repository Finder (<https://repositoryfinder.datacite.org/about>; <https://www.re3data.org/>): A project of the Enabling FAIR Data Project in partnership with DataCite that queries the re3data registry of research data repositories.
- Scientific Data Recommended Data Repositories (<https://www.nature.com/sdata/policies/repositories#general>): The journal *Scientific Data* has compiled a comprehensive list of trusted discipline-specific, community-recognized, and generalist research data repositories.
- Recommended versus Certified Repositories (<http://doi.org/10.5334/dsj-2017-042>): A research article that examines both recommended and certified repository characteristics. Husen, S.E., de Wilde, Z.G., de Waard, A. and Cousijn, H., 2017. Recommended versus Certified Repositories: Mind the Gap. *Data Science Journal*, 16, p.42.

Section VII: Compliance and Enforcement

Elsevier is committed to continuing to evolve and enhance our Research Data Management Services to support the implementation of and compliance with Plans put forward by NIH-supported researchers.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

We commend the NIH for explicitly acknowledging within this policy that making data accessible and reusable for others may require costs above and beyond the routine costs of conducting research, and for proposing allowances for reasonable costs to be included in NIH budget requests when associated with: curating data and developing supporting documentation; preserving and sharing data through established

Elsevier's Response to [NOT-OD-20-013](#)

Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

repositories; and local data management considerations, such as unique and specialized information infrastructure. An indication of the approximate amount (in dollars or percentages) of the overall project budget that can be allocated toward these costs would be extremely useful for researchers. An example of explicit funding allocation for data curation is given in a recent NSF 'Dear Colleague' letter on Effective Practices for Data (see <https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp>).

As noted above, we recommend the policy mandate Plans include a Cost section, when applicable, and that the costs included in this section are accounted for within the overall proposed project budget.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan

The list of elements is minimal. OSP may also wish to review the Elements of a Data Management Plan (https://www.lib.ncsu.edu/data-management/how_to_dmp) at North Carolina State University, an abbreviated compilation of data management plan elements from several sources. There are additional elements in that compilation that could be considered for inclusion in policy: Roles and Responsibilities; Data Formats and Metadata; and Costs.

OSP might consider working with the California Digital Library to add an NIH Plan template to the DMPTool (<https://dmptool.org/>), which is used by many research universities and institutions to prepare consistent, quality data management plans.

While it is premature to require researchers to develop machine-readable data management plans (machine-readable DMPs focus on assigning identifiers and machine-actionable components of a plan), it might be worthwhile for the policy to include language encouraging researchers to develop them when possible. These can be integrated with the funding application and tracked during the period of funding, enabling a better way to track compliance throughout the project. For additional information on machine-readable tools and standards, we recommend the following resources, as well as others available from the Research Data Alliance (RDA) web site:

- NSF 'Dear Colleague' letter on Effective Practices for Data, <https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp>
- Miksa, T., Rauber, A., Ganguly, R., & Budroni, P. (2017). Information Integration for Machine Actionable Data Management Plans. *International Journal of Digital Curation*, 12(1), 22. <https://doi.org/10.2218/ijdc.v12i1.529>
- Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. *PLoS Comput Biol* 15(3): e1006750. <https://doi.org/10.1371/journal.pcbi.1006750>
- Research Data Alliance (RDA). (2017). DMP Common Standards WG | RDA. Retrieved from <https://www.rd-alliance.org/groups/dmp-common-standards-wg>

Other Considerations Relevant to this DRAFT Policy Proposal

We recommend that OSP review and update this policy on a more regular cycle, perhaps every 3-5 years, continuing to seek input from the community broadly. A shorter review/revise cycle will allow

Elsevier's Response to [NOT-OD-20-013](#)Request for Public Comments on a
DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

OSP to be nimble and keep the policy up to date with advances in both technical and technology capabilities. A shorter cycle will also allow for timely revisions should unforeseen negative consequences result, or if previously unconsidered limitations are brought to light. Moreover, since the NLM has recently commissioned the National Academies of Sciences, Engineering, and Medicine (NASEM) to conduct a study on forecasting the long-term costs for preserving, archiving, and promoting access to biomedical data, it will be important to review this policy in consideration of the findings from that study.

We strongly encourage OSP to set a schedule for collecting data about research data sharing practices, evaluating the impact of sharing research data on both research and researchers, and work with RDA and other community partners to develop and establish research data sharing metrics—sharing the findings with the community. These efforts underpin an evidence-based approach to science policy consistent with the ‘science of science policy’ (see https://en.wikipedia.org/wiki/Science_of_science_policy and https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505610&org=SMA&from=home) and will provide data to inform future policy changes and revisions.

Submission ID: 1390

Date: 1/10/2020

Name: Mary Jo Hoeksema

Name of Organization: The Population Association of America/Association of Population Centers

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: social science, biomedical, and longitudinal large-scale datasets

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: Government Affairs

Domain of Research Most Important to You or Your Organization:

population research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

PAA APC comments on NIH data mgt and sharing policy 1-20.docx

Description:

Population Association of America Association of Population Centers

Office of Government and Public Affairs

8630 Fenton Street, Suite 722 • Silver Spring, MD 20910
www.populationassociation.org • www.popcenters.org • 301-565-6710 x 1006



Population Association of America President

Dr. John Casterline
Ohio State University

Vice President

Dr. Noreen Goldman
Princeton University

President-Elect

Dr. Eileen Crimmins
University of Southern California

Vice President-Elect

Dr. Sara Curran
University of Washington

Secretary-Treasurer

Dr. Bridget Gorman
Rice University

Past President

Dr. Wendy Manning
Bowling Green State University

Dr. David Bloom

Harvard University

Dr. Jennifer Dowd

King's College, London, UK

Dr. Pamela Herd

Harvard University

Dr. Emily Hannum

University of Pennsylvania

Dr. Jeffrey Morenoff

University of Michigan

Dr. Jenna Nobles

University of Wisconsin, Madison

Dr. Mary Beth Ofstedal

University of Michigan

Dr. Krista Perreira

University of North Carolina

Dr. Zhenchao Qian

Brown University

Dr. James Raymo

University of Wisconsin, Madison

Dr. Jenny Trinitapoli

University of Chicago

Dr. Kathryn M. Yount

Emory University

Association of Population Centers

President

Dr. Kathleen Cagney
University of Chicago

Vice President

Dr. Debra Umberson
University of Texas at Austin

Treasurer

Dr. Andrew Foster
Brown University

Secretary

Dr. Jeffrey Morenoff
University of Michigan

January 10, 2020

The below comments are submitted on behalf of the over 3,000 members of the Population Association of America (PAA) (www.populationassociation.org) and the over 40 federally supported population research centers at U.S. based research institutions comprising the Association of Population Centers (APC) in response to 84FR60398, "Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance."

Our organizations are gratified to see that several of the recommendations that we raised in our [December 7, 2018 letter](#) have been incorporated into the draft policy. To recap that letter, we recommended that the policy: (1) articulate a system for sharing and archiving data extracts; (2) address the management of paradata (i.e., data about the data collection process); (3) reward data collection and sharing by incentivizing data citations; (4) address the costs of data sharing; and, (5) specify a timeline for data sharing.

The draft policy goes a long way to addressing key issues relevant for NIH-funded population scientists and demographers. For example, our organizations were pleased with the way the latest policy articulates a system for sharing and archiving data extracts in such a way that offers flexibility across fields and gives NIH Institutes discretion on how to implement it. As the draft rightfully points out, various fields have developed data and metadata standards and that, where possible, these should be used. It also allows fields that have not developed such standards to name standards, which will likely spread best practices to fields currently without standards.

Our organizations were also happy to see the emphasis on the management of data, including the management and sharing of metadata. In general, the current draft is consistent with the FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles.

We also support the draft plan and supplemental draft guidance to use established repositories when possible. These repositories are important institutional commitments to ensuring that data remain accessible even as the technology used for data storage and data finding changes. The investments that Institutes, such as NIA, NICHD and NIDA, have made in such repositories through the Inter-university Consortium for Political and Social Research (ICPSR) benefits the population sciences as well as many other disciplines. NIH grantees should be encouraged to use these repositories because they provide

avenues for data sharing in perpetuity and can provide secure access to restricted data, which is otherwise a major challenge.

We also were pleased that the dissemination of restricted data is mentioned explicitly, along with the need to safeguard respondents' identities and sensitive information. We encourage NIH to give greater emphasis to this matter in its final policy and state that the dissemination of restricted data, with appropriate protections of the privacy and confidentiality of respondents, should be standard practice, rather than an elective option.

We appreciate that the draft plan largely meets the needs of the research communities we represent and addresses the central challenge that resources and correct incentives are necessary to ensure good data stewardship and extensive data sharing occurs. However, we believe the policy could be expanded to include the following recommendations, which we included in our [December 2018 letter](#).

- In our earlier comments, the PAA and APC recommended rewarding data collection and sharing by incentivizing researchers to use citations to acknowledge the work that was done to curate and share data files. We continue to believe that the final policy should provide clear citation guidance, including recommendations for how to cite secondary data that are created and shared with the research community. Both primary and secondary data that are eligible for citation should receive an NIH data catalog record analogous to a PMID or PMCID (in addition to be cataloged using DOIs or other persistent identifiers). Such a mechanism aligns the incentives of academic rewards for principal investigators with the scientific community's data sharing needs.
- We also recommended that the final plan address the cost of data sharing. We recognize that the supplemental draft guidance on allowable costs for data management and sharing stipulates some expenses associated with data management and sharing could be budgeted in grant applications. We do not feel this goes far enough. Simply allowing these expenses does not ensure that researchers will budget sufficient resources for data sharing, especially when faced with funding caps that make it challenging to fully fund data collection and analysis. We continue to believe that funding for data management and data sharing needs to be provided outside the regular budget process, by separate supplements to cover data sharing costs and/or separate data sharing and archiving grants similar to a successful NICHD R03 program ([PAR-16-149](#)). An alternative would be to allow each NIH grant to have a separate budget for data management and sharing beyond any existing budget caps.
- The draft guidance is vague about what the expectations are regarding completeness and comprehensiveness of the data to be shared. There are

suggestions that investigators don't need to share every piece of data they collect and can make subjective decisions about what to include and exclude. There is a worry that investigators could withhold important data, either deliberately or inadvertently, and thus meet the letter but not the spirit of the policy. We suggest that NIH include language in its policy stipulating that all of the data and measures collected under its grants be shared with the broader research community, subject to standard restrictions related to protection of respondents' privacy and the confidentiality of the information they have provided.

Thank you for considering these comments. For more information, please contact Ms. Mary Jo Hoeksema, Director, PAA/APC Government and Public Affairs, at maryjo@popassoc.org.

Submission ID: 1391

Date: 1/10/2020

Name: Gerald J. Perry, Assoc Dean University Libraries/Lori Schultz, Sr Dir Research, Innovation & Impact

Name of Organization: University of Arizona

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All listed data types are relevant to the institution

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Representatives of University Libraries, and of the Sr. Vice President for Research

Domain of Research Most Important to You or Your Organization:

All domains in biomedical research and areas funded by the NIH

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We appreciate any guidance around this policy and how it interacts with other current and future NIH/Institute & Center policies for data sharing, resource sharing, and management. The December 16th webinar indicated the NIH's desire to have a consistent set of expectations around the various sharing policies, which we support. In particular, does this policy sit above other data sharing/management policies as a way to provide general guidance, while policies like the Genomic Data Sharing Policy provide specifics? Will this policy provide overarching guidance, or is it intended to be separate?

The instructions indicate that this will be a trans-NIH data management policy. The policy should be reviewed for congruency with other, more-specific data policies, and include features of other policies that create understanding, such as (specifically from the Genomic Data Sharing Policy):

- The ability to provide a basic plan at the proposal stage, with revisions/updates at the Just-in-Time (JIT) stage. Standards for JIT requests should be consistent.
- More clarity on what "sharing in a timely manner" means for types of data.

- More reminders of compliance responsibilities around privacy and consent, among other regulatory issues.

Section II: Definitions:

Comments on definitions are included in our response to Section VI: Data Management and Sharing Plans, and the Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan. We will appreciate clarification on the definitions of "scientific data", "publicly available", and timeframes for availability.

Section III: Scope:

The December 16th webinar on this topic indicated that training grants may be excluded from this policy. We appreciate all efforts to clarify and evaluate requirements for different funding mechanisms, for which the primary goal may not be the generation of scientific data.

Section IV: Effective Date(s):

Will the policy be issued under the Notice of Proposed Rulemaking Process, i.e, will there be an opportunity to comment on the draft policy in its final form?

Section V: Requirements:

We appreciate the NIH's effort to minimize researcher administrative burden, but we are concerned that a Just-In-Time approach to submission of the Data Management Plan (DMP) will impact various parts of the application lifecycle, including the investigator's ability to get feedback on the DMP from review, and allowing them the chance to fully consider, and budget for, the needs under the proposed plan. We hope that the draft policy will come with supplemental information to provide robust guidance for preparation of the plan (including best practices and samples of a good plan), and sample costs to consider in the proposal budget. We reiterate support for reducing researcher burden. More information is listed in the comments on the supplemental draft guidance on allowable costs.

Section VI: Data Management and Sharing Plans:

As mentioned, we appreciate the efforts to minimize burden, but are concerned that shifting the requirement to the JIT stage will not encourage thoughtful planning. For NSF applications, the plan is subject to the formal peer review process. It will be helpful to understand how feedback on the plan will be conveyed to the investigator.

How will regular updates to plans be published/tracked/monitored for compliance?

Section VII: Compliance and Enforcement:

Including the DMP as a Term and Condition in the award, contract, or other funding agreement indicates that the grantee institution is responsible for compliance with the plan. In addition, submission of materials at the "Just-In-Time" phase requires institutional signature/approval. Both points imply that the institution is responsible for compliance. Please provide guidance on the expectation to monitor and assure compliance with data management plans, including assurances that NIH Institutes and Centers will handle compliance review consistently. Clarity on this process will help ensure institutions can effectively advise investigators, and not add layers of administrative burden.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We recommend a balance between submission of the DMP at Just-In-Time (reducing researcher burden) and ensuring that the researcher has adequate guidance to budget allowable direct costs at the application stage. The NIH and awardee institutions could work together to develop resources to estimate costs to provide needed information to researchers and the administrative staff who support them.

Without having information on estimating costs, several issues arise:

- Will grantees resort to submitting detailed budgets when the costs for data management exceed the modular budget direct cost cap? Will this ultimately reduce the number of awards available for investigators?
- The recovery of the Administrative piece of the F&A rate is capped at 26%, which does not represent full recovery of allowable administrative costs. These requirements will stretch an already over-extended rate.
- Without good examples of multiple data management/sharing scenarios, researchers will not have adequate planning information both for their plans, and for their budgets.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We look forward to seeing the additional guidance for management plans referenced in the December 16th webinar. As mentioned above, understanding how updates to the plan will be handled is important, particularly if certain elements of the plan are listed as "TBD" at the time of submission.

We would like additional clarification/definition on several items:

- Publicly available, including any issues in complying with this policy when a data use agreement is in effect.
- Acceptable established repositories, or a site dedicated to this and preferred metadata minimums.

- Standards for long-term preservation, including decision-making around when data will no longer be available
- Expectations on local development to make data available when it is housed locally
- Acknowledgment of costs that extend beyond the award end date
- Tribal sovereignty and Tribal ownership of data for American Indians and/or Alaska Natives, in accordance with CARE principles

Other Considerations Relevant to this DRAFT Policy Proposal:

We would like to see more guidance and information on:

- Expectations of long-term accessibility/usability of shared data. Specific guidance for a lower boundary would be helpful.
- Retention of data and standards of practice for retention time
- Draft agreements for publishers to mitigate copyright issues
- How evolving standards for sharing data, (specifically scientific data including certain individual level and summary or aggregate data as well as metadata) retention, etc. will be updated independently of the over-arching policy
- Proposed language for informed consent documents and guidance for studies collected data prior to the revised Common Rule requirements
- Evaluation of the policy over time, including risks, benefits, and costs
- Requirements for sharing data from internal data management systems, such as REDCap.
- Consideration to eliminate, or make flexible, the proposed two-page limit, given the potential complexity of a project's data sharing needs

Attachment:

Description:

Submission ID: 1392

Date: 1/10/2020

Name: Heather Joseph

Name of Organization: SPARC (The Scholarly Publishing and Academic Resources Coalition)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All Research Data

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: Executive Director

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Making data widely available is an essential element of scientific research. We applaud the NIH for continuing to highlight the critical importance of proactive, systematic management and sharing of data arising from its funded research. As the steward of \$39 billion in annual taxpayer funding, the NIH has an obligation to ensure that its policies are designed to maximize the public's return on this investment.

Recommendation Ia: We recommend strengthening this section by setting a clear expectation that researchers receiving funding from the NIH will be required to manage and share data resulting from NIH-funded or conducted research. In addition, we recommend that this section is an ideal place to signal the potential for data sharing to accelerate the pace of scientific discovery, and to call for it to be shared as soon as practicable. This will indicate that data sharing is an issue of urgency, and this text should replace the word "timely," which is too vague to signal any meaningful timeframe expectations.

Section II: Definitions:

Data Management and Sharing Plan

While we appreciate the increased focus on proactive, prospective attention to creating plans for managing and sharing research, the definition of "Data Management and Sharing Plan"

currently included in this policy could be strengthened to better reflect the how integral DMP/DSPs are to the research process.

Recommendation IIa: The policy should make it clear that data sharing is not an add-on, but an ongoing management process that must be fully integrated into the scientific research process.

Data Management.

The definition of Data Management is a critical element of this policy and should reflect this importance.

Recommendation IIb: We support the 2018 AMIA definition of data management and recommend that the NIH adopt it, replacing the current definition text with the following text: "The upstream management of scientific data that documents actions taken in making research observations, collecting research data, describing data (including relationships between datasets), processing data into intermediate forms as necessary for analysis, integrating distinct datasets, and creating metadata descriptions. Specifically, those actions that would likely have impact on the quality of data analyzed, published, or shared."

Scientific Data

The policy defines scientific data as the "recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications." While we support this in concept, we are concerned that this wording still might reinforce the notion that the only research data of value is data that results in a journal article or similar publication.

This data tends to be data that positively supports a researcher's hypothesis; it's rare for journals to report on negative results. Negative results are too often considered "failures," yet basic research (and clinical trials) that fall into this category often result in data that may be extremely valuable to other researchers and research efforts.

Recommendation IIc: The NIH's general data sharing policy should address this problem by developing policies that would encourage scientists to share data and results even from projects that went unpublished. The NIH could consider rewarding cases in which researchers

take the trouble to share data even if otherwise unpublished. We recognize that this may have a significant effect the scope of the policy as well.

New Definition for Scientific Software Artifacts Should be Added

In order to maximize the utility and value of NIH-funded research data, scientific software artifacts, such as the code underlying the algorithms and models that process data, should be required to be shared as well.

Recommendation IId: We support AMIA's call for a definition of "scientific software artifacts" and recommend the NIH include the following definition in this policy:

"Scientific software artifacts: the code, analytic programs, and other digital, data-related knowledge artifacts created in the conduct of research. These can include quantitative models for prediction or simulation, coded functions written within off-the-shelf software packages such as MatLab, or annotations concerning data or algorithm use as documented in 'readme' files."

Section III: Scope:

As noted in Recommendation IId in Section II, scientific software artifacts are critically important research findings. Managing and sharing the means of manipulating data from one form to another, transforming raw inputs into valuable outputs, is also important to the end goal of rigorous, reproducible, and reusable science.

Recommendation IIIa: We recommend that the Scope section include the following statement:

"NIH funded research produces new scientific data and metadata, as well as new scientific software artifacts (e.g. the code of algorithms and models used to manipulate data). Software artifacts are outputs of research as much as data, and it is just as important to manage and share them in the interest of rigor, reproducibility, and re-use. NIH's commitment to responsible sharing of data extends to scientific software artifacts. As such, throughout this policy, the use of the term "data" should be understood to include scientific software artifacts, per the definition established in Section II."

Section IV: Effective Date(s):

We recognize that it difficult to establish a blanket timeframe for this policy's implementation across the broad range of institutes, research project types, and disciplines that the NIH is responsible for. In some instances, data sharing practices are well established, and trusted repositories are easily identifiable. In others, neither may yet be the case. It's important that the NIH establish a framework to enable those research projects/areas where readiness levels are high to begin to comply as quickly as possible, and also help accelerate the readiness levels of other research areas.

Recommendation IVa: We recommend that the NIH consider having the policy go into effect immediately (i.e., in the next funding cycle) for those research areas that have established practices of data sharing and well-recognized data repositories. For those that do not, the NIH should consider utilizing the expertise resident in scholarly societies/communities to meet with IC's to develop appropriate timeframes for implementation.

Section V: Requirements:

In keeping with Recommendation IVa, we recognize that it would be extremely challenging to establish a blanket policy applying to all grant types and sizes. This policy will require a change in behavior from both researchers and institutions, the ability of investigators to comply will certainly vary widely.

Recommendation Va: We recommend that the NIH create a tiered approach to implementing the data sharing requirement, using tiers that start with the NIH's current strata (i.e., starting with awards over \$500,000) and working downwards, tailoring implementation times to each tier.

Section VI: Data Management and Sharing Plans:

Encourage vs. Require

As we noted in our comments in Section I, making data widely available is an essential element of scientific research. Yet the language in this section notes that the "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." We find this language problematic.

Recommendation VIa: We recommend strengthening this policy by setting a clear expectation that researchers receiving funding from the NIH will be required to manage and share data resulting from NIH-funded or conducted research.

Require Data Management and Sharing Plans in Funding Application

As this draft policy is written, Data Management and Sharing Plans are essentially submitted on a "Just-in-Time" basis. This sends the message that that these plans are not an essential or valued part of the application.

Recommendation VIb: The quality Data Management and Sharing Plans should be an integral consideration in the NIH funding decision process. We recommend that these plans be required as a scoreable part of the application so that appropriate costs can be budgeted at the time of application, and also so that these plans can be included in the critical review process.

Make DMSPs Publicly Available

This section states that, "NIH may make Plans publicly available." We believe that the NIH should ensure transparency with the public who has funded the work, and take advantage of transparency as a means for encouraging compliance.

Recommendation VIc: We recommend this section state that "NIH will make Plans publicly available."

Section VII: Compliance and Enforcement:

As with any policy, the inclusion of both positive and negative incentives will greatly improve compliance rates. Currently, the compliance elements of this Policy are limited to negative reinforcement. The Policy currently notes that "The NIH ICO approved Plan may be taken into account by the funding NIH ICO for future funding decisions for the recipient institution." We fully support this approach and this specific language.

Recommendation VIIa: We also recommend adding language to incentivize compliance through positive reinforcement. We recommend that if a recipient institution has performed well toward a robust DMSP -- that is, it planned for extensive data sharing and followed through by sharing data extensively -- that institution should have an additional advantage toward future applications. The language in this section should be updated to reflect this incentive. Additionally, we recommend that an application in which more robust data management and sharing activities are clearly planned should be scored higher than those with weaker or no activities planned. The criteria for scoring plans should be clearly articulated in supplemental guidance.

We appreciate the fact that the NIH has indicated that this comment period is just one part of a longer process of iterating with community members to refine this data management and sharing process over time.

Recommendation VIIb: One key element that will greatly improve the policy's chances of success is to outline what specific evaluation mechanism(s) it will use to monitor and evaluate investigator progress towards achieving the goals of their Data Management and Sharing Plans. Ideally, the evaluation process should be transparent so that the community can help refine it over time.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

As libraries, we are keenly interested in ensuring long-term preservation of and access to research data. While we'd love to see research outputs preserved indefinitely, we recognize that preservation has costs, and that those costs will extend far beyond the duration of the funding period for any given NIH grant. The current policy does not specify whether costs to preserve data beyond the duration of the funded grant will be allowable costs.

Recommendation VIIIa: We recommend that this section provide detail as to whether NIH will cover data preservation costs after the funding period and, if so, for how long.

A key element of successful research data management that is too often overlooked is the expertise needed to collect, verify, format, standardize, annotate, etc. that data. This requires the time and effort of not only investigators, but also of research staff internal to the investigating institution. The draft guidance does not specify whether personnel costs are allowable expenses related to data sharing.

Recommendation VIIIb: We recommend that this section provide detail as to whether NIH will cover such personnel costs, and what allowable costs levels might look like.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The draft policy offers some helpful initial guidance for investigators on which data to share and preserve, based on scientific utility, validation of results, availability of suitable data repositories, privacy and confidentiality, cost, consistency with community practices, and data security. However, we encourage the NIH to expand this list to include other dimensions that are important to consider.

Recommendation IXa: Specifically, we recommend that the NIH consider providing additional detail on considering formats for data sharing, along with software artifacts, algorithms, and other complimentary materials. As with the choice of which data to preserve and share, the NIH should offer criteria for decisions in each of these areas as well.

The draft policy current asks investigators to describe "the degree of data processing that has occurred (i.e., how raw or processed the data will be)." Since the definition of Scientific Data provided in this guidance does not address this question, it appears to be up to individual investigators to choose both the level of processing and/or curation of the data to share, and how much data to share at different levels of processing/curation.

Recommendation IXb: We recommend that the NIH encourage (or even require) the sharing of data at all levels along with descriptions of appropriate data processing at each level. It might also be useful to consider a new category to deal with raw data that includes case report forms as well as the underlying raw data types for other studies.

The question of where to deposit data is central in the minds of all investigators. Currently, this draft policy guidance is very light on details that will help investigators locate and choose trusted repositories.

Recommendation IXc: We recommend that NIH work with experts from the library, archive, and repository communities to provide a list of specific criteria for investigators to use to judge whether or not a repository is an appropriate choice. Additionally, many researchers are unaware of existing repository infrastructure. We recommend that NIH provide a list of existing data repositories in this guidance – again in collaboration with experts from the library, archive, and repository communities. Over time, NIH may also want to consider providing guidance and resources to help investigators choose appropriate repositories.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1393

Date: 1/10/2020

Name: Maryrose Franko

Name of Organization: Health Research Alliance

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: biomedically-relevant data

Type of Organization: Other

Type of Organization - Other: Nonprofit nongovernmental research funders

Role: Other

Role - Other: Executive Director - Administrator

Domain of Research Most Important to You or Your Organization:

All biomedically-related research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

These comments are my own and do not represent those of the Health Research Alliance. However, they are informed through discussions with members of the HRA and my involvement in the National Academies of Sciences, Engineering, and Medicine's Roundtable on Aligning Incentives for Open Science, among other sources.

To summarize all of my comments, the NIH Policy for Data Management and Sharing and Supplemental Draft Guidance is far too cautious and as written is a missed opportunity to increase the impact of NIH funded research. This is NIH's opportunity to catalyze real change in this area and should not be squandered by using half-hearted and equivocal language in this policy.

In the Purpose section (as in all sections) words with no teeth like "should" need to be replaced by words notifying that there are consequences to noncompliance like "must." NIH must do more than "encourage" or "expect" – it must "require."

It must also enforce these requirements and penalize those who do not comply all along the process, from application stage, through the term of the award, and even after the funding period has ended. It is not unusual for researchers to need to share their data after the term of the grant has ended in a way that is findable, accessible, interoperable and reusable.

The Purpose section actually highlights the benefits of sharing data much better than other parts of the draft. It is well-stated that data sharing is critical to not only test the validity of research findings, but also to strengthen datasets through combining datasets, reuse hard-to-generate data, and explore new frontiers of discovery. Datasets alone are valuable outputs of research and that point must be made clear in the document and in practice by rewarding both sharing data and using other's data for the purposes above. NIH has several powerful examples of where reuse of data by others has been critical to pushing new discoveries. Datasets do not need to be cited in a publication to be valuable to the scientific community. NIH needs to make it clear that datasets (and other research outputs such as code) need to have digital object identifiers so that researchers can highlight and get credit for the development and public sharing of these outputs, and others can cite the reuse of these outputs. Citation of publicly available datasets (and other outputs such as code) by others can be a powerful incentive for researchers to publicly share their data.

In this sentence "Shared data should be made accessible in a timely manner..." not only should the word "should" be changed to "must" but the word "timely" needs to be clarified. This should be defined as no later than the time of publication. Exceptions to this timing must be justified. In the absence of a publication, the timing is a bit trickier. One option would be to require that researchers make data available in a repository that complies with the FAIR principles within 1-3 years of the approval of the Data Management Plan. As above, exceptions to this timing must be justified.

Section II: Definitions:

The FAIR principles are mentioned at a very high level in the Purpose section but do need to be defined and their importance highlighted either in that section or in the Definitions section.

The definition of a "Data Management Plan" needs to be broadened to stress the importance of the plan across the entire scientific research process and must include several pieces of basic information (though the details can be spelled out in the Supplemental Guidance as done in this draft.)

- The plan must describe clearly how scientific data will be managed across the entirety of a research project.

- It must include details for how and when resulting data will be shared.
- It must include in which FAIR-compliant repositories the data will be stored.

"Data Management" is a term that needs more detail than that given in this document. The 2018 AMIA definition (<https://www.amia.org/sites/default/files/AMIA-Response-to-Draft-DMSP-RFI.pdf>) is much more helpful in clarifying what is really meant and expected when using the term Data Management and its inclusion would be a valuable clarification. "The upstream management of scientific data that documents actions taken in making research observations, collecting research data, describing data (including relationships between datasets), processing data into intermediate forms as necessary for analysis, integrating distinct datasets, and creating metadata descriptions. Specifically, those actions that would likely have impact on the quality of data analyzed, published, or shared."

As stated above, when defining "Scientific Data" it is important to explicitly state that scientific data is more than just the factual material commonly accepted in the scientific community as necessary to validate and replicate research findings. The NIH policy needs to put in place policies that incentivize researchers to share data and results, EVEN FROM PROJECTS THAT WENT UNPUBLISHED OR YIELDED NEGATIVE RESULTS. Applicants and grantees should be rewarded for their use of preprint servers and FAIR-compliant repositories to post results and accompanying data and code from projects even when the results were negative and/or not published in a peer-reviewed journal. The research community can only benefit from data that was collected as part of a project but not used in a publication if the data and results are shared with the community according to FAIR principles.

In addition, if one of the goals is reproducible and reusable science, sharing scientific data needs to specifically include sharing the means of manipulating data from one form to another, and transforming raw inputs into valuable outputs. Researchers must not be allowed to withhold key artifacts necessary to make data valuable for reuse. This clarification can be accommodated either in this Definition section or in the Scope section.

Section III: Scope:

It is a significant improvement that NIH is not limiting the policy to awards requesting \$500,000 or more in direct costs per year and that the policy specifically applies to all research funded by NIH regardless of NIH funding level or mechanism. It needs to be made clear that this policy applies to funding for fellowships, training grants and other mechanisms that focus on career development and training in addition to grants to support research.

It also needs to be specified that this policy applies to scientific data produced by NIH funding in part or in whole, even after the term of the award has ended.

Section IV: Effective Date(s):

Of course, the effective date is dependent on the community's feedback on this draft policy. However, there is no reason NOT to include a specific "no later than" date. Ideally, the policy should state that implementation will be no later than 12 months after issuance of the final policy. If a tiered adoption of the policy is to be implemented (such as projects funded over \$500,000 per year would have to comply within one year of approval of the DMSP, those between \$250,000-\$500,000 within two years, and those below \$250,000 within three years) that needs to be stated explicitly in the final policy.

Section V: Requirements:

The NIH requirement should state explicitly that "NIH requires ALL NIH-funded researchers to share ALL scientific data generated in a study, with exceptions only when justified to a panel that includes subject matter and data science experts."

The policy should state that scientific data must be deposited in a FAIR-compliant repository and receive an accession code or DOI. Researchers may embargo this data until publication. If the data is not to be cited in a peer-reviewed publication, for any work that would be reported on a progress report, researchers must use preprint servers or other methods to share the datasets, highlight the work, and provide the metadata so that the scientific community and a much broader audience and can benefit from the work funded by the NIH.

"All researchers" means just that – all those who are awarded NIH research project grants, contracts, career grants, fellowships, intramural scientists, etc.

NIH should also specify that the repository for the data (and the software artifacts mentioned above) must comply with the FAIR principles.

Also, as part of the policy and application process, it is important to highlight that sharing data and other research outputs will be rewarded.

The NIH can encourage sharing by adding wording to applications such as:

"If applicable, describe

- 1) instances where you have engaged in "open" activities (such as making articles open access and sharing data/code according to FAIR principles)
- 2) examples of how your open research outputs have been used by others in your discipline, in other disciplines, and/or outside of academia (include DOIs if possible), and
- 3) plans to engage in open activities in the future"

The NIH can also modify the biosketch instructions to include language such as:

"If (public) sharing of your research outputs such as data, code, or material led to scientific advances by others, you are encouraged to detail that as well."

A powerful incentive is "getting credit" that a researcher can point to in a funding application, promotion and tenure package, etc. Adding the above language to NIH applications and to biosketch instructions which are widely used across sectors would give researchers the appropriate credit for creating and sharing a dataset that others found useful.

Of course, asking researchers to use DOIs to cite data from others they have used in their research is very important as well. For this sharing to become the norm and accelerate the rate of discovery, datasets must be shared and have digital object identifiers.

In summary, the policy should specifically highlight the fact that data sharing and other open behavior will be rewarded by the NIH and the NIH must put in place that reward structure.

Section VI: Data Management and Sharing Plans:

One of the most important changes to this document as detailed above must be to REQUIRE not EXPECT or ENCOURAGE NIH-funded researchers to share data.

Another significant change is that NIH must include data management and sharing plans as part of the REVIEW PROCESS. NIH must require a DMSP as part of the initial application for funding. It cannot be "part of Just-in-Time for extramural awards, or as part of the technical evaluation for contracts, as part of the NIH Intramural Annual Report, or prior to the release of funds for other funding agreements."

DMSPs need to be included in the regular submission of an application. Because EFFECTIVE DMSPs will increase the impact of the research, they need to be rewarded by increasing the overall score of an application. Reviewers receive guidance on how to score significance and approach and should also receive guidance on scoring effective DMSPs. The guidance should be based on how effectively the plan addresses the 15 FAIR principles.

NIH can create and encourage the use of templates for DMSPs which can help to standardize the essential elements and layout of DMSPs (these can be IC specific). Researchers who do not use the standard template are not penalized, as long as the essential elements are included, and the plan addresses the FAIR principles sufficiently.

This section states that "NIH MAY make Plans publicly available." This is not sufficient. The NIH MUST make plans publicly available. Making awardees' plans public will ensure transparency, which can help to encourage compliance. The NIH should publish DMSPs for funded awards (grants, contracts, fellowships, etc) with the abstracts in the NIH RePORTER. Knowing that these plans are available to the public will increase compliance among researchers. NIH can't police compliance 100%, though random audits would be valuable along with review of compliance during annual reports. Publishing plans in RePORTER also provides transparency and enables community scrutiny after the grant term as ended. It should also be made clear that all scientific data must be shared, and that any exceptions to this must be justified and subsequent funding conditioned on approval by and NIH advisory committee of scientific subject matter and data management experts.

A Board or Advisory council must also be responsible for oversight of DMSPs for intramural researchers. It is not prudent to give a single NIH official (such as the Scientific Director or Clinical Director) the ability to review and approve these plans.

Section VII: Compliance and Enforcement:

As mentioned many times above, if the DMSPs are included as part of the application, there is a strong positive incentive for developing a robust plan for data management and sharing.

Asking researchers to provide the DOIs for datasets should be integral to the progress report process. Datasets cited in progress reports should be administratively reviewed for compliance to the approved DMSP included in the application. If the data is not shared as detailed in the plan, and there is not sufficient justification as to why not, future funding can be withheld. DMSPs can be modified during the term of the award, but these modifications must be reviewed and approved before they can take effect.

To help compliance during and after the award term, the DMSPs need to be included (in a machine-readable fashion) in NIH RePORTER. That record needs to also have contact information to request corrective action for violations of the NIH's Data Management and Sharing Policy, or the published DMSP. The contact information needs to include PIs' or project directors' email addresses, and contact information for the institutional official at the grantee institution. It is also important to include an NIH contact to whom queries should be sent. This gives enough information to handle the issue at the PI or grantee institution level first, but also enables escalation to the NIH when necessary. Similar contact information should be included with the DMSP in NIH RePORTER for all mechanisms.

Violating the policy, even after the end of the award period, can result in sanctions. Sanctions can include prohibiting application for subsequent funding, prohibition of service on NIH advisory committees, board or peer review committees, publishing a notice describing the violation in the NIH Guide to Contracts and Grants, or debarment from contracting, subcontracting, or financial assistance from the NIH.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Guidance should specify that costs related to curating, developing supporting documentation, and preserving data are allowed beyond the funding period (including personnel).

It should also stipulate that only costs to deposit data into repositories that comply with the FAIR principles will be allowable.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The guidance on the elements of a plan is very important, and the NIH should be commended for including as much guidance in this area as possible. I will reiterate one point – that DMSPs MUST be part of the original application and the plan's effectiveness scored as part of the overall application score.

Without scoring this as part of the application, the incentive to develop a robust plan is significantly decreased. In addition, to say that "if certain elements of a Plan have not been determined at the time of submission, an entry of "to be determined" may be acceptable if a justification is provided ..." completely defeats the purpose of implementing this policy. In addition, what is gained by stating "NIH does not expect researchers to share all scientific data generated in a study"? Again, this sentence completely undermines the purpose of implementing this policy.

This entire section asks researchers to "consider" doing things like describing the types and amounts of data, which data will be preserved, and shared, specifying how needed tools can be accessed, or providing the name and URL of the repository to be used, or indicating why an existing repository will not be used. These need to be REQUIREMENTS – not mere considerations.

The section needs to be rewritten to say as part of the plan "you must" instead of "you should consider".

For example:

You must describe the types and estimated amounts of data (section 1)

You must describe which scientific data will be preserved and shared (section 1)

You must specify how needed tools can be accessed (section 2)

You must specify when scientific data will be archived to ensure long-term preservation (section 4)

You must provide the name and URL of the repository (section 4)

You must justify why an existing FAIR-compliant repository will not be used (section 4)

If scientific data will be shared through a restricted mechanism, you must describe the terms of access for that data (section 4)

A related note, if an existing repository is not to be used, justification as to why not must be included and needs to be subject to approval by the NIH.

Also, at the end of section 4, the statement "IN GENERAL, scientific data should be made available as soon as practicable, independent of award period and publication schedule. There is absolutely no need to include the qualifier "in general."

In section 5, "CONSIDER INDICATING" needs to be changed to "MUST INDICATE".

"In describing proposed plans for managing data sharing agreements and other types of agreements, YOU MUST indicate:

- restrictions imposed by exiting agreements

- etc

Another important point is that the second bullet under section 5 enables applicants to choose to enter into restrictive agreements to avoid sharing data. To close this loophole, the plan needs to be part of the initial review, as mentioned before. Reviewers can evaluate the impact of these restrictive agreements and consider the decrease in impact as part of the overall score. Restrictive agreements entered into AFTER the grant is made (or the funding mechanism is approved) must be approved by an advisory committee that includes content and data science experts.

Other Considerations Relevant to this DRAFT Policy Proposal:

It is critical that incorporated in the philosophy of the policy and made explicit in the final NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance that THE DEFAULT IS THAT DATA MUST BE MADE AVAILABLE AS SOON AS PRACTICABLE. No qualifiers should be added. The use of "should" and "consider" and "in general" create the impression that the NIH is not really committed to increasing data sharing among its research community. Let me reiterate what I said at the beginning of my comments, this is NIH's opportunity to catalyze real change in this area and should not be squandered by using half-hearted and equivocal language in this policy.

Attachment:

Description:

Submission ID: 1394

Date: 1/10/2020

Name: Twila Reighley

Name of Organization: Michigan State University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Complying with Administrative and Regulatory Requirements

Type of Organization: University

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

We are writing in support of the Council on Governmental Relations (COGR) comments on the draft policy. COGR has covered most of the concerns we have in reviewing the policy and has made suggestions for improvements. We want to emphasize a few points COGR made in their response. We would like to see consistency between the NIH Institutes and Centers, support to metadata collection, NIH host data repositories, and clarifications that protect privacy and security, e.g., when human subjects are involved, please acknowledge the role of the Institutional Review Board (IRB). We would be happy to be included in discussions of improving costing methodology and encourage NIH to further work to reduce the increased costs and administrative burden associated with the implementation. We believe there should be an appropriate embargo period to allow time for data analysis and intellectual property advancement.

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We would like to see additional discussion between NIH and impacted organizations to discuss the allowable costs guidance of the data sharing policy at future roundtable meetings or other public forums. For instance, we request clarification that although there have been budgeting restrictions placed on what can be charged as data management and sharing that does not limit budgeting for costs associated with research to projects. In other words, the statement that follows from the guidance should relate only to data management and sharing costs, but not to budgets in total and clarification would be helpful: Estimates should not include . . . costs associated with the routine conduct of research. Costs associated with collecting or otherwise gaining access to research data (e.g., data access fees) are considered costs of doing research and should not be included in budgets.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:**Other Considerations Relevant to this DRAFT Policy Proposal:****Attachment:**

MSU comments NIH Data Sharing Policy.pdf

Description:

Michigan State University Comments on NIH Data Sharing Policy

MICHIGAN STATE **UNIVERSITY**

January 10, 2020

Dr. Andrea Jackson-Dipina
Director of the Division of Scientific Data Sharing Policy
National Institutes of Health
Office of Science Policy
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Subject: Comments to DRAFT NIH Policy for Data Management and Sharing

Dear Dr. Jackson-Dipina:

Thank you for the opportunity to provide input on the subject document. We recognize that data sharing will benefit long-term research progress and results. At the same time, implementation needs to take an affordable, pragmatic approach.

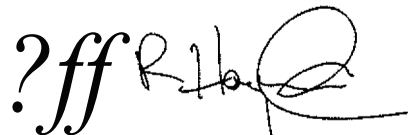
We are writing in support of the Council on Governmental Relations (COGR) comments on the draft policy. COGR has covered the issues we see with the policy and has made suggestions for improvements. We want to emphasize a few points COGR made in their response. We would like to see consistency between the NIH Institutes and Centers, support to metadata collection, NIH host data repositories, and clarifications that protect privacy and security, e.g., when human subjects are involved, please acknowledge the role of the Institutional Review Board (IRB). We would be happy to be included in discussions of improving costing methodology and encourage NIH to further work to reduce the increased costs and administrative burden associated with the implementation. We believe there should be an appropriate embargo period to allow time for data analysis and intellectual property advancement.

We share NIH interest in making results available to the public and improving the reproducibility and reliability of research results. Thank you for your efforts.

Sincerely,



Twila Fisher Reighley
Asst. VP for Research and Innovation



Joseph R. Haywood
Asst. VP for Regulatory Affairs



**Office of the Senior
Vice President for
Research and
Innovation**

Sponsored Programs
Administration

Hannah Administration Building
Michigan State University
426 Auditorium Road, Room 2
East Lansing, MI 48824

517-432-355-5040
Fax: 517-432-3337

Submission ID: 1395

Date: 1/10/2020

Name: Collaborative Study on the Genetics of Alcoholism (COGA)

Name of Organization: COGA

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

genomics, clinical course, psychiatry, neuroscience, epidemiology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

(a) As scientists, we recognize the necessity and benefits of data sharing. We view this movement of data sharing and management as inspirational and forward thinking. As a matter of policy, we share all our genomic data and a very extensive array of our phenotypic data. However, there are some limits on what can be shared because of the nature of our dataset, which is based on multigenerational family pedigrees, and because we collect extensive personal health information. Safeguards to avoid re-identification are fundamental to the trust needed for human subject research and issues of identifiability are complex. We follow the best guidance available in de-identifying data, and we consult regularly with our IRBs prior to data release internally as well as to NIH repositories. Fuller use of scientific data will enhance the work that we do and accelerate discovery. However, legacy studies - such as COGA - cannot be retrospectively engineered to meet these standards.

(b) Datasets such as COGA are rich and longitudinal. We have shared lifetime diagnostic outcomes within repositories. However, additional nuanced characterization of data can be difficult to communicate. COGA maintains extensive documentation of data collection timelines, coding, versioning; however these are written at a level of detail for COGA researchers. Again, as a legacy project, it is not clear what expectations are for previously collected/archival data as collections continue. As noted, we cannot always go back to re-consent subjects. Retro-fitting previously collected human subjects data for broader sharing may not be possible and may contribute to inadvertent errors in subsequent data analysis in the absence of collaborations with investigators from the study.

(c) Methods for full de-identification at best require substantial resource investment (e.g., GUIDS) and may require data points that are either unavailable or, by HIPAA regulation, cannot be shared outside IRB approved settings in legacy projects.

Section VII: Compliance and Enforcement:

While future projects may be designed to consent participants appropriately, mandated public sharing from legacy projects (COGA) may be precluded in the absence of major funding to locate and re-consent subjects, and some subjects cannot be re-consented because of refusal, not being located or death. The many complexities outlined above include financial, regulatory and administrative burdens on investigators and staff for projects involving human subjects in legacy projects.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

COGA has always had a strong internal data management plan, which is what has made our multi-site effort so successful. We use common linking variables to connect various data modalities. We have funds dedicated towards data management and this is a critical component of our large-scale collaboration. Data management goes beyond merging types of data – it involves uniform assessment development, evaluation, piloting, harmonization of quality assurance protocols, algorithms, data, investigator response. It is a significant undertaking requiring considerable effort and resources, and the current proposal does not provide clear guidance on whether allowable costs will account for the extent of management that will be required to provide standardization and preservation not just for internal use but for external sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The requirements for documentation are too extensive, requiring prospective documentation of the lifespan of the data beyond the availability of funding resources. Several of these aspects may adapt to changing scientific insights and therefore, such a data sharing plan, while meaning to provide a general plan of action, should not be seen as the only approach to data management, preservation and sharing. Such a process is best served when dynamic, allowing

for flexible alterations as science changes and sufficient funding to meet university, state and federal regulatory needs.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1396

Date: 1/10/2020

Name: Juliane Baron

Name of Organization: Federation of Associations in Behavioral & Brain Sciences

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Basic behavioral

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: Non-profit

Domain of Research Most Important to You or Your Organization:

Basic behavioral science

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The Federation of Associations in Behavioral and Brain Sciences (FABBS) strongly supports the purpose of the DRAFT NIH Policy for Data Management and Sharing. FABBS shares the commitment to rigorous research, protection of privacy, and effective and efficient data management. FABBS recommends that the purpose also include encouraging researchers to work to reproduce findings as well as to use existing data rather than collecting new data. Current incentives lead researchers to pursue new research with new data.

Section II: Definitions:

No Comment

Section III: Scope:

No Comment

Section IV: Effective Date(s):

No Comment

Section V: Requirements:

No Comment

Section VI: Data Management and Sharing Plans:

FABBS appreciates the flexibility provided in this section, allowing researchers to determine where to deposit their data and which data to share. However, FABBS encourages NIH to provide additional guidance on the characteristics of acceptable repositories and about expectations regarding which data need be shared.

Section VII: Compliance and Enforcement:

No Comment

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

FABBS appreciates the inclusion of supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing. Researchers will almost certainly need additional time and resources to prepare and store data.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

FABBS appreciates the flexibility of required data management plans. Basic behavioral studies can be iterative in nature, this flexibility allows for evolving studies.

Other Considerations Relevant to this DRAFT Policy Proposal:

As NIH moves forward finalizing the data management plan, please be mindful of other requirements for NIH-funded research, such as Clinical Trial reporting, so as not to produce unintended, adverse consequences.

Attachment:

Description:

Submission ID: 1397

Date: 1/10/2020

Name: Dr. Lisa Simpson

Name of Organization: AcademyHealth

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: all of the above

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Health services research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

AcademyHealth believes it is crucial to pay careful attention to context and rationale provided for the purpose of data sharing. The argument to support enhanced data sharing must be compelling and it needs to resonate with constituents to garner support. The first sentence of the policy describes NIH's commitment to "making results and outputs of the research it funds and conducts available to the public." We would argue that the scope of data management and sharing should be broadened to encompass data creation and collection, and not solely focus on results and research outputs. At the outset, the policy should state not only the role of shared data in optimizing research results, but also underscore broader implications for the field, including enhanced collaboration, transparency and accountability.

In terms of setting context, how the benefits of data sharing are communicated, including the order in which they are described, are important. Currently, "to test the validity of research findings" is listed as the first (and presumably most important) benefit of data sharing. We would argue the benefits may be better expressed tied to the goals of NIH, and the value of reusing data for deeper or broader discovery should be the emphasis.

Further, while FAIR principles are valuable for considering data sharing, it is unclear whether investigators subject to the proposed policy for data management and sharing are responsible for each of these principles. We believe that the focus of sharing should be to make the data accessible and interoperable, and the role of the NIH is to make data findable and usable. As currently written, investigators sharing the data take on the full burden, which may be counter-productive. We suggest revising the plan such that it diminishes investigator burden through technical, infrastructure and cost sharing across the spectrum of stakeholders in the publication and sharing of data.

Section II: Definitions:

Data Management and Sharing Plan (Plan): The definition of the Plan in the proposed draft should be considered more broadly. Rather than just defining how the data will be "managed, preserved and shared," it is important that the Plan also identify at least one target "other" who would use the data, and describe how that user would actually use it. Otherwise, the Plan will only describe how to make it possible, rather than to enable it. The definition also does not address how the scientific data will be collected or described in the shared data set, which are important components of the Plan.

Data Sharing: The proposed definition for data sharing is limited, and should more explicitly state that sharing is more than increasing access. Sharing data must involve enabling the access and reuse of data to facilitate and optimize research. In addition, because it is important to understand how results were determined, and considering that current data analytics provide unique methods of data analysis and interpretation, sharing the associated code that was used to determine the accuracy of analyses should be required.

Scientific Data: The proposed definition of scientific data explicitly excludes "preliminary analyses" as eligible material. However, many data will be rendered futile for the purposes defined in data sharing without a clear understanding of preliminary analyses, and in some cases access to that data. At a minimum, the definition could state the scientific data "may exclude" preliminary analyses, rather than an unequivocal exclusion.

Section III: Scope:

We believe an important opportunity inherent in this policy is to expand access to data extracted from electronic health records for secondary analyses, which represents a growing area of important research. It is possible that many studies using secondary data analysis will not explicitly de-identify the data if not legally required (e.g., if the data do not leave the institution). To mitigate this potential limitation, the scope should be broadened to recommend the creation of shareable data sets to support these analyses.

Section IV: Effective Date(s):

no comments

Section V: Requirements:

NIH guidance on the submission of the Plan should be as specific and prescriptive as possible. For example, the "Requirements" section notes that the submission of the Plan should outline how scientific data will be managed and shared, "taking into account any potential restrictions or limitations." Instead of taking into account potential restrictions or limitations, the Plan should clearly describe what those restrictions and limitations are. For example, if the data contains proprietary information that imposes restrictions on sharing, this should be clearly described in the Plan prior to NIH funding decisions. The "Requirements" section should clearly state which elements of the Plan are required and which are optional.

We agree with the statement that "additional or specific information" may be requested to meet expectations for data management and sharing. NIH may consider elaborating on this statement and stating that the creation of shareable de-identified data sets may be requested (even if not required for observational studies). We encourage NIH to consider how to ensure sufficient resources (budget) exist to support that aspect of the Plan.

Section VI: Data Management and Sharing Plans:

AcademyHealth members recommend several suggestions related to the Plan, which are organized around topic area below.

Data security and privacy. The draft policy suggests that investigators are responsible for ensuring data security and compliance with privacy protections throughout the life of the scientific data, even after it has been shared. This is a tremendous and daunting requirement. Further, with changing discoveries in security and privacy protections, this requirement may be beyond the capability of most researchers and institutions, and may limit sharing. One strategy to alleviate this burden would be the use of NIH repositories (see next subsection for additional thoughts and comments).

Use of repositories. The draft policy states that NIH "encourages the use of established repositories for preserving and sharing scientific data." NIH should elaborate on the utility of repositories, why repositories are encouraged, and the criteria for establishing a repository. Given the statement that Plans should identify strategies or approaches to ensure data security and compliance with privacy protections through the life of the data, repository use could be incentivized (e.g. by making administrative supplements available, providing the investigator extra "points" for a track record of using repositories as part of subsequent grant reviews, etc...), not just encouraged. Use of established repositories or repositories within NIH would lift the

burden of privacy and security protections from the researcher and be assumed by NIH, who can better represent the public need for the data.

Plan Elements. The proposed guidance suggests that investigators "consider" addressing specific Plan elements outlined in the supplemental guidance. We believe that requiring applicants to address all elements listed in the supplemental guidance would provide clarity to applicants on expectations for an adequate Plan as well as assist NIH with their review of the Plan.

Making plans public. The policy states that NIH may make Plans publicly available. Revising this statement to indicate that NIH will publish Plans for public consumption removes any uncertainty that this may or may not happen, and promotes transparency and accountability among the community of NIH researchers.

Peer review. As currently written, the proposed policy does not require that Plans be evaluated as part of the peer review process. This separation from Peer Review suggests that specific elements of the Plans are not subject to an acceptable level of scientific scrutiny, and therefore may not carry as much importance as the rest of the study plan. We believe that the Plan for making data accessible for evaluation and further research should be regarded with the same level of importance and integrity as the study plan, and note that doing so also requires the engagement of reviewers with appropriate knowledge and expertise to evaluate the Plan.

Section VII: Compliance and Enforcement:

no comments

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

NIH's guidance on allowable costs for data management and sharing should include generating shareable datasets when legal privacy protections would otherwise restrict the sharing of data.

NIH should also collect information about the anticipated costs to researchers to access the study's scientific data in the intended repository. Data that will carry higher than average access costs become essentially inaccessible.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

AcademyHealth members recommend additional suggestions related to elements of the Plan, which are organized around topic area below.

Introductory Language. The draft guidance states that NIH does not expect researchers to share all scientific data generated in a study. NIH should provide more clarity on what types of specific scientific data researchers are not expected to share. AcademyHealth feels strongly

that scientific data created through NIH funding that will not be shared should include a justification for its exclusion.

Data Type. Descriptions of the modality, level of aggregation, and degree of data processing should be considered basic requirements of the Plan. These descriptions would form the minimally necessary metadata that will make the scientific data understandable to others. Without this information, other researchers would be burdened with producing this basic level of information, impeding and delaying further use and reuse, slowing the progress of science, and undermining the objective of data sharing.

Access to Data. With regard to access to restricted scientific data, the applicant should be required to explain their process for obtaining approval to the restricted data. NIH should be able to determine from the Plan how likely it is that others will also be able to access the restricted data.

Other Use Limitations. There should be specific consideration of limitations imposed by sharing protected health information and identifiable information. These are significant limitations and should be addressed directly, along with proposed strategies to mitigate these challenges.

Other Considerations Relevant to this DRAFT Policy Proposal:

The draft policy alludes to the significance of the HIPAA Privacy Rule, but does not explicitly address the implications for data management and sharing. The HIPAA Privacy Rule identifies research as a public interest and benefit activity, and if such allowances are made, the Plan should include how to address protected health information. Otherwise, it may be easy to exclude research using protected health information from this policy.

An additional element and benefit of data sharing is the importance of open data to spark cross-disciplinary research. Members noted that the availability of valid data resources is a key strategy for bringing new investigators from other fields into health and health care research.

We believe the guidance should be prescriptive enough to provide funding panels clear insight into the likely success of the research project in terms of data sharing-- that is, the likelihood that the study's data could be used to validate or replicate study findings or be reused for further research. The NIH should make it clear to researchers that insufficient data management and sharing plans will affect the likelihood of receiving funding.

Finally, beyond and in addition to the whole of our comments above, we want to reemphasize the importance of providing centralized support – infrastructure, operational guidance, resources, logistical support and training – to the field in order to support meaningful implementation of this policy.

AcademyHealth appreciates the opportunity to provide additional input on the Draft Policy for Data Management and Sharing. We believe clarity and specificity will be paramount to a sound and effective policy – one that clearly communicates context, rationale, requirements, and expectations and advances progress toward the widespread and responsible sharing of data.

Attachment:

AH Response_NIH Policy for Data Mgmt.docx

Description:

AcademyHealth response NIH Data Management

AcademyHealth Comments on the NIH Draft Data Management and Sharing Policy and Supplemental Guidance

Submitted by Dr. Lisa Simpson, President and CEO

January 10, 2020

AcademyHealth represents 4,000 individuals and organizations in the research community using evidence and data to improve health and health care for all. Our organization recognizes the crucial role of data sharing in advancing scientific research, and ultimately improving clinical care and patient outcomes. We commend the National Institutes of Health for their efforts to optimize access to and use of shared data in research. Below, we provide our comments and suggestions on NIH's draft Policy for Data Management and Sharing. These comments intend to build on and reinforce [our feedback](#) provided in December 2018 on NIH's proposed provisions for this policy. At a high level, our feedback offers suggestions for enhancing clarity and specificity with respect to terminology, requirements, and components of a data management and sharing plan (Plan), and perspective on the use of resolute language to effectively communicate NIH's expectations, and commitment to upholding Plans to rigorous scientific standards.

Section I: Purpose

AcademyHealth believes it is crucial to pay careful attention to context and rationale provided for the purpose of data sharing. The argument to support enhanced data sharing must be compelling and it needs to resonate with constituents to garner support. The first sentence of the policy describes NIH's commitment to "making results and outputs of the research it funds and conducts available to the public." We would argue that the scope of data management and sharing should be broadened to encompass data creation and collection, and not solely focus on results and research outputs. At the outset, the policy should state not only the role of shared data in optimizing research results, but also underscore broader implications for the field, including enhanced collaboration, transparency and accountability.

In terms of setting context, how the benefits of data sharing are communicated, including the order in which they are described, are important. Currently, "to test the validity of research findings" is listed as the first (and presumably most important) benefit of data sharing. We would argue the benefits may be better expressed tied to the goals of NIH, and the value of reusing data for deeper or broader discovery should be the emphasis.

Further, while FAIR principles are valuable for considering data sharing, it is unclear whether investigators subject to the proposed policy for data management and sharing are responsible for each of these principles. We believe that the focus of sharing should be to make the data accessible and interoperable, and the role of the NIH is to make data findable and usable. As currently written, investigators sharing the data take on the full burden, which may be counter-productive. We suggest revising the plan such that it diminishes investigator burden through technical, infrastructure and cost sharing across the spectrum of stakeholders in the publication and sharing of data.

Section II: Definitions

Data Management and Sharing Plan (Plan): The definition of the Plan in the proposed draft should be considered more broadly. Rather than just defining how the data will be “managed, preserved and shared,” it is important that the Plan also identify at least one target “other” who would use the data, and describe how that user would actually use it. Otherwise, the Plan will only describe how to make it possible, rather than to enable it.

The definition also does not address how the scientific data will be collected or described in the shared data set, which are important components of the Plan.

Data Management: No comment

Data Sharing: The proposed definition for data sharing is limited, and should more explicitly state that sharing is more than increasing access. Sharing data must involve enabling the access and reuse of data to facilitate and optimize research. In addition, because it is important to understand how results were determined, and considering that current data analytics provide unique methods of data analysis and interpretation, sharing the associated code that was used to determine the accuracy of analyses should be required.

Metadata: No comment

Scientific Data: The proposed definition of scientific data explicitly excludes “preliminary analyses” as eligible material. However, many data will be rendered futile for the purposes defined in data sharing without a clear understanding of preliminary analyses, and in some cases access to that data. At a minimum, the definition could state the scientific data “may exclude” preliminary analyses, rather than an unequivocal exclusion.

Section III: Scope

We believe an important opportunity inherent in this policy is to expand access to data extracted from electronic health records for secondary analyses, which represents a growing area of important research. It is possible that many studies using secondary data analysis will not explicitly de-identify the data if not legally required (e.g., if the data do not leave the institution). To mitigate this potential limitation, the scope should be broadened to recommend the creation of shareable data sets to support these analyses.

Section IV: Effective Date(s)

No comments.

Section V: Requirements

NIH guidance on the submission of the Plan should be as specific and prescriptive as possible. For example, the “Requirements” section notes that the submission of the Plan should outline how scientific data will be managed and shared, “taking into account any potential restrictions or limitations.” Instead of taking into account potential restrictions or limitations, the Plan should clearly describe what those restrictions and limitations are. For example, if the data contains proprietary information that imposes restrictions on sharing, this should be clearly described in the Plan prior to NIH funding decisions. The “Requirements” section should clearly state which elements of the Plan are required and which are optional.

We agree with the statement that “additional or specific information” may be requested to meet expectations for data management and sharing. NIH may consider elaborating on this statement and stating that the creation of shareable de-identified data sets may be requested (even if not required for observational studies). We encourage NIH to consider how to ensure sufficient resources (budget) exist to support that aspect of the Plan.

Section VI: Data Management and Sharing Plans

AcademyHealth members recommend several suggestions related to the Plan, which are organized around topic area below.

Data security and privacy. The draft policy suggests that investigators are responsible for ensuring data security and compliance with privacy protections throughout the life of the scientific data, even after it has been shared. This is a tremendous and daunting requirement. Further, with changing discoveries in security and privacy protections, this requirement may be beyond the capability of most researchers and institutions, and may limit sharing. One strategy to alleviate this burden would be the use of NIH repositories (see next subsection for additional thoughts and comments).

Use of repositories. The draft policy states that NIH “encourages the use of established repositories for preserving and sharing scientific data.” NIH should elaborate on the utility of repositories, why repositories are encouraged, and the criteria for establishing a repository. Given the statement that Plans should identify strategies or approaches to ensure data security and compliance with privacy protections through the life of the data, repository use could be incentivized (e.g. by making administrative supplements available, providing the investigator extra “points” for a track record of using repositories as part of subsequent grant reviews, etc...), not just encouraged. Use of established repositories or repositories within NIH would lift the burden of privacy and security protections from the researcher and be assumed by NIH, who can better represent the public need for the data.

Plan Elements. The proposed guidance suggests that investigators “consider” addressing specific Plan elements outlined in the supplemental guidance. We believe that requiring applicants to address all elements listed in the supplemental guidance would provide clarity to applicants on expectations for an adequate Plan as well as assist NIH with their review of the Plan.

Making plans public. The policy states that NIH may make Plans publicly available. Revising this statement to indicate that NIH *will* publish Plans for public consumption removes any uncertainty that this may or may not happen, and promotes transparency and accountability among the community of NIH researchers.

Peer review. As currently written, the proposed policy does not require that Plans be evaluated as part of the peer review process. This separation from Peer Review suggests that specific elements of the Plans are not subject to an acceptable level of scientific scrutiny, and therefore may not carry as much importance as the rest of the study plan. We believe that the Plan for making data accessible for evaluation and further research should be regarded with the same level of importance and integrity as the study plan, and note that doing so also requires the engagement of reviewers with appropriate knowledge and expertise to evaluate the Plan.

Section VII: Compliance and Enforcement

No comments.

Supplemental DRAFT Guidance: [Allowable Costs for Data Management and Sharing](#)

NIH's guidance on allowable costs for data management and sharing should include generating shareable datasets when legal privacy protections would otherwise restrict the sharing of data.

NIH should also collect information about the anticipated costs to researchers to access the study's scientific data in the intended repository. Data that will carry higher than average access costs become essentially inaccessible.

Supplemental DRAFT Guidance: [Elements of a NIH Data Management and Sharing Plan](#)

AcademyHealth members recommend additional suggestions related to elements of the Plan, which are organized around topic area below.

Introductory Language. The draft guidance states that NIH does not expect researchers to share all scientific data generated in a study. NIH should provide more clarity on what types of specific scientific data researchers are not expected to share. AcademyHealth feels strongly that scientific data created through NIH funding that will not be shared should include a justification for its exclusion.

Data Type. Descriptions of the modality, level of aggregation, and degree of data processing should be considered basic requirements of the Plan. These descriptions would form the minimally necessary metadata that will make the scientific data understandable to others. Without this information, other researchers would be burdened with producing this basic level of information, impeding and delaying further use and reuse, slowing the progress of science, and undermining the objective of data sharing.

Access to Data. With regard to access to restricted scientific data, the applicant should be required to explain their process for obtaining approval to the restricted data. NIH should be able to determine from the Plan how likely it is that others will also be able to access the restricted data.

Other Use Limitations. There should be specific consideration of limitations imposed by sharing protected health information and identifiable information. These are significant limitations and should be addressed directly, along with proposed strategies to mitigate these challenges.

Other Considerations Relevant to this DRAFT Policy Proposal

The draft policy alludes to the significance of the HIPAA Privacy Rule, but does not explicitly address the implications for data management and sharing. The HIPAA Privacy Rule identifies research as a public interest and benefit activity, and if such allowances are made, the Plan should include how to address protected health information. Otherwise, it may be easy to exclude research using protected health information from this policy.

An additional element and benefit of data sharing is the importance of open data to spark cross-disciplinary research. Members noted that the availability of valid data resources is a key strategy for bringing new investigators from other fields into health and health care research.

We believe the guidance should be prescriptive enough to provide funding panels clear insight into the likely success of the research project in terms of data sharing-- that is, the likelihood that the study's data could be used to validate or replicate study findings or be reused for further research. The NIH should make it clear to researchers that insufficient data management and sharing plans will affect the likelihood of receiving funding.

Finally, beyond and in addition to the whole of our comments above, we want to reemphasize the importance of providing centralized support – infrastructure, operational guidance, resources, logistical support and training – to the field in order to support meaningful implementation of this policy.

Conclusion

AcademyHealth appreciates the opportunity to provide additional input on the Draft Policy for Data Management and Sharing. We believe clarity and specificity will be paramount to a sound and effective policy – one that clearly communicates context, rationale, requirements, and expectations and advances progress toward the widespread and responsible sharing of data.

AcademyHealth consulted with a committee of members and thought leaders to offer a response to NIH's request for comments on their draft Policy for Data Management and Sharing. We thank and acknowledge Greg Downing for his valuable contributions and guidance as Chair of the Ad-Hoc Committee.

Submission ID: 1398

Date: 1/10/2020

Name: Tina Koplinski

Name of Organization: Versiti Wisconsin

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Data of Primary Interest - Other:

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Distinguish between published and un-published data: It is difficult to envision a small lab being able to make all of its unpublished scientific data broadly available in an interpretable fashion to members of the research community. As such, we would recommend that unpublished and published data be distinguished within the policy.

Section III: Scope:

Clarify if requirements apply to subrecipients and who is responsible for retaining data.

Section IV: Effective Date(s):

Section V: Requirements:

1. Better define the time frames. The stated purpose of the policy is to provide guidance to scientists on how their data should be managed and shared to ensure rigor, reproducibility, and that it can be shared with the broader public. While this is a worthy goal, the policy does not stipulate specific time-frames for data management and sharing. Specifically, how long should scientific data be retained, and how long should it be broadly accessible. Given the explosion in data generation by most NIH-funded, indefinite data retention and sharing is unlikely to be feasible. For smaller research laboratories this will likely represent a substantial, unfunded burden. We would recommend a defined period of time for which data should be retained and

be shared widely. This will maximize the utility of the data generated while limiting the burden for most laboratories/research universities.

2. Clarify Regulatory boundaries for Data transfer and sharing such as GDPR and Privacy Shield: Provide clarity on Data Sharing between international collaborators impacted by International Regulations

Section VI: Data Management and Sharing Plans:

1. Provide examples of how to properly share data. Given the myriad of types of data generated, how to generate and share these types of data are unlikely to be captured and easily shared using a single approach. Unlike next-generation sequencing (NGS) based data, the "output" of the vast majority of experiments is unlikely to be reduced to a single type of file which can be readily interpreted. As such, we would recommend the broadest possible view of how data can be shared, but also provide examples for NIH-funded investigators to use as a guide. For example, it may be appropriate to make data available upon request, so long as investigators agree to share data.

2. Address data management models that will allow for distributed storage using hosted services such as Amazon: The architecture of hosted services are highly distributed across various geographies, and "always on" data storage is more the norm. Provide guidance on data management and sharing using hosted services.

3. Include a risk-based model for data management and sharing: Including risk analysis based on a standard risk model in the plan will help clarify and ensure the proper security levels are included for data management and sharing, and ensure that small labs are not burdened with unnecessary data management costs.

4. Provide guidance on the management of structured vs. unstructured data: To allow for consistent management of data that are structured and unstructured or semi-structured data (Such as those stored in Hadoop or Document databases) and move such data without losing integrity, provide examples and proposed structures, data models and formats.

Section VII: Compliance and Enforcement:

Clarify how compliance will be monitored via the RPPR process.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1399

Date: 1/10/2020

Name: Susan Meyn

Name of Organization: Association of Biomolecular Resource Facilities (ABRF)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Broad scope of genomic, imaging and other basic biomedical research data

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: Chair, ABRF Committee for Core Rigor and Reproducibility

Domain of Research Most Important to You or Your Organization:

ABRF represents >700 members working within or in the support of >300 shared resource and research biotechnology laboratories.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The NIH draft policy for research data management and sharing reflects the importance of data stewardship, good research data management, FAIR principles, flexibility in Data Plans and the autonomy and protection of research participants. Highlighting the need to consider data preservation and sharing as part of the research process is critical to foster culture change. The draft lists an expectation of "timely" data sharing. This should be defined at the time of publication. Funding opportunities specifically designated to create a shared resource should specify a date by which data must be available even in the absence of a publication. The previous NIH policy clearly defines "timely" as "no later than the acceptance for publication of the main findings from the final data set." The relaxation of this existing requirement is not justified.

Section II: Definitions:

The Association of Biomolecular Resource Facilities (ABRF) appreciates the definition of scientific data. By supporting all findings contributing to a line of research inquiry, it enables researchers to combine data types to strengthen analyses, facilitates reuse of hard to generate data or data from limited sources, and accelerates ideas for future research inquiries. Central to sharing scientific data is the recognized need to make data as available as possible while

ensuring that the privacy and autonomy of research participants are respected, and that confidential/proprietary data are appropriately protected. The definitions should include definitions of FAIR data and the 15 FAIR principles (Wilkinson M. et al, SCIENTIFIC DATA 3:160018 | DOI: 10.1038/sdata.2016.18)

Section III: Scope:

The scope applies to all research funded or conducted by NIH that results in generation of scientific data, The scope should make clear that the policy continues to apply for scientific data produced by funding in whole or in part from NIH after the NIH funding period is over.

Section IV: Effective Date(s):

ABRF recommends the final NIH policy for research data management and sharing have a "no later than" date for implementation, ideally 12 months after issuance of the final policy. Many of the issues faced with research data management are related to the absence of an effective data management and sharing policy.

Section V: Requirements:

ABRF, a member society of the Federation of American Scientists and Experimental Biologists (FASEB), supports their recommendation for trans-NIH coordination of supplemental requests and listing ICO-specific requirements as part of centralized resources associated with the final data management and sharing policy to minimize confusion and administrative burden.

Section VI: Data Management and Sharing Plans:

ABRF notes that data management plans will still be largely freeform under the proposed policy. ABRF recommends the NIH Guidance highlight the need to address issues of data integrity and quality in Data Management Plans submitted to NIH, as many investigators are not aware of the scope and discipline specific requirements for true data integrity and durable quality. For example, the importance of metadata associated with highly dense file types. The metadata ensures the data can be both effectively shared, reanalyzed or merged with other complex data sets. While there are many advantages to a freeform approach, a more structured approach would help guide applicants towards creating a useful and sustainable Data Management Plan. An explicit template of required elements would be a useful resource in addition to the suggested elements provided in the supplemental material of the draft policy. This will also intersect with guidance for Allowable Costs, due to the need to budget for process and resources to ensure data integrity and quality. We propose that the allowed costs be more flexible. As proposed, they will cover some established repositories, but internal infrastructural support is not adequately included.

The NIH's draft guidance proposes to collect data management and sharing plans as part of Just-In-Time (JIT) documentation. Although JIT submission is offered to minimize administrative burden at the proposal stage for both the applicant and peer reviewers, and provide more

flexibility for grantees to make real-time updates to their plans, the inclusion of data management and sharing plans at an earlier stage in the process allows concurrent planning with experimental design and facilitates planning. Additionally, we consider it an oversight in the proposed draft that data is considered without methodology. We believe methodology should be treated as data, and, as such be standardized accordingly with the purpose of improving reproducibility.

ABRF requests NIH ensure that credit for data sharing is keeping pace with calls to increase access to data. Further development of the concept of and criteria for recognition of the contributions of data generators is timely and will propel data sharing for the advancement of science and ensures that the available data set follows FAIR Guiding Principles, which instruct that the data and metadata meet criteria of findability, accessibility, interoperability, and reusability are standard practice. Shared resources/core facilities are integral to effective data management, sharing and cost effectiveness. NIH Guidance should emphasize the importance of investigators working with these (often NIH-funded) resources not only to generate high quality data, but to also advise, guide, and in many cases appropriately manage data for, investigators to ensure long term data integrity and quality.

One area in the final policy that would benefit from further clarification is whether NIH will make data management and sharing plans or limited details about the plan to increase awareness of the work, particularly if the work leads to outputs other than publications. To truly fulfill the FAIR data principles, plans should be made publicly available; however, NIH is encouraged to engage with the stakeholder community to determine possible unintended consequences of this strategy.

Section VII: Compliance and Enforcement:

The strategy of making the data management and sharing plan a term and condition of the grant award demonstrates NIH's commitment to fostering a culture of data sharing among investigators and institutions supported by NIH funding and support. To assist in compliance and education, we propose the designation of Research Data Management officer at each NIH funded academic institution, trained in how to structure and maintain adherence to good data management under the new policy in addition to the proposed guidance by NIH provided during the regular reporting intervals.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

ABRF commends the inclusion of supplemental guidance to assist understanding of the desired elements of a data management and sharing plan. The proposed guidance offers investigators flexibility to adapt plans to their specific research needs. Additional guidance in the form of examples and/or case studies will help investigators understand how costs to support data management are appropriately classified as direct costs of research (as opposed to indirect or

infrastructure costs). For instance, the section Local Management Considerations is ambiguous with regard to the categories of costs allowable and whether they be considered infrastructure or overhead. "Unique and specialized information infrastructure necessary to provide local management, preservation, and access to data" may include one or more of the following: external hard drives; a large, off-site server that is backed-up to tape; expansion of existing departmental server space; purchase of commercial server space (e.g. Amazon, Google); high-speed optical fiber connections; etc.

The guidance should specify that fees that preserve data beyond the funding period are allowed, as are personnel expenses related to data sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

ABRF is supportive of including supplemental guidance defining possible allowable costs. However, many researchers may not be aware of the issues surrounding data management for many different experimental approaches. For instance in imaging, there is a substantial fraction of our NIH-funded faculty who are unaware of the metadata associated with image files, or that the information is lost when those files are converted to a different format. Therefore, those researchers may not be aware of how the discussion of "metadata" in the Guidance Elements Supplement would relate to their work. It would be helpful if the NIH would provide concrete examples (perhaps with help from the ABRF community) that could help researchers identify the issues related to the data they are generating as they prepare this document.

ABRF appreciates NIH's recognition of the costs associated with data management and sharing and applauds the inclusion of the supplemental guidance defining possible allowable costs. However, the guidance only addresses those costs incurred during the term of the award but does not address costs associated with long-term data retention and accessibility. ABRF requests that NIH consider the cost of data management over the lifetime of research records, which will often pre-date the NIH project budget period and may have the potential to extend long past the project end date. This would also intersect with the guidance for Data Management and Sharing Plans, in emphasizing the importance of continued data management after all data for a project are collected.

Other Considerations Relevant to this DRAFT Policy Proposal:

ABRF supports data management policies and deposition of data, metadata and methods to prevent digital meddling, either through repositories specific for techniques (i.e. <https://flowrepository.org>) or general data repositories sponsored by the NIH and Center for Open Science (<https://osf.io>). The shared goals of the NIH, other research stakeholders and research institutions are more likely to be achieved when shared resource core scientists and

research scientists work together to identify and minimize risk to research data, thereby improving research quality, rigor and reproducibility. At an institutional level, shared resource core laboratories/facilities generate the majority of research data at many institutions. Core science inherently supports transparency and scientific reproducibility through unbiased acquisition, minimizing interoperable variability and promoting transparent processes and reporting (detailed experimental materials and methods) for publications and grants. Data provenance is assured— detailing who performed what experiment on which instrument; instrument standardization and maintenance; QA/QC (required controls, standards, documentation and tracking of buffers, reagents, components, lot numbers, version, expiration dates); location of source data and shared data (curation in compliance with Data Storage Standards for Research Core Laboratories, OMB Circular A-110, NIH GDS Policy and FAIR Guiding Principles).

Finally, ABRF reiterates conclusions and recommendations from NOT-OD-16-091 gathered for NIH from shared resource core laboratories and stakeholders including:

1. Facilitate training and education of researchers, technical staff and data scientists on data management with respect to basic knowledge of data types, annotation methods, databases
2. Encourage intra-institutional and multi-institutional partnerships among cores and bioinformatics centers to promote standardization and use of best practices for data annotation and management
3. Support community development of guidelines for metadata standards, best practices in data annotation, and key elements for methods, as well as assisting with dissemination and implementation of these guidelines
4. In partnership with other funding agencies, provide guidance on data management (including data annotation) that would foster high-quality data, metadata and methods across scientific communities.

Attachment:

ABRF-Response_NIH RFI Data Management and Sharing_FIN01102020-submitted.pdf

Description:

ABRF response to NIH draft policy for data management and sharing

Response to RFI on DRAFT NIH Policy for Data Management and Sharing

Submitted on behalf of the Association of Biomolecular Resource Facilities (ABRF)

<https://abrf.org/>

Contact:

Susan Meyn

s.meyn@vumc.org

Chair, ABRF Committee for Core Rigor and Reproducibility

Section I: Purpose (limit: 8000 characters)

The NIH draft policy for research data management and sharing reflects the importance of data stewardship, good research data management, FAIR principles, flexibility in Data Plans and the autonomy and protection of research participants. Highlighting the need to consider data preservation and sharing as part of the research process is critical to foster culture change. The draft lists an expectation of “timely” data sharing. This should be defined at the time of publication. Funding opportunities specifically designated to create a shared resource should specify a date by which data must be available even in the absence of a publication. The previous NIH policy clearly defines “timely” as “no later than the acceptance for publication of the main findings from the final data set.” The relaxation of this existing requirement is not justified.

Section II: Definitions (limit: 8000 characters)

The Association of Biomolecular Resource Facilities (ABRF) appreciates the definition of scientific data. By supporting all findings contributing to a line of research inquiry, it enables researchers to combine data types to strengthen analyses, facilitates reuse of hard to generate data or data from limited sources, and accelerates ideas for future research inquiries. Central to sharing scientific data is the recognized need to make data as available as possible while ensuring that the privacy and autonomy of research participants are respected, and that confidential/proprietary data are appropriately protected. The definitions should include definitions of FAIR data and the 15 FAIR principles (Wilkinson M. et al, SCIENTIFIC DATA 3:160018 | DOI: 10.1038/sdata.2016.18)

Section III: Scope (limit: 8000 characters)

The scope applies to all research funded or conducted by NIH that results in generation of scientific data, The scope should make clear that the policy continues to apply for scientific data produced by funding in whole or in part from NIH after the NIH funding period is over.

Section IV: Effective Date(s) (limit: 8000 characters)

ABRF recommends the final NIH policy for research data management and sharing have a “no later than” date for implementation, ideally 12 months after issuance of the final policy. Many of

the issues faced with research data management are related to the absence of an effective data management and sharing policy.

Section V: Requirements (limit: 8000 characters)

ABRF, a member society of the Federation of American Scientists and Experimental Biologists (FASEB), supports their recommendation for trans-NIH coordination of supplemental requests and listing ICO-specific requirements as part of centralized resources associated with the final data management and sharing policy to minimize confusion and administrative burden.

Section VI: Data Management and Sharing Plans (limit: 8000 characters)

ABRF notes that data management plans will still be largely freeform under the proposed policy. ABRF recommends the NIH Guidance highlight the need to address issues of data integrity and quality in Data Management Plans submitted to NIH, as many investigators are not aware of the scope and discipline specific requirements for true data integrity and durable quality. For example, the importance of metadata associated with highly dense file types. The metadata ensures the data can be both effectively shared, reanalyzed or merged with other complex data sets. While there are many advantages to a freeform approach, a more structured approach would help guide applicants towards creating a useful and sustainable Data Management Plan. An explicit template of required elements would be a useful resource in addition to the suggested elements provided in the supplemental material of the draft policy. This will also intersect with guidance for Allowable Costs, due to the need to budget for process and resources to ensure data integrity and quality. We propose that the allowed costs be more flexible. As proposed, they will cover some established repositories, but internal infrastructural support is not adequately included.

The NIH's draft guidance proposes to collect data management and sharing plans as part of Just-In-Time (JIT) documentation. Although JIT submission is offered to minimize administrative burden at the proposal stage for both the applicant and peer reviewers, and provide more flexibility for grantees to make real-time updates to their plans, the inclusion of data management and sharing plans at an earlier stage in the process allows concurrent planning with experimental design and facilitates planning. Additionally, we consider it an oversight in the proposed draft that data is considered without methodology. We believe methodology should be treated as data, and, as such be standardized accordingly with the purpose of improving reproducibility.

ABRF requests NIH ensure that credit for data sharing is keeping pace with calls to increase access to data. Further development of the concept of and criteria for recognition of the contributions of data generators is timely and will propel data sharing for the advancement of science and ensures that the available data set follows FAIR Guiding Principles, which instruct that the data and metadata meet criteria of findability, accessibility, interoperability, and reusability are standard practice. Shared resources/core facilities are integral to effective data management, sharing and cost effectiveness. NIH Guidance should emphasize the importance of investigators working with these (often NIH-funded) resources not only to generate high quality data, but to also advise, guide, and in many cases appropriately manage data for, investigators to ensure long term data integrity and quality.

One area in the final policy that would benefit from further clarification is whether NIH will make data management and sharing plans or limited details about the plan to increase awareness of the work, particularly if the work leads to outputs other than publications. To truly fulfill the FAIR data principles, plans should be made publicly available; however, NIH is encouraged to engage with the stakeholder community to determine possible unintended consequences of this strategy.

Section VII: Compliance and Enforcement (limit: 8000 characters)

The strategy of making the data management and sharing plan a term and condition of the grant award demonstrates NIH's commitment to fostering a culture of data sharing among investigators and institutions supported by NIH funding and support. To assist in compliance and education, we propose the designation of Research Data Management officer at each NIH funded academic institution, trained in how to structure and maintain adherence to good data management under the new policy in addition to the proposed guidance by NIH provided during the regular reporting intervals.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing (limit: 8000 characters)

ABRF commends the inclusion of supplemental guidance to assist understanding of the desired elements of a data management and sharing plan. The proposed guidance offers investigators flexibility to adapt plans to their specific research needs. Additional guidance in the form of examples and/or case studies will help investigators understand how costs to support data management are appropriately classified as direct costs of research (as opposed to indirect or infrastructure costs). For instance, the section *Local Management Considerations* is ambiguous with regard to the categories of costs allowable and whether they be considered infrastructure or overhead. "Unique and specialized information infrastructure necessary to provide local management, preservation, and access to data" may include one or more of the following: external hard drives; a large, off-site server that is backed-up to tape; expansion of existing departmental server space; purchase of commercial server space (e.g. Amazon, Google); high-speed optical fiber connections; etc.

The guidance should specify that fees that preserve data beyond the funding period are allowed, as are personnel expenses related to data sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (limit: 8000 characters)

ABRF is supportive of including supplemental guidance defining possible allowable costs. However, many researchers may not be aware of the issues surrounding data management for many different experimental approaches. For instance in imaging, there is a substantial fraction of our NIH-funded faculty who are unaware of the metadata associated with image files, or that the information is lost when those files are converted to a different format. Therefore, those researchers may not be aware of how the discussion of "metadata" in the Guidance Elements Supplement would relate to their work. It would be helpful if the NIH would provide concrete examples (perhaps with help from the ABRF community) that could help researchers identify the issues related to the data they are generating as they prepare this document.

ABRF appreciates NIH's recognition of the costs associated with data management and sharing and applauds the inclusion of the supplemental guidance defining possible allowable costs. However, the guidance only addresses those costs incurred during the term of the award but does not address costs associated with long-term data retention and accessibility. ABRF requests that NIH consider the cost of data management over the lifetime of research records, which will often pre-date the NIH project budget period and may have the potential to extend long past the project end date. This would also intersect with the guidance for Data Management and Sharing Plans, in emphasizing the importance of continued data management after all data for a project are collected.

Other Considerations Relevant to this DRAFT Policy Proposal (limit: 8000 characters)

ABRF supports data management policies and deposition of data, metadata and methods to prevent digital meddling, either through repositories specific for techniques (i.e. <https://flowrepository.org>) or general data repositories sponsored by the NIH and Center for Open Science (<https://osf.io>). The shared goals of the NIH, other research stakeholders and research institutions are more likely to be achieved when shared resource core scientists and research scientists work together to identify and minimize risk to research data, thereby improving research quality, rigor and reproducibility. At an institutional level, shared resource core laboratories/facilities generate the majority of research data at many institutions. Core science inherently supports transparency and scientific reproducibility through unbiased acquisition, minimizing interoperable variability and promoting transparent processes and reporting (detailed experimental materials and methods) for publications and grants. Data provenance is assured— detailing who performed what experiment on which instrument; instrument standardization and maintenance; QA/QC (required controls, standards, documentation and tracking of buffers, reagents, components, lot numbers, version, expiration dates); location of source data and shared data (curation in compliance with Data Storage Standards for Research Core Laboratories, OMB Circular A-110, NIH GDS Policy and FAIR Guiding Principles).

Finally, ABRF reiterates conclusions and recommendations from **NOT-OD-16-091** gathered for NIH from shared resource core laboratories and stakeholders including:

1. Facilitate training and education of researchers, technical staff and data scientists on data management with respect to basic knowledge of data types, annotation methods, databases
2. Encourage intra-institutional and multi-institutional partnerships among cores and bioinformatics centers to promote standardization and use of best practices for data annotation and management
3. Support community development of guidelines for metadata standards, best practices in data annotation, and key elements for methods, as well as assisting with dissemination and implementation of these guidelines
4. In partnership with other funding agencies, provide guidance on data management (including data annotation) that would foster high-quality data, metadata and methods across scientific communities.

Submission ID: 1400

Date: 1/10/2020

Name: Heidi Imker

Name of Organization: University of Illinois at Urbana Champaign

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Multidisciplinary

Type of Organization: University

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

University has many strengths in NIH areas, including in genomics, cognitive neuroscience, etc. Our feedback represents all areas.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

No comments.

Section II: Definitions:

Data Sharing - it may be helpful to note here that sharing may be on a continuum between completely open for anyone to access to restricted sharing with approved parties (e.g. for some human subjects data). It also may be helpful to articulate if NIH believes some mechanism are not appropriate, for example, does NIH consider "on request" from author appropriate? Or is sharing within a journal article appropriate? Several surveys have shown that researchers think these are adequate mechanisms. If not, it should be clarified.

Scientific Data - per webinar and provided documents "NIH does not expect researchers to share all scientific data generated in a study." However, the use of the word "all" in the last sentence of this definition implies otherwise. Why digitize all if not expected to be shared all? This will create confusion and potentially waste resources. Consider this revision: "NIH expects that reasonable efforts will be made to digitize any scientific data of relevance to the community."

Given that data curation is an allowable cost, a definition of data curation should also be included.

Section III: Scope:

No comments.

Section IV: Effective Date(s):

Given costs for data management and sharing are allowable, effective dates should be staged to take into account the time needed to determine government costing rates.

Section V: Requirements:

For the purposes of this section, the brevity is appropriate. However, please see below for comments on restrictions/limitations and compliance.

Section VI: Data Management and Sharing Plans:

Paragraph 2 - we greatly appreciate the sensitivity to human subjects projection and are very happy to see the language in this paragraph.

Plan Assessment - it is a constant question how these plans will be evaluated. Please include a sentence on how NIH staff will be prepared to evaluate plans (e.g. training? checklist?).

Section VII: Compliance and Enforcement:

During the Funding or Support Period - we appreciate the clarity provided in this section that Plans will become part of Terms and Conditions and reviewed during RPPRs. Please see comment above, however. It is unclear what program staff will be looking for. For example, should a section in the RPPR on "Plan Compliance" be included? Would NIH staff compare the Plan in the Terms and Conditions with that section and if deviation is observed, a justification will be requested?

Post Funding or Support Period - NIH has acknowledged that the plans need to be flexible and responsive to the sometimes unpredictable outcomes of research. How would approval for Plan revisions be obtained after the funding period? If someone can be deemed non-compliant after a funding period concludes, there must be a mechanism for them to update the Plan post funding period, as well.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

While we greatly appreciate the acknowledgement that Data Management and Sharing costs are above and beyond, this section would benefit from concrete examples, especially for

expected F&A costs vs. allowable costs here categorized as "local data management considerations."

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The guidance provided in this document is excellent. We appreciate its flexible but helpful language, with clear acknowledgement of varying maturity in community standards and expectations, as well as opportunities to provide justifications when faced with limitations.

Other Considerations Relevant to this DRAFT Policy Proposal:

Please see attached letter.

Attachment:

VCRIWolinertNIHLtr1-10-20.pdf

Description:

supplemental letter emphasizing and summarizing critical points important to our university

January 10, 2020

Carrie D. Wolinetz, Ph.D.
 Associate Director and Acting Chief of Staff
 NIH Office of Science Policy
 National Institutes of Health
 6705 Rockledge Drive #750
 Bethesda, Maryland 20817

Dear Dr. Wolinetz:

Thank you for the opportunity to comment on proposed provisions for a draft Data Management and Sharing Policy for NIH-funded or supported research. The University of Illinois at Urbana-Champaign is committed to stewarding the data resulting from our federally funded research and making this data as available as possible while safeguarding the privacy of research participants and protecting confidential or proprietary data. We applaud NIH's initiative in putting forward the proposed provisions and the commitment to considering feedback from the community.

The policy and associated guidance documents are a marked improvement on the documents provided for the RFI in late 2018. We believe the more nuanced treatment will help our university staff and researchers understand the spirit of the new policy and appropriate actions to take.

In our feedback provided for the 2018 FI, we noted several critical recommendations that are especially important to us. These recommendations are also reflected in the [APLU-AAU Public Access Working Group Report and Recommendations](#). While we see improvements in all cases, we would like to offer updated comments for these recommendations, specifically:

- Data to be Shared - from the APLU-AAU report "Agencies should provide clear information on expectations regarding what data do and do not need to be shared"
 - We appreciate the clarity that not "all" data must be shared and that it is up to researcher to determine "which" data will be shared. This pragmatic approach is likely to reflect, in most cases, established community norms. However, we note that is not as likely to spur new community norms or be responsive to emerging norms. As such, we trust that NIH plans to accept highly variable interpretations of which data to share and let communities evolve at their own pace.
- Data Retention - from the APLU-AAU report "Agency expectations for data access after the funding period has ended should be specific and finite in duration..."
 - Retention expectations remain a common question with our researchers. We maintain that the policy should offer a minimum window (such as the 3 year retention period in OMB Circular A-110 as a *minimum* with a reference to HHS's RCR site: https://ori.hhs.gov/education/products/rcradmin/topics/data/tutorial_11.shtml), with strong wording that longer retention may be warranted based on community expectations and reuse potential.

- Plan Compliance - from the APLU-AAU report "Agencies should provide clear information on how compliance with data sharing requirements will be monitored, evaluated, and enforced ..."
 - We appreciate the significant improvement in regard to compliance. We ask that additional information on how NIH staff will be prepared to evaluate plans and the process of reviewing Terms and Conditions and RPPRs be provided. For example, should a section in the RPPR on "Plan Compliance" be included? Would NIH staff compare the Plan in the Terms and Conditions with that section and, if deviation is observed, a justification will be requested? Additionally, the documentation suggests that researchers can be found non-compliant post funding/support period. In this case, a mechanism for researchers to update Plans post funding period should be provided.

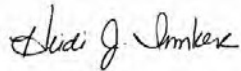
Finally, we also appreciate the acknowledgement that Data Management and Sharing costs are above and beyond. However, this section would benefit from concrete examples, especially for expected F&A costs vs. allowable costs here categorized as "local data management considerations."

We are very appreciative of NIH's consultative approach in developing this policy. Please let us know if we can be of further assistance as it is finalized.

Sincerely yours,



Susan A. Martinis
Stephen G. Sligar Endowed Professorship in the School of Molecular and Cellular Biology
Vice Chancellor for Research and Innovation



Heidi Imker
Associate Dean for Research
University Library

Submission ID: 1401

Date: 1/10/2020

Name: Robert R. Montgomery

Name of Organization: Blood Research Institute

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: genomic and clinical

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Comments on DRAFT NIH Policy for Data Management and Sharing

1) Better define the time frames. The stated purpose of the policy is to provide guidance to scientists on how their data should be managed and shared to ensure rigor, reproducibility, and that it can be shared with the broader public. While this is a worthy goal, the policy does not stipulate specific time-frames for data management and sharing. Specifically, how long should scientific data be retained, and how long should it be broadly accessible. Given the explosion in data generation by most NIH-funded, indefinite data retention and sharing is unlikely to be feasible. For smaller research laboratories this will likely represent a substantial, unfunded burden. We would recommend a defined period of time for which data should be retained and be shared widely. This will maximize the utility of the data generated while limiting the burden for most laboratories/research universities.

2) Provide examples of how to properly share data. Given the myriad of types of data generated, how to generate and share these types of data are unlikely to be captured and easily shared using a single approach. Unlike next-generation sequencing (NGS) based data, the "output" of the vast majority of experiments is unlikely to be reduced to a single type of file which can be readily interpreted. As such, we would recommend the broadest possible view of how data can be shared, but also provide examples for NIH-funded investigators to use as a

guide. For example, it may be appropriate to make data available upon request, so long as investigators agree to share data.

3) Distinguish between published and un-published data: It is difficult to envision a small lab being able to make all of its unpublished scientific data broadly available in an interpretable fashion to members of the research community. As such, we would recommend that unpublished and published data be distinguished within the policy.

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1402

Date: 1/10/2020

Name: Idan Gabdank

Name of Organization: Stanford

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Functional Genomics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

While I agree with most of the statements in this section, I think that instead of saying "data should be made accessible in a timely manner" it should be "data have to be made accessible at the time of publication".

Section II: Definitions:

I think the section should include FAIR principles.

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

In line with the inclusion of the FAIR principles in the Definitions section, researchers should be required to describe how they address different FAIR principles.

Section VI: Data Management and Sharing Plans:

I think the sentence " NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public" is a little vague and could be re-worded to be more definitive, something along the lines of "NIH requires scientific data to be shared".

Section VII: Compliance and Enforcement:

I think the sanctions listed in the draft are very weak and will be un-effective.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:**Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:**

I am not sure why "NIH does not expect researchers to share all scientific data generated in a study"? I think all the data should be required to be shared.

Other Considerations Relevant to this DRAFT Policy Proposal:

The draft policy's requirements are too vague and weak and will not serve their purpose without enforcement. I strongly believe that the scientific community as a whole will benefit from the enforcement of the policy, and that will educate the scientists to make their data findable, accessible, interoperable and reusable.

Attachment:**Description:**

Submission ID: 1403

Date: 1/10/2020

Name: Cole Allick (Turtle Mountain Band of Chippewa Indians), MHA

Name of Organization: Washington State University, Institute for Research and Education to Advance Community Health (IREACH) and Partnerships for Native Health (P4NH)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All research with Tribal Nations and urban Native populations

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Practice Based Research Network Coordinator/Tribal Liaison

Domain of Research Most Important to You or Your Organization:

Community Health Research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

In the background for this draft policy, NIH describes its past and future commitment to working with Federally acknowledged Tribal Nations (Tribal Nations) to develop culturally sensitive data management and sharing resources. Additionally, NIH recognizes that research conducted with Tribal Nations may have unique data sharing concerns and has consulted, and will continue to consult, with Tribal Nations. It is extremely important to continue this process to operationalize meaningful consultation and to ensure future policy decisions facilitate appropriate and effective data management and sharing strategies.

Existing literature suggests that Tribal Nations' acceptability of and participation in research projects depend, in large part, on the following considerations: 1) whether the data management and collection methods account for the cultural importance of the data collected (biological specimens, stories, etc.); 2) whether tribal perspectives inform the analysis and interpretation of data; 3) whether security measures and conditions of storage provide adequate protections against loss of privacy and subsequent harm; 4) whether community control of data extends to new proposals and/or dissemination of research; 5) whether the terms of data withdrawal and disposal fully account for participant- and community-specific

cultural beliefs; and 6) to what extent community engagement in decisions and activities relates to each of these data management components. In this way, community engagement in the creation and governance of research can reasonably be thought of as a necessary prerequisite for developing data management practices and policies that respect the legal rights and interests of Tribal Nations by allowing them to set the terms of data ownership, control, access, and possession in research.

Tribal Nations, as experts on their communities and culture, employ a spectrum of strategies to exert their sovereignty regarding regulation of research; some may have tribal IRBs, research review boards, cultural committees, or other research groups while others may not. Nonetheless, tribal communities are rich in history and knowledge about the types of research that have been conducted and that are appropriate for their members. Tribal Nations and their reviewing bodies, whatever they might be, are also keenly attuned to protecting their people not only from individual harm – the purview of most academic IRBs – but also from community harm. These dual concerns coincide directly with the general purpose of this draft proposal. The unique status of Tribal Nations creates an opportunity for Indian Country and NIH to begin to operationalize the creation of a Data Management and Sharing Plan that reflects a joint interest in safeguarding tribal sovereignty while promoting culturally and locally appropriate research that improves the health of Native people.

Section II: Definitions:

No comment.

Section III: Scope:

No comment.

Section IV: Effective Date(s):

No comment.

Section V: Requirements:

This section may provide an opportunity for the funding NIH ICO to request an assurance that the proper tribal entities were consulted in the formulation of the Data Management and Sharing Plan if it has not already been explicitly stated. Tribal Nations involved in research should be active participants in data management, sharing, and ownership efforts. For example, IREACH ensures that data sharing is done with tribal approval – sometimes at the beginning of the study, but just as often, later for secondary use. Researchers who request data held at IREACH for secondary use have to go through an application process with an internal committee (the Publications and Presentations Committee); part of the process ensures that all Tribal Nations and entities that approved the original research also approve the secondary use. Researchers are responsible for obtaining these approvals before data can be released to them.

A description of a similar process could be included in the plan for research done with Tribal Nations. This may be a natural evolution of the continued work and consultation with Tribal Nations but could be a supplemental resource that can be developed for future researchers.

Section VI: Data Management and Sharing Plans:

We appreciate the proposed flexibility that would allow the Plan to be developed and/or revised as the funded research study evolves. This is highlighted in the draft policy language on applicable tribal law regarding human participants and in giving Tribal Nations the authority to govern research conducted on their lands and with their citizens living on tribal lands.

Section VII: Compliance and Enforcement:

As indicated in the draft policy, reviewers will conduct annual evaluations, at a minimum. While the draft policy language appears to recognize the sovereignty of tribes, it should also require the use of a "Tribal Nation lens" when reviewing research that includes Native Americans. A best practice of the Washington State Department of Health is the inclusion of a tribal liaison who works directly with Department staff and tribes. This common thread allows for a general Department of Health decision to be viewed through a "Tribal Nation lens" and considers how the decision will impact tribal communities. While this may not be a feasible option, at least operationally, in an annual grant review, it is crucial to operate with the understanding that Tribal Nation projects require a rigor that other projects may not necessarily need. This may be another opportunity for Tribal Nations to collaborate with the NIH Tribal Health Research Office to find a best practice that works for all parties involved.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

While Tribal Nations have expertise and solutions germane to the issues presented in the research studies, there are nearly 600 sovereign Tribal Nations across the United States, each with its own history, practices, and solutions. As such, Tribal Nations exist on a spectrum of ability to self-govern, due to the administrative and financial resources necessary to successfully operate an independent nation. In terms of allowable costs as they relate to data management and sharing – something at the core of tribal sovereignty – NIH should carefully consider some of the unique opportunities it has to engage with Tribal Nations that wish to rightfully govern this matter, and work closely with these Nations to find solutions that may not fit within the already robust language provided in this section of the policy.

For those researchers (including those at IREACH) who wish to work with Tribal Nations in a secondary data analysis, additional costs are incurred for obtaining approvals from the Tribal Nation; such costs are not otherwise common in research that was not conducted with Tribal Nations. These secondary approvals by Tribal Nations are part of a culturally appropriate data management and sharing plan, and apply to both Native and non-Native researchers wishing to utilize Tribal Nation data for a secondary data analysis.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The rigor and structure of this section as it relates to specific data elements suggested in the NIH Data Management and Sharing Plan is paramount to Tribal Nations in providing a starting guide. In particular, the inclusion of applicable tribal laws, regulations, statutes, guidance, and institutional policies allows Tribal Nations to exert their sovereignty while also using the tool as a guide for the elements needed for a well-rounded and thoughtful plan.

As mentioned previously, the existing literature presents a spectrum of data management and community involvement in research studies conducted with Tribal Nations. This specific guidance gives Tribal Nations and their partners a roadmap for a robust proposal. As NIH continues its commitment to working with Tribal Nations, we strongly encourage NIH and its partners to think carefully about additional guidance as it relates specifically to those projects across Indian Country. This careful consideration will benefit Tribal Nations who wish to apply for NIH projects and will likewise benefit NIH as it funds and regulates these projects.

Other Considerations Relevant to this DRAFT Policy Proposal:**Attachment:****Description:**

Submission ID: 1405

Date: 1/10/2020

Name: Diane Lehman Wilson

Name of Organization: University of Michigan Medical School

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All types of data are of interest to the Medical School

Type of Organization: Other

Type of Organization - Other: Medical School

Role: Other

Role - Other: Regulatory Manager

Domain of Research Most Important to You or Your Organization:

All areas of medicine, public health, and behavior may be of interest to the Medical School.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We fully endorse the broad purpose of the policy to make NIH funded research broadly available to the public. We would simply recommend that the first sentence be supplemented as follows to make clear from the outset that data sharing, while hugely important, never supersedes the concerns for human dignity and privacy. "The NIH Policy for Data Management and Sharing (herein referred to as the Policy) reinforces NIH's longstanding commitment to making the results and outputs of the research that it funds and conducts available to the public to the extent possible consistent with legal and ethical considerations for research participant privacy."

Implementing FAIR (Findable, Accessible, Interoperable, and Reusable) data principles requires collaboration, education, and standardization. Interoperability and reusability are particularly challenging and labor intensive. This is a global endeavor to advance the way science is conducted, and we applaud the NIH for trying to lead the way in this effort.

Given the breadth of research that NIH sponsors, we can understand that the policy itself may not discuss which standards, technical tools, etc. to utilize, but without more detail than this draft contains or contemporaneous guidance from NIH or from the ICOs, it will be difficult for even the most conscientious researchers and institutions to know if they are "hitting the mark". Without standardization, there is a lot of room for inefficiency in use of the data; misuse of the data; or misinterpretation of the data. Some fields may have much more developed standards

than others, so it may be challenging for the research community to guess how NIH is going to determine whether or not the way we define and share our data is going to be helpful to other institutions.

Thus, while we commend the efforts NIH has made within this draft, to balance intentionality and expectations with leeway and variability to account for the nearly infinite variety of circumstances in which data that may be referenced in research proposals will be generated, we encourage still more development of both of these approaches, along with additional resources and guidance documents. We believe that providing great clarity with regard to required minimum standards, and clear encouragement or announcement of future levels of heightened minimum standards will best allow institutions, large and small, to put their most efficient institutional effort into improving data sharing, while continuing to allow scientific researchers to focus primarily on their areas of expertise.

Section II: Definitions:

We appreciate having a definitional set from which to start. But we are concerned that in an effort to not make the policy overly burdensome, certain fairly well established elements of some of these terms have been omitted or specifically excluded. We believe that taking common vocabulary and constraining it into terms of art risks greater misunderstanding.

Regarding "Data Management": Data identification/definition and collection are important parts of data management. We recommend the following definition: "The process of identifying, collecting, reviewing (validating), securely storing and delivering data to the intended users, using processes that promote data integrity, quality, and reliability." Perhaps further discussion should be included to indicate that in an ideal state of significant data sharing, the distinction between "intended users" and "a broad public scientific audience" might ultimately disappear. And if the desired Data Management plans do not need to include specific sections about the identification and collection of the data, then that narrowing could be done within the term, "Data Management and Sharing Plan" by modifying it to "NIH-mandated Data Management and Sharing Plans" and then specifying that NIH-mandated Data Management and Sharing plans would need to include only the following elements: ...

Regarding Metadata: We encourage NIH to consider separating "supporting documentation" for items like the protocol, variable descriptions, etc. from metadata that is more within the category of digital requirements, field requirements (e.g., # of digits) and information about coding or software that would be essential to using the data.

We are still more concerned with the definition of Scientific Data. It appears that the use of the term scientific data as the operable term for definition is trying to meet opposing goals unsuccessfully. It seems to bely common sense understandings of the term, in order to have the policy sound very broad while not changing the norms of data sharing too dramatically at one time. It would be better to keep the terms more consistent with scientific norms. Specifically, one would normally consider scientific data to include laboratory notebooks and

completed case report forms, as those are the ultimate source documents upon which the validation of research findings would depend. Indeed, it's hard to imagine that the FDA would endorse the notion that Case Report Forms are not scientific data. Therefore, we believe it would be preferable to define a narrower term to which the data sharing expectations apply. In the 2003 policy, the term "Final Research Data" was used – and might be re-used here effectively, with minor definitional modification if needed. Alternatively, a new term such as Expected Sharable Scientific Data could be useful to clarify that this is defining the limits and expectations around data sharing under the policy, and to specify that digitization expectations do not necessarily apply to case report forms and laboratory notebooks.

Given the caveat, "regardless of whether the data are used to support scholarly publications," it becomes difficult to determine what "research findings" mean. Does that mean any interim determinations or hypotheses upon which the researcher chooses to rely? We believe it would make more sense to say something like, "research findings, whether published, presented to a community or audience, or shared within a report back to NIH, subsequent to final submission to NIH."

It would be helpful to clarify what is meant by "reasonable efforts to digitize all scientific data". Is the reasonableness standard a function of the grant size, the data size, the institutional infrastructure available for such digitization? Further, if digitization expectations will become a part of this policy, as proposed, standards for the attributability, security, and traceability of that digital information should be presented within the policy. Additionally, even some digital data may need to go through a transformation process to be usable by others. In some cases, intellectual property or proprietary limitations may constrain how certain kinds of data (e.g., instrument specific) can be shared.

Section III: Scope:

We believe it would be helpful for the scope section to explicitly confirm that it does NOT include research that simply indirectly or peripherally takes advantage of infrastructure which is partially funded by NIH.

Section IV: Effective Date(s):

For this policy to effect the significant change in culture and practice that it envisions above and beyond present application of the 2003 policy, substantial change in education and infrastructure needs to occur nationwide. This effort is not as straightforward as it might appear. Indeed, at the Academies of Medicine meeting in November (<http://www.nationalacademies.org/hmd/Activities/Research/DrugForum/2019-Nov-18.aspx>) several speakers, including Jeffrey Drazen of the New England Journal of Medicine, who years ago were "all for" increased broad requirements for data sharing, addressed the point that effective and meaningful data sharing is far more difficult than expected, that a staged or tiered approach to different sorts of data sharing based on the size, scale, and immediacy of clinical

implementation may make more sense than an all-in approach, especially for clinical trials. And the requirement to be good stewards of taxpayer dollars requires not only that data be reusable but also that the effort involved to make it reusable be proportionate to the value that the data may contribute.

Based on other change management efforts in which the University of Michigan Medical School has engaged for data management, it is not an exaggeration to state that it takes a large research institution two to four years to go from scoping and planning necessary purchases through training and finally to implementation of data management changes. Thus we would request a two to four year effective date with an additional period for enforcement activities. Smaller institutions may be even more challenged to comply with this policy. To avoid winner-take-all effects, it may behoove NIH to give even more time for knowledge bases and practices to spread, or to have some other sliding time frames.

We appreciate the statement built into the definition of scientific data that the policy does not seek sharing ALL scientific data. We would encourage additional language to acknowledge and support step-wise improvement over time. Indeed, some sort of planned cyclical review on the part of NIH might be an efficient means to adapt to new information about the effectiveness and the security of the data sharing that will occur. If such a plan were announced, provisions that might seem too much to require at present could be mentioned in the present policy as aims or intentions for the following period (e.g., 5 years hence), so that the community could work toward the necessary preparations, but if the data preservation world changed so dramatically within that period that those provisions needed to be changed (in any direction), there would opportunity to do so.

Section V: Requirements:

We appreciate that the policy is drafted in such a way as to accommodate variety between different types of research funded by different ICOs. We also greatly appreciate the specific recognition of costs of data management, preservation and data sharing as allowable – but we would ask that the data preservation costs be included for whatever time frame would be a "standard" base – whether that is 5 or 10 years. Further, we would like to flag that these costs might in some instances not be in the form of fees, but of partial salaries for support personnel in one fashion or another.

Section VI: Data Management and Sharing Plans:

We appreciate that the draft policy acknowledges that plans may need to change throughout the life of a grant for a variety of reasons. We appreciate that the draft policy references not only the ultimate sharing, but also the ongoing management of the data and specifies that plans must explain how data will be managed. We note that within the guidance about the data sharing plans there is reference to specifying who will do this data management. We note here, as well as in our comments to the guidance, that the amount of explanation and cataloguing expected for a complex study may not be able to be represented accurately enough

in 2- page plans except in broad language, even for grants that only cover one type of research with a predetermined set of data sources.

We greatly appreciate the decision to have data sharing plans become a part of the Just-in-Time submission to avoid a lot of detailed work for many projects that may never get funded. However, given the seasonality to some grant cycles, the resource constraints for an institution to effectively assist scores of researchers in a short window of time with thinking through the data architecture and data carpentry, as well as the legal, ethical, and technological constraints, the repository options etc. that may exist in some cases but not in others, may be extreme. It would be helpful if there were some alternative surrogate approach in certain instances for a simple "pledge" to meet certain minimum standards or to follow at least a basic NIH-created template with a subsequent submission of the full data plan. This might apply to very small scale pilot studies or to proposals whose first aim doesn't even include creation of any sort of large data base (for example behavioral studies whose first aim is to develop a tool for something or focus groups or identifying local partners and interviewing them – where data architecture will need to come soon, but not until after question or tool validation). A consultation with some of our grant specialists revealed that at present some grants are allowed a six month window in which to submit a data sharing plan. This sort of flexibility may still be wise, at least in certain domains.

Further, with grants that have multiple aims, some of which may generate very different data from others, the forecasting that would be necessary in the beginning to foresee all the types of data involved could be especially challenging. Therefore, it may be appropriate to create a specific method for mid-grant specifications for data management plan refinement – or to have separate plans drafted at the beginning for separate stages of the grant, with the first stage needing full detail, but the subsequent stages being only in broad strokes, to be refined once the data is better defined after the first stage is nearly completed and the second one is just ahead.

We also appreciate the acknowledgement in the draft policy that there are multiple intersecting requirements and responsibilities for human subject data. Academic medical centers and universities are working hard to develop systems and processes for reviewing and checking on de-identification – but these are challenging in a constantly changing data environment. With the growing "internet of things" and use of "wearables" in trials, the risk of reidentification based on data patterns alone continues to multiply and grow. We appreciate the acknowledgment that there may be instances where the ability to preserve or share data is limited by legal, ethical, and technical factors.

It therefore seems incongruous to conclude that paragraph with "NIH encourages the use of established repositories for preserving and sharing scientific data." This seems like a more appropriate opening statement prior to discussing human subjects issues which might argue against using repositories, at least open access ones, or having a clause within the sentence that acknowledges the concern for reidentification when combining large human data sets.

Where data being collected matches previously defined terms precisely, repositories can be enormously efficient, but when research is truly ground breaking, it may be hard to clarify what exactly is meant by a specific term – until someone else looks at it and raises a question. While we appreciate the benefits of established repositories, their use may pose more potential risks of reidentification than benefits, particularly for small scale studies that occur in a single environment and which may be only piloting concepts to allow for future development and testing; in that context, repositories may not be an efficient or safe use of resources. For these smaller scale pilot studies in particular, a simpler expectation of having a mechanism for data sharing and review upon reasonable request subject to a two party data use agreement may be more effective, secure, and efficient, as suggested by speakers at the National Academy of Medicine workshop in November.

(<http://www.nationalacademies.org/hmd/Activities/Research/DrugForum/2019-Nov-18.aspx>)

This allows researchers doing exploratory work to pay more attention to their project development without an outsized burden of trying to fit their early work into square data holes, especially in the most newly developing fields where terminology and standards may be less well established. It also encourages collaborative work in any subsequent re-analysis, which may go farther than someone pulling the data from a repository might achieve. In collaboration, these clarifications can occur naturally, but using a repository, the possibility of misunderstanding a term and therefore misconstruing the data is heightened. (One analogy to this is the use of the term sponsor within the federal government in the grant and drug development area. In one context it means funder, whereas in another it means the author of a protocol, and in a third it may mean the grantee of a grant, who allows a researcher to carry a protocol forward.)

Further clarification of what NIH means by "established repositories", either by listing criteria or having a website in which they will be referenced, would be helpful. It would also be helpful for NIH to provide guidance in the policy or in guidance documents about who would make determinations, or upon what criteria the determination could be made, that data are "no longer useful to the research community or the public". While at one level such determinations are necessary, one would get very different answers from consulting with people of different generations or professions; "history of science" scholars might even find data sets more useful and significant precisely because someone else felt they were no longer useful!

Finally, although beyond the scope of the Draft policy itself, even as it has been helpful for NIH to establish certain repositories in which it wishes parties to deposit data, and it may be helpful for it to do this more broadly, it would be helpful if NIH were to create and maintain an easily searchable database of at least the legal and ethical precedents that limit human subject data sharing. In particular, it would be helpful to have a centralized resource and analysis of concerns about the intersection of GDPR (which applies not only to European citizens but to

Americans whose data may be collected while they are in Europe), tribal concerns, and newly or future arising state law human subject data protections with NIH data sharing goals, with checklists and/or decision trees for researchers to use to determine if these constraints apply – and what the best alternative approaches to data sharing might be.

Section VII: Compliance and Enforcement:

It would be helpful to give examples of how the NIH would seek to monitor compliance with the management plan, what evidence or proofs would be sought, especially during the funding period. Under many circumstances the data may not be ready to be shared until very late in, or after, the funding period is over. So we fully appreciate that statements about non-compliance after the funding period are necessary and appropriate. That said, we would request that clarification be made regarding for what periods these issues "may be taken into account" and that those be directly related to the plans themselves. In other words, in those instances where a grantee has placed the data in a repository in accordance with a plan, subsequent actions by the repository, even if in conflict with the plan, should not redound to the harm of the institution, unless the repository was controlled by the same institution and knew of the conflict with the plan.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We greatly appreciate the recognition that the data sharing accessibility and reusability expectations will require costs above and beyond the (present and historic) routine costs of conducting research. We do question whether, indeed, this calls into question the wisdom of applying this policy across the board to all NIH funded research, regardless of project or sample size. Thus, in the case of small, trickle-down grants, issued through a CTSA for a pilot project, the funds involved can only be parsed so many ways; and if data management, curation, and deposition fees amount to \$10,000, there might be no money left to do the pilot project itself. We have already seen instances where the NIH policy regarding ClinicalTrials.gov data sharing has had a chilling effect in small pilot projects.

Regarding, "Reasonable costs regarding curation and developing supporting documentation, and data deposit fees, local data management considerations, separate from infrastructure costs typically included in overhead", our biggest concern is that these costs are extremely hard to estimate as a portion of them are contingent on data set size and the actual lived success of gathering the data, which by definition are in the future. Particularly in the case of multi-part grants, this difficulty will be intense. Conversely, in the case of very small pilot trials, the fixed costs of establishing the data base structure and the fees for deposition may be outsized to the grant awards.

Additionally, it would be helpful to have clarification of what, if any, components of data management are considered "routine costs of conducting research" beyond data access fees. Data management is not yet universally recognized as a profession. To implement and maintain a comprehensive data management plan for each NIH-funded project would require a

significant investment by most universities proportional to their grant budget to provide sufficient training and staff dedicated to this work. This could not be done overnight. Institutions need time to introduce data management as a specific and necessary study team role, create infrastructure, recruit, hire, and train data managers. This will take time and money.

Data security costs are not mentioned, and yet data security and who will provide it are considered elements of a data management plan in the other guidance. While at one level institutional data security is a part of ordinary operating expenses, if a description and discussion of the security for a given project is to represent anything more than institution-wide boilerplate, this will take specific time and attention, and therefore should be an allowable cost. Indeed, with many different forms of data collection, which may use many different types of software and hardware, having a data security review could be useful – but would be expensive. NIH should consider what is a reasonable approach to this dilemma and provide funding commensurate with the position it takes.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

While we greatly appreciate the utility of this draft guidance, we are concerned that for many projects, there may be a tension between fitting a plan into two pages, and covering the elements listed in the draft guidance with enough specificity to satisfy reviewers. Further – there are multiple levels of regulatory and technological sophistication required to implement this data management and sharing and to craft it into data sharing plans. While grants that relate to repositories that their ICO has already mandated may find this task relatively straightforward, large swaths of research will still be charting relatively new territory. For them, each grantee institution will need to decide whether it is going to try to operate on a centralized model, a semi-centralized model, or a decentralized model of adhering to these goals, or some mixture. These in turn may vary based on the nature of the data being gathered; the constraints and oversight required for human subject data is obviously greater than that of most other data – but some very sensitive biohazard or dual use data might also face constraints about sharing that require regulatory oversight.

The goals of the Elements of A NIH Data Management and Sharing Plan guidance are useful and lofty, but some of them, such as common data elements, remain for many areas of research still largely undeveloped and therefore aspirational. Even the relatively small number of data management professionals that academic medical centers and universities do have, have little time to "develop data standards" or maintain them. They will need to have serious staff time allocated to consulting with and learning from repositories for new norms to develop. New data types will be added continually to standards and thus, they will change over time. This is yet another reason why implementing the policy in an iterative, stepwise fashion may make sense.

The guidance (and indeed the policy) also speaks surprisingly little about the data quality and checking for accuracy which are the traditional domains of a data management plan. Without these practices and assurances, placing data in repositories potentially risks expanding problems rather than solutions.

Finally, the question of identifying who will be doing the data management may be nearly impossible, either if staff have to be hired to perform these functions, or if a service will be used (unless NIH policy clarifies that an institutional resource can be identified as the "who").

Other Considerations Relevant to this DRAFT Policy Proposal:

Once again, we support the NIH's efforts to shepherd tax dollars wisely by trying to get the most value out of government funded research. The speed at which technology, cyber-security, and social norms are changing legitimates creating sections within such a policy that establish clear and precise new floors, that are achievable in the present, and goals or encouragements that no less clearly point to potential future requirements. This entire area requires new forms and levels of collaboration with specialists who did not exist, or were peripheral to the scientific enterprise, when nearly half of the faculty at a medical center earned their tenure: data architects and carpenters, informationists, regulatory specialists, much less grant managers to help with Just-in-Time additions to proposals. For NIH staff and reviewers, as for researchers, there is a brave new world ahead. Some institutes have, for understandable reasons, blazed trails in the areas of common data elements, federally hosted repositories, and the like. Other areas may be far more diffuse in their expectations. To design a policy with clarity and latitude for these differences is not easy. We hope in that in taking this effort forward, NIH will consider with great care the infrastructure internal to NIH that will be necessary to facilitate these bold steps forward nationwide, as well as the additional levels of infrastructure that medical centers, schools, and universities will need to develop to follow that path. The more simplicity, clarity, and centralized resources to find and understand expectations that the NIH can provide, the easier it will be for researchers and institutions to get on board. We believe that a thoughtful step-wise approach, with centralized web-accessible databases of templates, checklists or decision trees, and lists of validated repositories and recommended practices will be the most efficient way to climb this mountain. Thank you for seeking widespread input to this draft.

Attachment:

Description:

Submission ID: 1406

Date: 1/10/2020

Name: Dee Dee Aubourg

Name of Organization: Acumen, LLC

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Healthcare Administrative Records

Type of Organization: Other

Type of Organization - Other: Healthcare Policy Government Contractor

Role: Other

Role - Other: MedRIC Project Manager

Domain of Research Most Important to You or Your Organization:

Healthcare and Social Policy, Drug and Vaccine Safety, Health Insurance Markets, Value-Based Purchasing, Healthcare Fraud

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Introduction

Acumen, LLC (Acumen) leverages vast data resources to improve information available to policymakers on topics such as epidemiological studies of drug and vaccine safety; health insurance markets; value-based purchasing; and healthcare fraud. As the recipient of multiple NIH Small Business Innovation Research (SBIR) grants as well as an NIH contract (No. HHSN271201500131U), our firm understands the criticality of data management and sharing to NIH's mission 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" (NIH.gov). In support of this mission, our firm has used NIH funds to establish the Medicare & Medicaid Resource Information Center (MedRIC), which provides the Centers for Medicare & Medicaid Services (CMS) data of NIH-sponsored survey participants to NIH-sponsored researchers for the study of aging, health care, and health outcomes. More specifically, MedRIC develops research-oriented versions of NIH-sponsored survey participants' Medicare and/or Medicaid records; creates linking crosswalks that enable researchers to effectively combine CMS data with NIH-sponsored survey data; and builds analytic tools—such as a remote-access data enclave—that improve researchers' ability to securely analyze sensitive CMS and NIH-sponsored survey data. This work requires our group and our NIH-sponsored researchers to produce standardized data management and sharing plans for both NIH and

CMS. In performing this work over the past 13 years, we found that researchers often lack the technical and security expertise to successfully produce such plans on their own, creating substantial burdens and costs for both the researchers that must compose these plans as well as the government offices that must certify them.

To reduce these burdens and costs, we recommend that NIH make the following four revisions to Section I. Purpose: (1) expand the definition of scientific data to include third party-owned data and their associated data sharing constraints; (2) establish a Data Management and Sharing Plan ("Plan") template file to improve Plan compliance and lower Plan costs; (3) integrate opportunities to consult with Trusted Third Parties (TTPs) about "[managing] scientific data resulting from NIH-funded or conducted research and prospectively [planning] for which scientific data will be preserved and shared" (Section I. Purpose), so that Policy compliance is feasible for all types of researchers; and (4) clarify two key phrases in Section I to enhance researchers' understanding of NIH's objectives.

Expand Scientific Data to Account for Third Party-Owned Data Sharing Constraints

We urge NIH to explicitly address third party-owned data in its Purpose statement because access terms for these data often complicate the data sharing that NIH's Policy mandates. For example, CMS requires DUA-authorized researchers to destroy all CMS data within 30 days of completing their study, but permits researchers to retain and reuse CMS data for other studies under a reuse request protocol. Convolved policies such as these make it impossible for researchers to meet NIH's objectives without additional guidance. For this reason, we recommend that NIH nuance its description of scientific data to account for data sets that either cannot be shared or have sharing limitations. More specifically, we suggest integrating access restriction language currently in the "Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan" ("Guidance") into Section I's references to scientific data.

Establish a Data Management and Sharing Plan Template File

We recommend creating a Plan template file—in addition to the Guidance—to give researchers a standardized framework through which they can more effectively and quickly compose their Plans, as well as to reduce Plan variations that drive up NIH review costs. To keep template creation costs low, NIH need only specify section headings and section instructions. In return, the template will reduce variations in Plan structure and content; lower the chances of non-compliant Plans; and decrease NIH review costs.

Integrate TTPs into Researchers' Data Management and Sharing Responsibilities

Integrating TTPs into NIH's description of researchers' data management and sharing responsibilities would offer critical support resources to researchers, without absolving researchers of their duty to carefully plan for and implement Policy-compliant data management and sharing. Indeed, identifying TTPs as a resource for helping researchers to comply with Policy terms would expand the number and type of researchers who propose

research projects to NIH, as researchers who lack substantial data management and sharing expertise would then have a means, via the TTP, to develop sound data management and sharing plans as part of their proposals. Even better, the TTP would be able to help researchers implement effective data management and sharing policies during their grant/contract award periods, providing ongoing and valuable help to researchers and NIH. Given these benefits, we think NIH should explicitly reference TTPs in two ways within Section I: (1) as a Policy support resource; and (2) when needed, as a provider of data management and sharing systems (such as remote-access data enclaves).

Relative to Policy support, a TTP could serve as a data management and sharing consultant to researchers, responding to their data management and sharing questions as well as recommending data management and sharing techniques. By doing so, a TTP would accelerate researchers' ability to produce Policy-compliant plans, reduce NIH costs in enforcing Policy terms, and lower the likelihood of researcher errors, or even breaches, in the management and sharing of their projects' scientific data. For these reasons, we think NIH should explicitly recognize and account for TTPs in its discussion of data management and sharing.

In addition to Policy support, a TTP could provide secure data enclaves to researchers, when those researchers lacked the technical and/or financial means to establish such environments on their own. This option is particularly critical to researchers who need sensitive federal information for their research projects (such as Medicare or Medicaid claims data) because this information requires the construction of complex, Federal Information Security Modernization Act (FISMA)-compliant systems; the implementation of high-cost FISMA procedures; and the acquisition of federal Authority to Operate (ATO). For researchers with limited budgets or know-how, the development of such systems is simply infeasible, making a TTP-provided enclave a burden-reducing option. This option could be critical in propelling researchers to apply for and comply with NIH grants/contracts. For that reason, we believe that NIH should explicitly discuss TTPs as secure data enclave providers in this section.

Clarify Two Key Phrases

Finally, we recommend that NIH clarify two key phrases in Section I—namely, (1)"results and outputs"; and (2)"made accessible in a timely manner." Relative to (1), NIH should clarify what qualifies as results and outputs by providing examples of each term. We also think it would be helpful to define what does not qualify as results and outputs, akin to how the Policy defines scientific data in Section II. Per (2), we think"timely" is too ambiguous in meaning for most researchers, as subjective interpretations of this term could result in Policy non-compliance issues. We therefore recommend setting a timeframe for data sharing (e.g., up to one year after the conclusion of a study) and a policy for requesting additional time, as we believe this level of specificity would ensure that researchers do not lag on their data sharing responsibilities.

Section II: Definitions:

We recommend clarifying three key terms in Section II. Definitions—namely, "Data Sharing," "Metadata," and "Scientific Data."

For "Data Sharing," we recommend adding "subject to data access authorizations and privacy protection requirements" to the end of the definition. Doing so will highlight the constraints on data sharing that researchers must account for in their Data Management and Sharing Plans. More critically, the clause will better balance NIH's underlying definition of scientific data between projects that create new data and those that use existing data—most of which have access authorization terms.

We think NIH should augment the definition of "Metadata" by adding "variable construction methods" and "analytic programs" to the list of metadata examples. Both of these materials provide researchers with crucial information on scientific data structuring and the methods by which a research team came to that structuring—key inputs to replication and validation studies. More specifically, these examples will remind researchers to account for the data manipulation and/or transformation work they have performed on raw data to produce their final scientific data set(s), as opposed to limiting their view of metadata to variable definitions. In other words, these examples will drive researchers to describe how they produced scientific data, not just what their scientific data is.

Likewise, for "Scientific Data," we recommend adding "methodologies for variable construction" to the definition, as these methods prove critical to data replication work (as explained above). We also recommend that NIH clarify "recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings" by providing specific examples of such material. These examples will give researchers a better sense of what these material consist of, as opposed to the list of things that the material is not.

Section III: Scope:

Given that the "DRAFT NIH Policy on Data Management and Sharing" ("Policy") will apply to thousands of NIH grants and contracts, we recommend engaging a Trusted Third Party (TTP) to support Policy implementation across NIH ICOs as well as all funding vehicles. Though such an entity would introduce additional Policy costs in the short term, the TTP would offset those costs in three main ways. First, the TTP would facilitate Policy consistency across ICOs, reducing the costs of variations in Policy implementations for NIH as a whole. Second, the TTP could create a Data Management and Sharing Plan template (described in our feedback on Section VI. Data Management and Sharing Plans), lowering the review costs of varying Plan structures and content for NIH. Third, the TTP could monitor researchers' compliance with Policy terms, cutting down the duplicative costs of multiple NIH ICOs performing the same monitoring work as well as reducing the likelihood of needless enforcement variations that drive up agency costs. For these reasons, we think a TTP would add value to NIH's Policy implementation while driving down the overall cost of Policy implementation.

Section IV: Effective Date(s):

We recommend making two enhancements to the current "IV. Effective Date(s)" guidelines, as follows: (1) create supporting documentation that explains the differences between NIH's past data management and sharing requirements and NIH's new requirements; and (2) establish a support resource (such as an email-based helpdesk) for questions about the new requirements. By making these two adjustments, we believe NIH will improve researchers' ability to successfully implement NIH's new Data Management and Sharing Policy on its effective date, while lowering the cost of Policy implementation on both researchers and NIH.

Creating a Policy comparison document will help existing grantees and contractors understand changes to the data management and sharing documentation that they've completed as part of their current or past NIH-funded work. With this understanding in place, existing grantees and contractors will be able to avoid costly mistakes when adapting to NIH's new Policy. Testifying to this outcome, our MedRIC team has successfully adapted our NIH-sponsored research clients to ever-evolving CMS guidelines by producing comparison documentation on revised policies. This documentation has helped our research clients make fewer mistakes when adapting to CMS's policy revisions. Using this type of documentation as a model, NIH could develop a comparison document that provides researchers with practical guidance on how to adapt their previous NIH data management and sharing documentation to NIH's new policy. Not only would this reduce the likelihood of new Policy-related errors, but it would also establish Policy reference documentation that could lower Policy support costs. For example, instead of having to craft lengthy emails explaining Policy changes to each inquirer, NIH could simply point researchers with comparison questions to its comparison documentation, saving support time and money.

Likewise, establishing a Policy-dedicated support resource, such as an email-based helpdesk, would give researchers a means of verifying and validating their interpretation of NIH's new Policy before composing their Data Management and Sharing Plans. They would then be less likely to produce non-compliant Data Management and Sharing Plans and, as a result, would be able to avoid the time and money needed to revise non-compliant Plans. Likewise, NIH would receive less problematic Plans for review and would expend less time and money on helping researchers to correct their work. A Policy-dedicated support resource would thus offer substantial and ongoing cost savings to both researchers and NIH.

Section V: Requirements:

Introduction

NIH faces two key challenges to the requirements outlined in the "DRAFT NIH Policy on Data Management and Sharing" (Policy)—namely, (1) non-standardized Data Management and Sharing Plans that increase both researcher and NIH Plan-related costs; and (2) datasets not owned by research teams that impede researchers' ability to share data in a timely manner. To overcome these challenges, we recommend establishing a lightweight Data Management and Sharing Plan template (described in greater detail in our response to Section VI. Data

Management and Sharing Plans) as well as reiterate our recommendation in Section I. Purpose and Section II. Definitions to better balance the definition of scientific data between new data that researchers create and existing data that researchers obtain.

Challenges to Requirements

Though NIH's "Supplemental Guidance: Elements of a NIH Data Management and Sharing Plan" ("Guidance") and Section VI. Data Management and Sharing Plans of the Policy contain good recommendations for structuring mandatory Data Management and Sharing Plans, phrases like "Elements of a Plan should consider" (Guidance), "Plans should explain" (Section VI), and "Plans should include" (Section VI) permit variations in Plan structuring and content that risk making Section VI. Requirements' mandate for Plans high-cost. Not only do these phrases increase the likelihood of researchers producing non-compliant Plans, they also increase the chances of high review costs for NIH ICOs, who must dedicate time and effort to vetting varying Plan structures and content. If an NIH ICO identifies substantial issues with Plan structure and/or contents, researchers could then face considerable revision costs to achieve Plan compliance.

Further, researchers interested in using third party-owned data as part of their NIH-funded projects often face convoluted restrictions on the data they can share with other researchers, driving up the effort to "[take] into account any potential restrictions or limitations" for these data. For example, the Centers for Medicare & Medicaid Services (CMS) bans researchers from sharing any CMS data with any individuals not explicitly authorized on researchers' CMS Data Use Agreements (DUAs), but does permit researchers to "disclose" non-small cell data in their publications (<https://www.cms.gov/Medicare/CMS-Forms/CMS-Forms/downloads//cms-r-0235.pdf>). CMS also requires DUA-authorized researchers to destroy all CMS data within 30 days of completing their study, but permits researchers to retain and reuse their CMS data for other studies under a reuse request protocol. With convoluted policies such as these, third party-owned datasets will, at a minimum, produce Plan-related questions on the part of researchers and, in some cases, could result in Policy compliance issues for researchers.

Methods for Overcoming Requirement Challenges

To address these challenges, we recommend establishing a lightweight Data Management and Sharing Plan template file that standardizes the main sections of Plans for all researchers. For more specific template recommendations, refer to our response to Section VI. Data Management and Sharing Plans. In addition, we think NIH should better balance the definition of scientific data between newly created data and existing data by creating additional guidance content that improves Policy implementation. This additional content should include (1) expanding the definition of metadata, such as adding data specifications and statistical programs as examples (per our response to Section II. Definitions); and (2) establishing a support resource, such as an email-based helpdesk, that will help researchers address the data

sharing constraints imposed by third party-owned data (per our response to Section VII. Compliance and Requirements).

Section VI: Data Management and Sharing Plans:

Introduction

NIH's stated goal of making research available to the public is inextricably linked to the level of effort required for researchers to produce successful Plans for accessing, managing, and sharing scientific data. The complexity of the NIH's Plan requirements will thus correlate with more, or less, research project proposals.

To incentivize as many researchers as possible to submit proposals, we think NIH should (1) establish a lightweight, file-based template for researchers' Data Management and Sharing Plan in addition to the "Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan;" and (2) engage a Trusted Third Party (TTP) to serve, among other things, as a Plan consultant for researchers.

Lightweight, File-based Template for Data Management and Sharing Plan

Though the "Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan" ("Guidance") contains good recommendations for Plan content, those recommendations permit too many variations in Plan structure and content, which drive up Plan composition and review costs. We therefore urge NIH to establish a lightweight, file-based template for the Plan containing standard section headings and specific section instructions. Developing such a file will lower not only the cost of producing Plans for researchers but also the cost of reviewing Plans for NIH ICOs.

Indeed, templates such as the one described above have proven invaluable to our Medicare & Medicaid Resource Information Center (MedRIC) for NIH-sponsored researchers. Our templates have directly improved our research clients' ability to successfully request data and/or enclave access from us, as our clients do not have to expend substantial time and effort determining what specific content they need to provide or how to structure that content. These templates have also substantially lowered our review costs, as we receive request information in a more predictable and streamlined way. Given these benefits, we think a Plan template would offer researchers critical support in complying with NIH's Policy directives as well as substantially lower the burden on NIH of having to review wildly varying Plans.

Trusted Third Party (TTP) Consultant for Plan Support

A TTP consultant tasked with providing both administrative and technical consultation on Data Management and Sharing Plans would offer three crucial benefits to researchers and NIH—namely, (1) reducing the likelihood of unacceptable Plans; (2) enhancing researchers' data management and sharing knowledge; and (3) reducing Plan support costs for all NIH ICOs.

First, the TTP would be able to respond to researchers' questions about Plan requirements, reducing the likelihood that these researchers either submit problematic Plans as part of their proposal materials or, worse, abandon proposals because they lack the expertise and resources to overcome data management and sharing challenges on their own. As a result, researchers would be less likely to face substantial revisions from NIH ICOs, lowering researchers' overall Plan production costs in the process.

Second, the TTP would be able to supplement researchers' knowledge of data management and sharing best practices, since many researchers lack substantial data management and sharing experience. Without such experience, researchers will struggle to develop sound Data Management and Sharing Plans on their own. Further, even research teams that can include data management and/or sharing experts as team members may still have questions about acceptable management and sharing methods when their scientific data includes third party-owned data. These types of data can introduce a number of legal, technical, and security challenges that the TTP could help a research team address.

Third, the TTP would reduce the burden on NIH ICOs of addressing Plan questions individually. With a single and centralized TTP, NIH would be able to ensure consistent responses across all NIH ICOs. The TTP would then be in a position to develop frequently asked questions and responses for public consumption, improving the self-help resources available to researchers at any time of the day or night and thus reducing the need for extensive and costly live support. In providing these services, the TTP would lower the overall cost of Plan support for NIH and its ICOs.

Section VII: Compliance and Enforcement:

Though a critical component of the "DRAFT NIH Policy on Data Management and Sharing" ("Policy"), Section VII. Compliance and Enforcement's terms pose three key implementation challenges to NIH and its ICOs. First, dispersing Policy enforcement to each NIH ICO risks Policy enforcement variations that increase costs, as NIH will not be able to streamline enforcement inconsistencies across NIH ICOs. As a result, NIH may face higher compliance costs than are actually warranted. Second, supporting a distributed data storage model—wherein each research team could store their collected data at their own remote locations—means that NIH must rely on researchers' self-reported compliance to verify that their NIH-funded research projects satisfy Policy terms. While this approach upholds NIH's mandates for researchers to honestly and thoroughly report on their compliance, the success of the approach wholly depends on researchers' ability to fully understand and effectively report on that compliance. If a research team misinterprets compliance terms and/or does not correctly report their compliance, NIH has no means of detecting these compliance issues. Third, emphasizing scientific data that a research team directly collects or generates, over data that a research team obtains under a license or agreement, means that NIH's current compliance and enforcement terms do not provide researchers with sufficient guidance on third party-owned data. Though Section VI. Data Management and Sharing Plans explicitly references data sharing

restrictions, the current guidance is still too high-level to help researchers address the extensive and convoluted legalese governing third party-owned data sets. For example, researchers' use of sensitive federal information, such as Medicare and Medicaid claims data, subjects those researchers to federal security mandates that can conflict with NIH's goal of making scientific data available to the public. Resolving those conflicts requires not only substantial federal security expertise but also out-of-the-box thinking on scientific data alternatives that replace the need for raw data sharing. As a result, many researchers will struggle to identify viable methods for both complying with the terms governing third party-owned data and meeting their obligations under NIH's Policy. Underemphasizing third party-owned data, limiting enforcement to researcher team's self-reported compliance, and dispersing compliance terms to each NIH ICO thus introduce considerable hurdles to the successful implementation of the "DRAFT NIH Policy on Data Management and Sharing."

To overcome these hurdles, we urge NIH to consider engaging a Trusted Third Party (TTP) that can facilitate the establishment of consistent compliance policies across NIH ICOs, offer NIH a pathway to independently validating the compliance of research projects, and support researchers in identifying innovative methods of sharing their scientific data, when that data includes restricted, third party-owned data.

Working closely with NIH, a TTP could establish consistent compliance policies across NIH ICOs and, when policy variations are needed, work with NIH to ensure that variations in policies are not only warranted but also clearly communicated to researchers. Likewise, a TTP could provide federal information systems, such as federally certified remote-access enclaves like our MedRIC remote-access enclave, to researchers for storing, processing, and sharing federal and non-federal data sets. Such systems would empower NIH ICOs to track and enforce compliance with researchers' Data Management and Sharing Plans through logs of user actions. As a result, NIH ICOs would be able to independently validate claims in annual Research Performance Progress Reports (RPPRs) and thereby transcend current self-reported methods. Finally, a TTP could serve as a central support resource for researchers facing restricted data sharing as part of their research projects. For example, a TTP could help researchers identify metadata resources (e.g., non-sensitive data specifications and researcher-developed analytic programs) that uphold third party restrictions on data sharing yet still provide other researchers with sufficient information to replicate scientific findings. Given these outcomes, we recommend that NIH consider a TTP as a means of developing more robust compliance and enforcement terms.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Introduction

The "Supplemental Draft Guidance: Allowable Costs for Data Management and Sharing" ("Guidance") provides critical budgeting guidance on data management and sharing to prospective NIH grant/contract awardees. However, we believe the Guidance underestimates certain costs related to data access, structuring, sharing, and storage. To address these costs,

NIH should engage a Trusted Third Party (TTP) tasked with, among other things, controlling and lowering data-related costs for both researchers and NIH ICOs. In the sections below, we describe the data-related costs, explain the value of a TTP in controlling those costs, and suggest clarifications to Guidance terminology to support researchers' implementation of budgeting terms.

Cost-Related Challenges of Current Guidance Draft

In our work as the Medicare & Medicaid Resource Information Center (MedRIC) for NIH-sponsored researchers, we've learned that our clients face the following four data-related costs.

1. **Costs of Accessing Data.** Researchers can face enormous fees for accessing third party-owned data, making the exclusion of such fees from NIH grant and contract budgets problematic. For example, CMS can charge over \$1,000,000 for longitudinal versions of its datasets, while the CDC can charge \$500,000 for information on 100,000 decedents. Given that NIH's R01 grants typically provide researchers with \$250,000.00 in direct costs per year, data access fees, such as those charged by CMS and the CDC, create insurmountable data access barriers for researchers who need these data as part of their research projects.
2. **Costs of Structuring Data.** Many researchers lack substantial experience structuring data for other researchers to use, making clear and specific data structuring guidance critical to NIH's data management and sharing goals. While resources such as Office of the National Coordinator (ONC)'s Interoperability Standards Advisory (ISA), CDISC, and Logical Observation Identifiers Names and Codes (LOINC) offer valuable guidance to researchers, these resources require time and effort to understand, implement, and adapt to project-specific data. As a result, many first-time researchers tend to underestimate the costs of implementing those techniques in the planning stages of their studies, resulting in unanticipated costs once they reach their projects' data structuring stage. More critically, many researchers can only identify questions about data structuring during the data structuring phase, increasing their work load and costs after grant or contract funds have been awarded.
3. **Costs of Sharing Data.** Many researchers encounter non-trivial costs in determining appropriate methods for sharing data. For data subject to HIPAA, appropriate data sharing methods require team members with proven skills in data de-identification methods that often go beyond Safe Harbor to Expert Determination. This work requires a strong understanding of the potential for publically available forms of information to pose a re-identification risk. For data subject to FISMA, data sharing methods must include, but are not limited to, identity proofing every researcher requesting access to these data, establishing multifactor authentication protocols, and segmenting network systems into multi-tiered zones—all of which require technical expertise, effort, and funds. As a result, the work needed to share scientific data often exacts an amount of time and effort that many researchers cannot accurately estimate in the planning stages of their study.

4. **Costs of Data Storage.** Hosting and managing secure data repositories can total several million dollars to the organizations that implement them. These organizations then seek to defray their high operational costs through high end-user fees. For example, CMS's Virtual Research Data Center (VRDC) charges researchers an annual fee of \$25,000 per user to offset the costs of maintaining the VRDC's multi-million dollar per year operational costs. In this way, then, storing scientific data can take a toll on both grant/contract funds during the grant/contract period and researchers' personal funds, when NIH grant/contract funds end.

A Trusted Third Party to Streamline Data-Related Costs

Given these cost challenges, we recommend that NIH engage a TTP to assist NIH in lowering data access fees, serving as researchers' consultants on data structuring questions, providing data sharing guidance, and offering affordable systems with predictable costs for data preservation.

A TTP could empower NIH to negotiate more favorable data access fees with their owners. For example, NIA contracted our group to centralize CMS data access for NIH-sponsored researchers interested in linking CMS data to NIH survey data. This contract work includes having our group negotiate and acquire CMS data for all of our NIH-sponsored research clients. As a result, NIA has also been able to limit our NIH-sponsored research clients' data acquisition costs to \$3,000 per year—that is, a \$2,000 payment to CMS for the receipt of a CMS data use agreement and \$1,000 to our group to cover the costs of extracting, documenting, and distributing CMS data. Using this model with other TTPs, NIH would be able to lower and streamline the overall costs of data acquisition for grant and contract applicants.

A TTP could also establish reliable, on-demand consultants for researchers' questions on methods of data structuring and data sharing. For example, the TTP could assist research teams in either adapting data structuring standards to their project-specific data or identifying HIPAA- and FISMA-compliant sharing methods. As a result, the TTP would accelerate researchers' ability to effectively structure and share data, lowering the costs of these activities for researchers. In doing so, the TTP would also lower the costs on NIH ICOs' review of these activities, as the TTP would help researchers avoid data structuring and sharing mistakes that, in turn, raise compliance and enforcement costs for NIH ICOs.

Finally, a TTP could offer low-cost data storage systems, such as remote access enclaves, to researchers. Such systems could achieve substantial cost savings through economies of scale and the defraying of data management costs across many researchers. Instead of each individual research team either developing or purchasing their own data management systems, a TTP could offer a remote-access enclave that reduces duplicative infrastructure costs and, through economies of scaling, keep usage fees low.

Content Clarification Comments

In "1. Curating and Developing Supporting Documentation," we recommend clarifying "accepted community standards" through examples such as those cited in the "DRAFT NIH Policy for Data Management and Sharing" statement. We also recommend that NIH add cost guidelines for researchers to update their datasets whenever acceptable community standards change.

In "2. Preserving and Sharing Data through Established Repositories," we think NIH should clarify the term "repositories" through a definition and examples to ensure that researchers understand what systems qualify as repositories and what systems do not. We also think NIH should clarify whether recurring repository fees are covered entirely or only for preserving and sharing data.

In "3. Local Data Management Considerations," we strongly recommend that NIH clarify the difference between data access and the data preservation fees specified in "2. Preserving and Sharing Data through Established Repositories," as this difference has substantial cost implications. We also urge NIH to allow a budget line item for acquiring data, as doing so will ensure that NIH can track the costs of various data access fees over time as well as understand whether a TTP-based data acquisition model offers cost savings to the agency.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Introduction

After reviewing the "Supplemental DRAFT Guidance: Elements of an NIH Data Management and Sharing Plan" (Guidance), our group foresees two key challenges to successful implementation—namely, (1) data preservation and sharing constraints common in third party-owned research data sets; and (2) the inability of many researchers to develop and/or operate systems for data archiving and sharing on their own.

In this response, we provide suggestions for improving the Guidance by (1) enhancing guidelines on third party-owned data; and (2) explicitly discussing Trusted Third Party (TTP)-operated data management and sharing systems—such as remote-access enclaves—as opposed to simply encouraging the use of shared data repositories.

Enhance Third Party-Owned Data Guidance

We think the "Data Sharing Agreements, Licenses, and Other Use Limitations" contains vital guidelines for data sharing restrictions, including clearly highlighting the importance of accounting for these restrictions in Policy-compliant Data Management and Sharing Plans. However, these restrictions have deeper implications for other portions of the Guidance. For example, federal agencies typically have dense, complex, and hard-to-understand regulations governing their data access and retention policies. Researchers often struggle to fully grasp these regulations, which makes "providing a rationale for decisions about which scientific data are to be preserved and made available for sharing" for federal data sets incredibly challenging for researchers to do without additional Guidance language.

To help researchers overcome these challenges, we recommend accounting for data access restrictions (such as banned and/or limited data sharing terms common to third party-owned data) in the 1. Data Type section more explicitly by providing examples of appropriate justifications when data sharing and preservation are restricted. For example, NIH could enhance its description of privacy and confidentiality protections—that is, "outlining plans for providing appropriate protections of privacy and confidentiality (i.e., through de-identification or other protective measures) that are consistent with applicable federal, tribal, state, and local laws, regulations, statutes, guidance, and institutional policies" (1.Data Type)—by creating citation guidelines for protections specified in third party-owned data access agreements available on the Web. In doing so, NIH would alleviate the burden of researchers having to rearticulate these protections and ensure that NIH ICOs receive accurate descriptions of these protections from the organizations that defined them.

TTP-Based Enclaves for Data Management and Sharing

Though the Guidance does cover data repositories (e.g., "availability of suitable data repositories" and "submitted to specified data repositories"), that coverage does not sufficiently counterbalance the strong emphasis on researcher-controlled and/or –developed systems. And while researchers must be accountable for planning the management and sharing of their scientific data, our MedRIC project has revealed, again and again, that most researchers struggle with the technical and financial challenges of establishing and/or operating their own data management and sharing systems. As such, a TTP could also offer researchers data management options that shift the technical burden and costs from researchers to a seasoned TTP. For example, MedRIC offers a NIST-compliant virtual data enclave that stores sensitive CMS and NIH-sponsored survey datasets for research purposes; that provides rights-based access for users to perform statistical research via remote connections; and that offloads the burdens of appropriately structuring data for reuse to our data experts.

Enclaves such as the MedRIC one offer researchers three key advantages over building and/or operating their own data management and sharing systems: (1) they reduce the infrastructure and equipment costs that researchers need to budget in support of NIH's new Policy; (2) they provide SOC II/Type II-or FISMA-certified environments for sensitive data research, reducing the substantial security costs that researchers must take on; and (3) they streamline compliance requirements for NIH ICOs. Use of such enclaves enables researchers to access and deploy proven infrastructure and equipment, rather than having to obtain NIH funds for, build, and operate this infrastructure from scratch. For researchers lacking substantial data systems experience, this option could be the difference between successfully fulfilling Policy requirements and not submitting a research request at all. Leveraging existing SOC II/Type II- or FISMA-complaint enclaves, rather than having to acquire the security expertise and resources to achieve this security compliance independently, means that researchers can optimize their security compliance and lower their overall security costs. Without such options, most researchers would require millions of dollars to establish SOC II/Type II- or FISMA-compliant

infrastructure. Finally, by supporting the use of existing enclave systems, NIH can define concrete and specific system requirements; validate those requirements through government or private-sector auditing firms; and lower the burden on NIH ICOs of having to assess numerous remote systems, based on self-reported statements about Policy compliance. TTP-operated enclaves would thus offer researchers a cost-effective option to weigh against the potentially prohibitive costs of building or operating a secure information system and documenting its compliance with federal standards on their own.

Other Considerations Relevant to this DRAFT Policy Proposal:

Introduction

In addition to the barriers and burdens we outlined in our other responses, we believe that researchers face crucial security barriers to complying with the "DRAFT NIH Policy on Data Management and Sharing" because they lack the security expertise and resources needed to establish secure mechanisms for protecting sensitive information—including Personally Identifiable Information (PII) and Protected Health Information (PHI)—when such information is necessary to their research project. These barriers consist of (1) lack of familiarity with various federal data management policies governing protected information; (2) expertise implementing information protections; and (3) lack of experience implementing security-compliant systems.

The remainder of this statement describes the security barriers, as well as the impact of these barriers on NIH ICOs, in detail. The statement then outlines the benefits of engaging a Trusted Third Party (TTP), including helping researchers address thorny security issues and establishing verifiable information security measures for NIH.

Barrier to Secure Mechanisms for Protecting Identifiable Information

Researchers interested in using government-owned data as part of their research project often lack practical familiarity with the government policies governing the acquisition, management, and sharing of protected information. If incorrectly understood, these policies create the potential for unintended violations on the part of researchers. As an example, the Centers for Medicare & Medicaid Services (CMS) bans the redistribution of CMS data to anyone not listed on the CMS data use agreement (DUA). This requirement creates both procedural challenges and legal risks for NIH-sponsored researchers, who need CMS approval to share CMS data used in their research projects. Researchers may face similar hurdles gaining access to, and documenting use of, research data approved under strict Institutional Review Board standards. While some more experienced researchers may be familiar with the rules governing data owned by different entities, others will limit their research plans to avoid the administrative difficulty of working with multiple federal and private entities to document compliance.

The challenge of protecting sensitive information does not end with documenting data access. Even when researchers are able to identify appropriate data sharing rules for data they have access to, they often lack the expertise to protect these data in accordance with multiple laws

and regulations. For instance, while many researchers can become well-versed in the Health Insurance Portability and Accountability Act (HIPAA) standards, implementing HIPAA's Expert Determination de-identification methods requires unique expertise that many research teams lack. Not only must those teams understand HIPAA standards and statistical determinations, they must also expend substantial time analyzing the potential for small cells when data users run statistical model after model on a data set. On top of these challenges, NIH-sponsored researchers will be responsible for what is arguably a much more onerous task—namely, assessing whether the combination of their scientific data with the universe of publicly available information potentially increases re-identification risks for their cohort.

Information system security standards create perhaps the most costly overhead for researchers. These standards require organizations possessing confidential information to satisfy complex and extensive federal security and privacy requirements, which may include the Federal Information Security Modernization Act of 2014 (FISMA Reform); NIST SP 800-171 Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations; FIPS 140-2 Security Requirements for Cryptographic Modules; the Privacy Act of 1974; and more. Such requirements not only demand a strong understanding of requirement specifications, but also necessitate extensive experience translating these specifications into compliant information systems, making compliance work time-consuming and costly. Smaller institutions and other organizations without significant investment in information security face even higher hurdles, as information security compliance would require directing substantial research funds to systems security. Finally, distribution and storage of protected information at numerous geographical locations blocks NIH from independently validating the compliance of each information system, converting compliance from independent verification of researchers' activities to unfounded assessments of researchers' self-reported statements.

A TTP for Security Support

We believe the NIH should engage a TTP to help both researchers and NIH ICOs overcome the information security challenges of protecting sensitive information. This TTP would offer researchers expert guidance and resources on security, while simultaneously relieving NIH ICOs of much of the administrative burdens associated with enforcing information security compliance in two key ways.

First, the TTP could facilitate communications between researchers and NIH ICOs by providing subject matter expertise in security and privacy requirements. In this role, a TTP could help researchers adapt existing security standards to their specific research objectives, paving the way for other researchers to access and use their scientific data. For example, a TTP, like MedRIC, could provide information sharing alternatives for data that cannot be archived and/or shared. When our NIH-sponsored researchers face such constraints, our team has recommended archiving unrestricted data specifications and researcher-developed statistical programs, in lieu of access-restricted data sets. These non-sensitive resources provide other

researchers with the information needed to request the exact same data sets from their owners, as well as replicate the methods (statistical programs) used to arrive at study findings.

Second, a TTP could provide researchers with a security-compliant data-access platform—or remote-access enclave—that reduces the burden of developing project-specific systems. This enclave would feature expert-driven access protocols that conform to both industry best practices (e.g., multifactor authentication) and, when applicable, federal security mandates (e.g., identity proofing). As a result, NIH would not only be able to define concrete security measures for a wide range of scientific data, but also independently verify those measures via 3PAOs available to the federal government or certified security auditors in the private sector (e.g., SOC II Type II auditors). A TTP would thus reduce the burden on NIH ICOs' Policy oversight, via standardization, and overcome the challenges of Policy verification, via TTP audits or certifications.

Attachment:

Description:

Submission ID: 1407

Date: 1/10/2020

Name: Alessia Daniele

Name of Organization: Cornell University and Weill Cornell Medicine

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of the above

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Research Administration and Information Technology

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We agree with the purpose of this policy and appreciate that data sharing underpins the advancement of biomedical research.

Section II: Definitions:

In the definition of "Data Management and Sharing Plan (Plan)," we appreciate the inclusion of the modifier "as appropriate." This acknowledges that not all data is useful and allows investigators and institutions the latitude to determine when and which data is appropriate to manage and share.

Section III: Scope:

We are concerned with the scope of this policy. As NIH's budget leveled off after the doubling in the early 2000s, inflation eroded away all of those gains. In the last five years, however, substantial bipartisan majorities in Congress have provided average annual increases of \$2 billion to NIH's appropriation. While this has helped considerably to make up lost ground, NIH is still operating below where it would have been had the budget kept pace with inflation. It makes little sense to divert dollars away from research that is improving our understanding of human disease, developing life-saving treatments and cures, and improving the lives of patients and their families. We will continue to advocate for robust funding for the NIH, but we know there are no guarantees that the agency will continue to receive comparable increases if economic growth tails off.

This unfunded mandate also runs contrary to recent directives by the Department of Health and Human Services (HHS) to reduce administrative burden by limiting the oversight requirements of Institutional Review Boards (IRBs). Because fewer projects now require IRB review, it will be difficult for institutions to oversee every single data management and sharing plan without implementing another board or screening mechanism to ensure investigators are submitting appropriate plans. Considering that Cornell University has over 220 active NIH awards and Weill Cornell Medicine has over 300, availing the resources required to screen, implement and monitor data management and sharing plans for every single grant will put a significant strain on our respective research administrations.

Section IV: Effective Date(s):

We recommend that the final policy adopted allow institutions at least one year to make changes to internal policies and procedures and at least one additional year thereafter before requiring full compliance. We encourage the Office of Science Policy to continue to carefully consider all the impacts on research output as it moves to adopt and implement this policy.

Section V: Requirements:

The draft policy and supplemental guidance documents are vague on several critical requirements of the policy. We know that an award will be able to include data management and sharing costs, but the guidance does not specify the maximum dollar amount or percentage of an award that can be allocated to the plan. Similarly, the draft does not provide sufficient information about the mechanism for allocating award funds after a project has concluded to allow for long-term data management costs.

It is possible that 2 CFR 200.461 (b), which allows researchers to incur publication costs during the period of an award even though they may be delivered later, could be interpreted to cover the cost of data storage, which is just as critical for sharing as publication. Without further guidance, however, we do not see any other currently available mechanism to include data storage costs as a direct cost of research.

Funding data storage and sharing through Facilities and Administrative (F&A) costs would not be feasible without a substantial increase to the allowed percentage for these costs. Otherwise, adding the significant expense of data management would drastically reduce the amount of funds available for other necessary F&A expenses.

Treating data storage as a direct cost which must be negotiated as part of establishing every individual research contract is not optimal. We believe it would be far better if NIH established a new mechanism for providing these services as an integral part of any new data sharing policy adopted by the agency. One possibility would be to create a recharge entity that would be operated on a break-even basis. This entity would estimate the costs and provide or procure appropriate services for data storage and sharing during the required period. NIH could also change its regulations to allow the creation of an institutional reserve for the same purpose. A recharge entity or an institutional allowance would provide a fair and uniform way to allocate storage costs while relieving individual researchers from tasks likely outside their expertise and interest.

Whatever cost mechanism is ultimately adopted, we encourage NIH ICOs to maintain their exemplary level of programmatic and project support without adding penalties or additional reporting burden that impedes research.

Section VI: Data Management and Sharing Plans:

We outline our concerns regarding data from human subjects in greater detail in subsequent sections.

Section VII: Compliance and Enforcement:

We are concerned compliance will require substantial resources in addition to the significant resources already invested in data management and sharing by both Weill Cornell Medicine and Cornell University. This is especially concerning given the volume of our respective NIH research portfolios.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

It is critical that NIH define "reasonable" as it pertains to allowable storage costs. We also need additional information on the maximum length of time research budgets can pay for recurring data repository fees. Similar to the point made in Section V, we need to know how to transfer funds after the end date of the award and if we will be expected to pay upfront costs for both short- and long-term storage.

We recommend that plan preparers have complete discretion in selecting long-term and short-term repositories.

We read the statement that "Costs associated with collecting or otherwise gaining access to research data (e.g., data access fees) are considered costs of doing research and should not be

included in budgets" to mean the data management plan budget and not the overall project budget.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We appreciate the flexibility provided to investigators and their research projects by the supplemental guidance.

We are concerned, however, that the "Data Preservation, Access, and Associated Timelines" outlined in Point 4 regarding human subject research does not address consent, thus effectively requiring that consent forms include language stating that data and specimens will be shared. Without some clarification, it is possible that data or specimens previously obtained without consent that acknowledges the parameters of a data management plan may not be shareable. To meet due diligence standards for protection of human subjects, institutions would have to go back and compare each consent to the Plan, incurring considerable additional burden and expense. We recommend that NIH provide more information on secondary research consent as it pertains to this policy.

We cannot state strongly enough that the protection of patient privacy is a higher ethical demand than the need to share data. This policy should not compromise privacy, nor should it prevent important research that can be performed ethically only if privacy is not compromised, even if that diminishes the ability to share. Because of the complexities of consent, it may be beneficial to separate or exempt human subject research from this policy.

Other Considerations Relevant to this DRAFT Policy Proposal:

As we read the proposed policy, we cannot confidently ascertain if institutions such as Weill Cornell Medicine and Cornell University can make data available through institutional repositories at a cost to anyone who wishes to obtain the data and whether there are any limits to costs for such access. It is our experience that costs to access data in cloud-based repositories are far greater than the costs of storing it. The ability to charge back the cost of data retrieval to those requesting the data could provide a critical element of sustainability to this policy. Additional guidance would help efforts to budget for the resources required to comply with the final policy.

We are also concerned the policy requires an increased understanding of information technology practices and principals with which investigators might not be familiar, such as understanding whether the security, backup or encryption practices of a data repository are vulnerable or adequate.

Additional information on a potential dispute resolution mechanism or process would also be helpful, as the policy does not appear to provide a means by which institutions could challenge or adjudicate rejection decisions by an ICO program manager. We strongly recommend that NIH develop such a process.

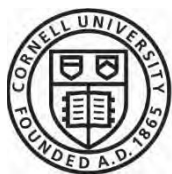
The importance of protecting human subject research data and privacy cannot be overstated. Based on the draft policy and supplemental guidance, we cannot confidently assess whether this policy conflicts with the Office for Human Research Protections (OHRP) policy on consent and data sharing. Additional guidance from OHRP would be useful so that institutions can ensure data management plans do not

Attachment:

Cornell NIH Data Management and Sharing Policy Letter 01-10-20.pdf

Description:

Response Letter



Cornell University

January 10, 2020

Andrea Jackson-Dipina, Ph.D.
Director
Division of Scientific Data Sharing Policy
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Dear Dr. Jackson-Dipina,

On behalf of Cornell University and Weill Cornell Medicine, we write to provide comments on the draft NIH Policy for Data Management and Sharing and supplemental draft guidance published in the Federal Register on November 8, 2019.

Cornell University is a world-class research institution known for the breadth and rigor of its curricula, and an academic culture dedicated to preparing students to be well-educated and well-rounded citizens of the world. Its faculty, staff and students believe in the critical importance of knowledge—both theoretical and applied—as a means of improving the human condition and solving the world’s problems. With campuses in Ithaca, New York and New York City, including Weill Cornell Medicine, and a location in Doha, Qatar, Cornell is a private Ivy League research university and the land-grant institution of New York State. 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 clinical care and cutting-edge research that connects patients to the latest treatment innovations and prevention strategies.

The NIH is the world’s foremost source of funding for biomedical research. It is a testament to the excellence of research by our scientists and physicians that Cornell University and Weill Cornell Medicine are so well-supported by the agency. We value and acknowledge that data management and sharing is the bedrock of scientific advancement and appreciate the judicious approach NIH has taken to obtain stakeholder input on the data management and sharing policy. We are particularly grateful that this draft reflects some of the comments submitted by the academic medicine and research communities, including Weill Cornell Medicine¹, to the 2018 Request for Information (RFI) on this subject. Despite

¹ [WCM 2018 NIH Data Management and Sharing Comment Letter.pdf](#)

these improvements, however, we remain deeply concerned by the depth and breadth of the proposal. If implemented in its current form, this policy represents a significant unfunded requirement that will siphon considerable funds away from scientific investigation to satisfy expanded data storage and sharing mandates, in addition to the significant resources we have already invested in developing and maintaining the data cores on our respective campuses. The policy also raises serious ethical and privacy concerns for human subjects.

This letter sets out our concerns and requests additional clarification and guidance on certain provisions.

We appreciate the opportunity to provide input and would be pleased to provide further assistance as you develop and implement this policy. If you have questions, require any additional information or would like to speak further about the subjects covered in this letter, please contact Alessia Daniele, Associate Director of Federal Affairs at Weill Cornell Medicine, at 646-962-9485 or via email at ALD2035@med.cornell.edu.

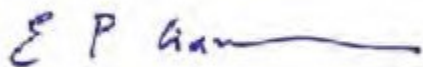
Sincerely,



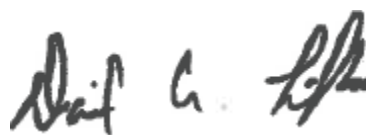
Curtis L. Cole, MD, FACP
Chief Information Officer
Assistant Vice Provost for Information Technology
Frances and John L. Loeb Associate Professor of
Libraries and Information Technology
Associate Professor of Clinical Medicine,
Healthcare Policy and Research
Weill Cornell Medicine



Hugh C. Hemmings Jr., MD, PhD, FRCA
Senior Associate Dean for Research
Chair of the Department of Anesthesiology
Joseph F. Artusio, Jr. Professor of Anesthesiology
Professor of Pharmacology
Weill Cornell Medicine



Emmanuel P. Giannelis
Vice Provost for Research
Vice President for Technology Transfer,
Intellectual Property and Research Policy
Walter R. Read Professor of Engineering, College
of Engineering
Cornell University



David A. Lifka
Chief Information Officer
Vice President for Information Technologies
Director, Cornell Center for Advanced Computing
Cornell University

Section I: Purpose

We agree with the purpose of this policy and appreciate that data sharing underpins the advancement of biomedical research.

Section II: Definitions

In the definition of “Data Management and Sharing Plan (Plan),” we appreciate the inclusion of the modifier “as appropriate.” This acknowledges that not all data is useful and allows investigators and institutions the latitude to determine when and which data is appropriate to manage and share.

Section III: Scope

We are concerned with the scope of this policy. As NIH’s budget leveled off after the doubling in the early 2000s, inflation eroded away all of those gains. In the last five years, however, substantial bipartisan majorities in Congress have provided average annual increases of \$2 billion to NIH’s appropriation. While this has helped considerably to make up lost ground, NIH is still operating below where it would have been had the budget kept pace with inflation. It makes little sense to divert dollars away from research that is improving our understanding of human disease, developing life-saving treatments and cures, and improving the lives of patients and their families. We will continue to advocate for robust funding for the NIH, but we know there are no guarantees that the agency will continue to receive comparable increases if economic growth tails off.

This unfunded mandate also runs contrary to recent directives by the Department of Health and Human Services (HHS) to reduce administrative burden by limiting the oversight requirements of Institutional Review Boards (IRBs). Because fewer projects now require IRB review, it will be difficult for institutions to oversee every single data management and sharing plan without implementing another board or screening mechanism to ensure investigators are submitting appropriate plans. Considering that Cornell University has over 220 active NIH awards and Weill Cornell Medicine has over 300, availing the resources required to screen, implement and monitor data management and sharing plans for every single grant will put a significant strain on our respective research administrations.

Section IV: Effective Date(s)

We recommend that the final policy adopted allow institutions at least one year to make changes to internal policies and procedures and at least one additional year thereafter before requiring full compliance. We encourage the Office of Science Policy to continue to carefully consider all the impacts on research output as it moves to adopt and implement this policy.

Section V: Requirements

The draft policy and supplemental guidance documents are vague on several critical requirements of the policy. We know that an award will be able to include data management and sharing costs, but the guidance does not specify the maximum dollar amount or percentage of an award that can be allocated to the plan. Similarly, the draft does not provide sufficient information about the mechanism for allocating award funds after a project has concluded to allow for long-term data management costs.

It is possible that 2 CFR 200.461 (b), which allows researchers to incur publication costs during the period of an award even though they may be delivered later, could be interpreted to cover the cost of data storage, which is just as critical for sharing as publication. Without further guidance, however, we do not see any other currently available mechanism to include data storage costs as a direct cost of research.

Funding data storage and sharing through Facilities and Administrative (F&A) costs would not be feasible without a substantial increase to the allowed percentage for these costs. Otherwise, adding the significant expense of data management would drastically reduce the amount of funds available for other necessary F&A expenses.

Treating data storage as a direct cost which must be negotiated as part of establishing every individual research contract is not optimal. We believe it would be far better if NIH established a new mechanism for providing these services as an integral part of any new data sharing policy adopted by the agency. One possibility would be to create a recharge entity that would be operated on a break-even basis. This entity would estimate the costs and provide or procure appropriate services for data storage and sharing during the required period. NIH could also change its regulations to allow the creation of an institutional reserve for the same purpose. A recharge entity or an institutional allowance would provide a fair and uniform way to allocate storage costs while relieving individual researchers from tasks likely outside their expertise and interest.

Whatever cost mechanism is ultimately adopted, we encourage NIH ICOs to maintain their exemplary level of programmatic and project support without adding penalties or additional reporting burden that impedes research.

Section VI: Data Management and Sharing Plans

We outline our concerns regarding data from human subjects in greater detail in subsequent sections.

Section VII: Compliance and Enforcement

We are concerned compliance will require substantial resources in addition to the significant resources already invested in data management and sharing by both Weill Cornell Medicine and Cornell University. This is especially concerning given the volume of our respective NIH research portfolios.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

It is critical that NIH define “reasonable” as it pertains to allowable storage costs. We also need additional information on the maximum length of time research budgets can pay for recurring data repository fees. Similar to the point made in Section V, we need to know how to transfer funds after the end date of the award and if we will be expected to pay upfront costs for both short- and long-term storage.

We recommend that plan preparers have complete discretion in selecting long-term and short-term repositories.

We read the statement that “Costs associated with collecting or otherwise gaining access to research data (e.g., data access fees) are considered costs of doing research and should not be included in budgets” to mean the data management plan budget and not the overall project budget.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan

We appreciate the flexibility provided to investigators and their research projects by the supplemental guidance.

We are concerned, however, that the “Data Preservation, Access, and Associated Timelines” outlined in Point 4 regarding human subject research does not address consent, thus effectively requiring that consent forms include language stating that data and specimens will be shared. Without some clarification, it is possible that data or specimens previously obtained without consent that acknowledges the parameters of a data management plan may not be shareable. To meet due diligence standards for protection of human subjects, institutions would have to go back and compare each consent to the Plan, incurring considerable additional burden and expense. We recommend that NIH provide more information on secondary research consent as it pertains to this policy.

We cannot state strongly enough that the protection of patient privacy is a higher ethical demand than the need to share data. This policy should not compromise privacy, nor should it prevent important research that can be performed ethically only if privacy is not compromised, even if that diminishes the ability to share. Because of the complexities of consent, it may be beneficial to separate or exempt human subject research from this policy.

Other Considerations Relevant to this DRAFT Policy Proposal

As we read the proposed policy, we cannot confidently ascertain if institutions such as Weill Cornell Medicine and Cornell University can make data available through institutional repositories at a cost to anyone who wishes to obtain the data and whether there are any limits to costs for such access. It is our experience that costs to access data in cloud-based repositories are far greater than the costs of storing it. The ability to charge back the cost of data retrieval to those requesting the data could provide a critical element of sustainability to this policy. Additional guidance would help efforts to budget for the resources required to comply with the final policy.

We are also concerned the policy requires an increased understanding of information technology practices and principals with which investigators might not be familiar, such as understanding whether the security, backup or encryption practices of a data repository are vulnerable or adequate.

Additional information on a potential dispute resolution mechanism or process would also be helpful, as the policy does not appear to provide a means by which institutions could challenge or adjudicate rejection decisions by an ICO program manager. We strongly recommend that NIH develop such a process.

The importance of protecting human subject research data and privacy cannot be overstated. Based on the draft policy and supplemental guidance, we cannot confidently assess whether this policy conflicts with the Office for Human Research Protections (OHRP) policy on consent and data sharing. Additional guidance from OHRP would be useful so that institutions can ensure data management plans do not

conflict with human subject protections. Careful consideration must be given to how this policy would impact research that involves the quickly evolving field of genomics and the collection of vast amounts of genomic data. It may never be possible to share this data without compromising privacy.

Submission ID: 1408

Date: 1/10/2020

Name: Brian Scarpelli & Alexandra McLeod

Name of Organization: Connected Health Initiative

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of the Above

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: Digital Health Multistakeholder Effort

Domain of Research Most Important to You or Your Organization:

Role of digital health tools in preventing and treating a variety of conditions (condition agnostic).

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Dear Dr. Jackson-Dipina:

The Connected Health Initiative (CHI) appreciates the opportunity to provide input on the National Institutes of Health's (NIH) Draft Policy for Data Management and Sharing and Supplemental Draft Guidance, intended to promote effective and efficient data management and sharing. Therefore, this draft policy and supplemental guidance will further NIH's commitment to making the results and accomplishments of the research it funds and conducts available to the public.

CHI represents a broad consensus of stakeholders across the healthcare and technology sectors whose mission is to support the responsible and secure use of connected health innovations throughout the continuum of care to improve patients' and consumers' experience and health outcomes. We advocate before the Department of Health and Human Services (HHS) on realizing the benefits of an information and communications technology-enabled American healthcare system. CHI is committed to advancing an interoperable healthcare system that enables the bidirectional flow of necessary health data between provider and patient, as well as between other important stakeholders who have a role in improving care coordination and decision-making.

The efficacy of precision medicine, population health, clinical decision support—and artificial/augmented intelligence (AI)- driven tools in particular—is dependent in large part on the availability of massive data sets. The free flow of information and interoperability are therefore important and potentially life-saving for patients.

NIH's proposed policy comes at an important time. There is no disputing that interoperability and patient access to health information prevent timely and informed care coordination and decision-making. Further, electronic health information and educational resources are critical tools that empower and engage patients in their own care regimens. CHI strongly believes that a truly interoperable eCare system includes patient engagement facilitated by store-and-forward technologies (ranging from connected medical devices to general wellness products) with open application programming interfaces (APIs) that allow the safe and secure introduction of patient-generated health data (PGHD) into electronic health records (EHRs). Data stored in standardized and structured formats with interoperability facilitated by APIs provides analytics as well as near real-time alerting capabilities. The use of platforms for data streams from multiple and diverse sources will improve the healthcare sector, helping to eliminate information silos, data blocking, and deficient patient engagement. Interoperability must not only happen between providers, but also between remote patient monitoring (RPM) products, medical devices, and EHRs. NIH's approach to data management and sharing are important for those stakeholders directly engaged with NIH, as well as to the wider healthcare community through the precedent NIH sets.

Based on the above, we provide the following viewpoints and recommendations on NIH's draft policy:

- CHI is generally supportive of NIH's efforts to update and improve its approach to data management and sharing. Specifically, we support NIH making scientific data publicly available at no (or nominal) cost in as timely a manner as possible. However, we believe that NIH's

approach, as proposed, may not align with information sharing norms in the public and private sector.

- A logical, objective approach is necessary to reduce confusion, and NIH should align its data management and sharing policy with the Office of the National Coordinator for Health IT's (ONC's) information blocking to the extent possible. While this rule is currently approaching finalization, it will represent the baseline for information sharing moving forward, and NIH should align its data management and sharing policies with these rules to the maximum extent possible to provide continuity across the healthcare ecosystem. For example, CHI recommends use of the Fast Healthcare Interoperable Resources (FHIR) standard (Release 4) as well as HL7 U.S. Core FHIR Implementation Guides (or in the alternative that NIH permit the use of such widely-accepted standardized approaches to information sharing).
- CHI generally supports preserving and sharing data through established repositories, but also encourages enabling APIs to facilitate streamlined data flows. However, NIH's data management and sharing policy completely omits discussion of APIs and how NIH contemplates APIs playing a role in its sharing of data. We believe this is an oversight that NIH needs to address before its policy is finalized. We strongly encourage NIH to facilitate the use of two-way APIs for management and sharing of data.
- CHI generally supports NIH's efforts to respect the autonomy and privacy of research participants and protection of confidential data. We again urge NIH to align its policies with the efforts of other key health sector agencies (e.g., ONC, HHS' Office of the Inspector General, etc.). CHI proposes that health data transparency can be advanced through the use of three "yes/no" attestations that NIH can share answers with for research participants to ensure they make informed decisions about how the technology being used handles privacy. Such questions should be to answer whether (1) the technology conforms to Xcertia's Privacy Guidelines; (2) the technology developer attests to the Federal Trade Commission's Mobile Health App Developers: FTC Best Practices and the CARIN Alliance Code of Conduct; and (3) the technology developer attests to adopting and implementing ONC's Model Privacy Notice. NIH should publicize these attestations to promote research participants' informed decision making and transparency.

CHI appreciates the opportunity to submit its comments to NIH. We look forward to assisting NIH in modernizing and improving its data management and strategy.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

CHI Comment re NIH Draft Data Mgmt & Sharing Policy (due 011020).pdf

Description:

CHI's Comments re NIH Draft Data Management and Sharing Policy

January 10, 2020

Andrea Jackson-Dipina, Dr.PH
Director of the Division of Scientific Data Sharing Policy
Office of Science Policy
National Institute of Health
Department of Health and Human Services
6705 Rockledge Drive, Suite 750
Bethesda, Maryland 20892

RE: Comments of the Connected Health Initiative on *Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance* (84 FR 60398)

Dear Dr. Jackson-Dipina:

The Connected Health Initiative (CHI) appreciates the opportunity to provide input on the National Institutes of Health's (NIH) Draft Policy for Data Management and Sharing and Supplemental Draft Guidance, intended to promote effective and efficient data management and sharing.¹ Therefore, this draft policy and supplemental guidance will further NIH's commitment to making the results and accomplishments of the research it funds and conducts available to the public.

CHI represents a broad consensus of stakeholders across the healthcare and technology sectors whose mission is to support the responsible and secure use of connected health innovations throughout the continuum of care to improve patients' and consumers' experience and health outcomes. We advocate before the Department of Health and Human Services (HHS) on realizing the benefits of an information and communications technology-enabled American healthcare system. CHI is committed to advancing an interoperable healthcare system that enables the bidirectional flow of necessary health data between provider and patient, as well as between other important stakeholders who have a role in improving care coordination and decision-making.

The efficacy of precision medicine, population health, clinical decision support—and artificial/augmented intelligence (AI)- driven tools in particular—is dependent in large part on the availability of massive data sets. The free flow of information and interoperability are therefore important and potentially life-saving for patients.

¹ [84 Fed. Reg. 60398 \(Nov. 27, 2019\)](#).

NIH's proposed policy comes at an important time. There is no disputing that interoperability and patient access to health information prevent timely and informed care coordination and decision-making. Further, electronic health information and educational resources are critical tools that empower and engage patients in their own care regimens. CHI strongly believes that a truly interoperable eCare system includes patient engagement facilitated by store-and-forward technologies (ranging from connected medical devices to general wellness products) with open application programming interfaces (APIs) that allow the safe and secure introduction of patient-generated health data (PGHD) into electronic health records (EHRs). Data stored in standardized and structured formats with interoperability facilitated by APIs provides analytics as well as near real-time alerting capabilities. The use of platforms for data streams from multiple and diverse sources will improve the healthcare sector, helping to eliminate information silos, data blocking, and deficient patient engagement. Interoperability must not only happen between providers, but also between remote patient monitoring (RPM) products, medical devices, and EHRs. NIH's approach to data management and sharing are important for those stakeholders directly engaged with NIH, as well as to the wider healthcare community through the precedent NIH sets.

Based on the above, we provide the following viewpoints and recommendations on NIH's draft policy:

- CHI is generally supportive of NIH's efforts to update and improve its approach to data management and sharing. Specifically, we support NIH making scientific data publicly available at no (or nominal) cost in as timely a manner as possible. However, we believe that NIH's approach, as proposed, may not align with information sharing norms in the public and private sector.
- A logical, objective approach is necessary to reduce confusion, and NIH should align its data management and sharing policy with the Office of the National Coordinator for Health IT's (ONC's) information blocking to the extent possible. While this rule is currently approaching finalization, it will represent the baseline for information sharing moving forward, and NIH should align its data management and sharing policies with these rules to the maximum extent possible to provide continuity across the healthcare ecosystem. For example, CHI recommends use of the Fast Healthcare Interoperable Resources (FHIR) standard (Release 4) as well as HL7 U.S. Core FHIR Implementation Guides (or in the alternative that NIH permit the use of such widely-accepted standardized approaches to information sharing).
- CHI generally supports preserving and sharing data through established repositories, but also encourages enabling APIs to facilitate streamlined data flows. However, NIH's data management and sharing policy completely omits discussion of APIs and how NIH contemplates APIs playing a role in its sharing of data. We believe this is an oversight that NIH needs to address before its policy is finalized. We strongly encourage NIH to facilitate the use of two-way APIs for management and sharing of data.

- CHI generally supports NIH's efforts to respect the autonomy and privacy of research participants and protection of confidential data. We again urge NIH to align its policies with the efforts of other key health sector agencies (e.g., ONC, HHS' Office of the Inspector General, etc.). CHI proposes that health data transparency can be advanced through the use of three "yes/no" attestations that NIH can share answers with for research participants to ensure they make informed decisions about how the technology being used handles privacy. Such questions should be to answer whether (1) the technology conforms to Xcertia's Privacy Guidelines;² (2) the technology developer attests to the Federal Trade Commission's *Mobile Health App Developers: FTC Best Practices* and the CARIN Alliance Code of Conduct;³ and (3) the technology developer attests to adopting and implementing ONC's Model Privacy Notice.⁴ NIH should publicize these attestations to promote research participants' informed decision making and transparency.

CHI appreciates the opportunity to submit its comments to NIH. We look forward to assisting NIH in modernizing and improving its data management and strategy.

Sincerely,



Brian Scarpelli
Senior Global Policy Counsel

Alexandra McLeod
Policy Counsel

Connected Health Initiative
1401 K St NW (Ste 501)
Washington, DC 20005

² XCERTIA MHEALTH APP GUIDELINES, <https://xcertia.org/wp-content/uploads/2019/08/xcertia-guidelines-2019-final.pdf> (issued on August 12, 2019).

³ *Mobile Health App Developers: FTC Best Practices*, F.T.C., <https://www.ftc.gov/tips-advice/business-center/guidance/mobile-health-app-developers-ftc-best-practices> (issued April 2016).

⁴ *Model Privacy Notice*, ONC, <https://www.healthit.gov/topic/privacy-security-and-hipaa/model-privacy-notice-mpn>.

Submission ID: 1409

Date: 1/10/2020

Name: Emily Harris

Name of Organization: Not Applicable

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All types of data from epidemiologic studies

Type of Organization: Not Applicable

Type of Organization - Other:

Role: Member of the Public

Role - Other:

Domain of Research Most Important to You or Your Organization:

Epidemiologic research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Comments_DraftNIHPolicyDataManagement&Sharing_PublicVersion_2020-01-10.pdf

Description:

Comments on the Draft NIH Policy for Data Management and Sharing and Supplemental Guidance

Emily Harris

January 10, 2020

Pluses

- Emphasis on making data from research projects available to the broad research community
- Expecting an acceptable data management and sharing plan prior to funding
- Budget allowed for implementing the data management and sharing plan
- For extramural awards, clear statement of enforcement actions
- Encourages use of established repositories for preserving and sharing data

Concerns

- No statement about the relationship to other NIH data-sharing policies. Need to be clear about whether this policy supersedes or is in addition to other NIH data-sharing policies that are more specific:
 - Genomic Data Sharing policy
 - 2003 policy for large budget awards
- No clear expectation of data to be shared. (“NIH does not expect researchers to share all scientific data generated in a study.” [1st paragraph, supplemental guidance on plan elements]) Need to state an expectation, such as data used to address the specific aims of the funded project.
- No clear expectation of the timeline for sharing data (“...as soon as practicable, independent of award period and publication schedule.” [last bullet, section 5., supplemental guidance on plan elements]). Need to state a minimum standard, such as with {xx} months of completion of data-cleaning, at time of acceptance of a publication, or before the end of the award period, whichever is sooner.
- Not stated whether data are to be shared “without strings attached”; that is, whether collaboration or authorship on manuscripts can be required, or scientific questions or methods can be limited. Need to state these expectations specifically. Suggest making this consistent with the 2003 policy under which collaboration and/or authorship cannot be required, and questions to be addressed and methods used cannot be limited (as long as they are consistent with informed consent, for human research).
- For research using human data and/or biologic specimens, consent expectations not stated. Suggest stating an expectation that informed consent for sharing data with the broad research community is expected going forward.
- Not clear whether the data management and sharing plan can be considered as part of the funding decision. Suggest making it clear that the data sharing plan can be considered in the funding decision, including the ability to share data from human research participants.
- While Programmatic review and approval is key, requiring a data management and sharing plan at Just-in-Time is too late in the process for extramural grant applications. Suggest

January 10, 2020

1 of 2

including the plan as part of the grant application, as is currently required for resource/data sharing plans.

- Peer reviewers will consider the budget, including the budget for the data management and sharing plan. Yet they will not have a plan to review on which the budget is based, if a plan is not required as part of the grant application. This is not a good situation for review.
- Some peer reviewers have expertise in data management and sharing, or in specific aspects (such as data standards for their communities). Their input may enhance Programmatic review, if a plan is included as part of the grant application.
- Data sharing is an important consideration in making funding decisions. To include that aspect in the funding decision process, Program needs to be able to review the data management and sharing plan early on. The Just-in-Time request is too late.

Submission ID: 1410

Date: 1/10/2020

Name: Seun Ajiboye

Name of Organization: American Association for Dental Research

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of the above

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: Director, Science Policy and Government Affairs

Domain of Research Most Important to You or Your Organization:

Dental, oral and craniofacial research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

As a general philosophy, greater openness and sharing to advance discovery and enhance accountability is a worthy goal. It is also wise that sharing falls under "guidance" and not a policy requirement, as there are several circumstances under which obligated sharing would not be advisable.

Sharing NIH-funded research data should not be required with parties that have a major financial interest in the outcomes or a history of undermining scientific research, such as the tobacco industry. For instance, some AADR researchers conduct research with minors related to tobacco, and in most cases, would insist on shielding that data from a tobacco company.

Knowing that there is free and open access to data that other investigators have collected may create a disincentive to engage in generating original primary data. There must remain incentives for conducting new, original investigations, such as assurances that data will remain under control of the initial investigators for a set amount of time and/or giving investigators fairly broad control over how data are disseminated.

In general, the draft guidelines and supporting information do not speak to institutional review boards, except to say that "institutional policies" "dictate how research involving human participants should be conducted and how the scientific data derived from human participants should be used." These same policies can dictate how the data are stored, retained, and shared. The NIH policies should give greater attention to the role of institutional review boards, and state/other entity rules, that may govern data storage and sharing.

Attachment:

Description:

Submission ID: 1411

Date: 1/10/2020

Name: Council on Governmental Relations

Name of Organization: Council on Governmental Relations

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other:

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other:

Domain of Research Most Important to You or Your Organization:

Council on Governmental Relations

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Refer to uploaded letter

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

COGR NIH DMSP letter 1-10-20 final.pdf

Description:

COGR response to NIH Draft DMSP Guidance



Dr. Andrea Jackson-Dipina
Director of the Division of Scientific Data Sharing Policy
National Institutes of Health
Office of Science Policy
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

January 10, 2020

Subject: Comments to Draft NIH Policy for Data Management and Sharing

Dear Dr. Jackson-Dipina:

The Council on Governmental Relations (COGR) is an association of 188 research universities and affiliated academic medical centers and independent research institutes. COGR concerns itself with the impact of federal regulations, policies, and practices on the performance of research conducted at its member institutions.

We thank you for the opportunity to respond to the Draft NIH Policy for Data Management and Sharing and Supplemental Draft guidance (NOT-OD-20-013). COGR recognizes the importance of data sharing and generally agrees with the NIH's draft policy. However, we believe that in order to promote a culture of data-sharing across all scientific disciplines, the NIH should also prioritize developing resources and tools to better facilitate data sharing.

COGR is pleased to see the proposal to submit data management and sharing plans (DMSP) as part of the "Just-In-Time" (JIT) documentation for extramural awards. Requiring submission of the DMSP at the JIT phase rather than at the proposal stage minimizes administrative burden for both the applicant and peer reviewers. We also assume that the details of the plans will not be considered as part of merit review.

While submitting DMSPs during JIT will allow researchers more time to focus on the science being proposed, one potential drawback is that it will be challenging to accurately budget data management costs for a plan at time of application when the details will be later finalized with NIH Program staff at JIT. We therefore recommend that NIH allow additional data management costs to be added to the budget at JIT based on the final negotiated DMSP. We also recommend an option that allows grantees to appeal NIH Institute, Center, and Office (ICO) mandated data sharing requirements to the NIH Policy Office should the requirements be considered unreasonable or inappropriate by the grantees involved, without fear of reprisal.

Furthermore, because many institutional offices are involved in reviewing and approving components of DMSPs, having feedback available on the status of the NIH review of the plan for those involved in the development and review process at the institution will be extremely helpful in order to manage a plan.

We appreciate that the Draft Policy allows for necessary flexibility across scientific disciplines by outlining minimal specific expectations for the NIH-wide DMSPs. We are concerned, however, that allowing each of the 27 ICOs to create separate supplemental requirements will create confusion in the awardee community. We urge NIH to harmonize among the ICOs via the use of a consistent format for collecting the minimal requested DMSP information. A common form for metadata organization with standardized data fields would also be helpful to ensure that the same relevant metadata is obtained for each study. Consistent collection of appropriate metadata (such as conditions under which studies were conducted, information about the research subject populations, and journal citations) may also enhance aspects of reproducibility.

For efficiency purposes, we further recommend that NIH establish a centralized location to host ICO-specific requirements as opposed to individual institute websites. One central location for all NIH information pertinent to data sharing would improve transparency and monitoring of practices for both public and grantee communities.

Allowing researchers to create the specific plans applicable to their data is important to ensure that data are not made public before any security or privacy restrictions or concerns are addressed. We thus strongly recommend that NIH include in the policy or in its implementation appropriate options to address the myriad legal, ethical, technical, security, and privacy considerations that may impact data sharing. We further recommend that NIH provide resource information to help researchers and the public understand the meaning and implications of these various restrictions. Leaving the coordination of restrictions across sensitive data sets to researchers alone could add significant unfunded administrative burden.

Thinking more broadly, NIH has the unique opportunity to lead the community by creating field-specific data repositories that capture data elements and metadata that are relevant for that field and have the added benefit of ensuring that relevant security and privacy concerns are addressed. NIH-led data repositories would also allow both the agency and the awardees to leverage resources, avoid duplication and disaggregation of valuable knowledge, and curate and provide data in ways that maximize the public benefit.

We appreciate NIH's recognition of protections for scientific data generated from humans or human biospecimens and ask that NIH explicitly acknowledge the role of the Institutional Review Board (IRB) in the review and approval of DMSPs, and in ensuring that such plans are appropriately disclosed in informed consent materials. NIH may want to consider the existing NIH Genomic Data Sharing (GDS) Policy and related guidance as a model, as it provides a framework for IRB considerations such as risks associated with data sharing and evaluation of informed consent, including identification of circumstances where informed consent may not adequately address data sharing. There must be consistency between the plan and the informed consent obtained from human participants.

We also ask NIH to consider issuing guidance on standards for uncontrolled access, de-identification, application of the NIH Certificate of Confidentiality Policy, consequences of participant withdrawal and ability for a participant to decline data sharing, and how requirements such as the Health Insurance Portability and Accountability Act, the European Union General Data Protection Regulation and other data protection laws apply, especially as the data could ultimately be used for commercial purposes through uncontrolled access.

The Draft Policy indicates that *“non-compliance with the NIH ICO-approved Plan may be taken into account by the funding NIH ICO for future funding decisions for the recipient institution.”* It would be helpful to gain clarity on how non-compliance will be assessed by NIH, particularly since (i) a DMSP is by definition a *plan*, subject to change, (ii) implementation of the DMSP is dependent on the progress of the research, and (iii) the DMSP requires descriptions such as anticipated timeframes and anticipated agreements that could limit the ability to share scientific data broadly. For example, if deposited data were not yet analyzed and ready for publication, the approved DSMP is unlikely to meet the overall intent of “reproducibility”.

The ability of NIH to make a finding of non-compliance any time after the end of the funding period creates, in effect, an unlimited and perpetual compliance obligation for PIs and grantees. We therefore recommend findings of non-compliance be limited to failures to follow a DSMP or other actions related to data sharing and management *during* the funding period.

Alternatively, NIH could consider whether the policy applies to the data set that is available at the end of the funding period, or whether the data desired and requested must necessarily rely on more fully contemplated resources needed after the end of the award period. One potential solution would be to create a data sharing mechanism using modular budgeting that could be a

supplement and extension to every award – *de facto* adding a sixth year to each standard R01 or an appropriate equivalent for each funding mechanism.

The Draft Policy contains the following statement, “*NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public*”. While we appreciate that NIH encourages rather than requires this practice, as the determination of usefulness is necessarily a subjective one best made by the investigator or other experts in the same field, we are concerned NIH’s “encouragement” will be used as a factor in approving a DSMP and in determining compliance with the DSMP. We ask that this statement be removed from the Draft policy or explicitly emphasized that it is only an encouragement.

On a related point, if a repository with recurring fees is the only viable option, the grantee could find itself needing to cover those costs once the project is over, potentially for years. We ask NIH to continue to discuss the allowable costs guidance of the data sharing policy with stakeholders at future roundtable meetings or other public forums.

We note the Draft Policy applies to all scientific data generated from NIH-funded research and is written with the expectation that reasonable efforts will be made to digitize all scientific data. The February 22, 2013, [memo from OSTP](#) to departments and agencies significantly applies only to digital data. This expectation that non-digital data will be digitized creates a new, complex and potentially costly burden for NIH and grantee institutions and could serve as a disincentive to participate in research.

The Draft Policy indicates that the plans should consider the life of the scientific data, and we applaud NIH for recognizing that each scientific discipline may have different life cycles for data. However, all fields will be affected by evolution of technology, which over time will render the current hardware and software necessary for accessing data obsolete. Migrating data to be compatible with future technology will be costly. In its policy guidance, NIH should recognize that technological changes are inevitable and should not require investigators to attempt to predict such changes nor require institutions to incur such costs in the future.

The recommendation to apply this Draft Policy to *all* projects, instead of those above the current \$500K threshold, will require significant additional resources, training, and time to implement. We ask that NIH choose a policy implementation date far enough in the future to allow the grantee community to prepare sufficiently. We recommend that implementation be effective with new proposals submitted in NIH fiscal year 2021 or later, assuming NIH releases




the final Policy by March 31, 2020. A delay of at least one year for the “effective date” will benefit all parties involved by allowing sufficient time to effectively implement the Plan against the standards to be established by the ICOs. We also ask that NIH provide a reasonable embargo period for data to provide for intellectual property protection. Finally, NIH should consider clarifying that the policy does not apply to awards (or activity codes) for which no data management plan is required to be submitted as a condition of the award.

We would suggest that NIH take into account the feedback that will be received by OSTP in response to its current [Request for Information](#), particularly with respect to research rigor and reproducibility. Prior to the implementation of the policy, NIH should consider the creation of a Good Research Data Practices (similar to Good Clinical Practices) standard that addresses DMSP, including standards for data collection/design/purpose and archival standards.

Lastly, COGR recommends that NIH consider the issues and potential solutions related to data sharing raised in the publication “[Good Practices for University Open-Access Policies](#)” published by the Harvard Open Access Project. While this work was primarily aimed at open access for scholarly articles, its principles can also be applied to data sets.

Thank you for the opportunity to comment. If there are questions, please contact Jackie Bendall at jbendall@cogr.edu.

Sincerely,


Wendy D. Streit
President

Submission ID: 1412

Date: 1/10/2020

Name: John Watts

Name of Organization: Texas A&M University Libraries

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: The University Libraries support a wide range of data types.

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Librarian

Domain of Research Most Important to You or Your Organization:

The University Libraries support a wide range of research domains.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Due to the specificity of the new draft policy, it would be most reasonable to enact the requirements therein to future awards only. As with all new policies, there will be a time of adjustment for researchers and those who provide research data management and sharing infrastructure. All institutions supporting research data management will need to build capacity for data sharing in compliance with this policy within their unique institutional contexts. Therefore it is crucial to provide space between the publication of the final policy and its implementation in future funding cycles in order to allow for planning and collaboration at the institutional level. Moreover, clear language regarding the obligations of primary investigators engaged in an existing funding cycle is crucial. Librarians stand ready to support the data management and sharing requirements, but there will be a period of growth and education in order to meet the needs of researchers working under the new policy.

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Compliance and enforcement with the terms and conditions of a Plan could be cumbersome as plans are reviewed and revised in interval. What instruments or metrics will NIH ICO use to measure compliance over the life span of the funding, and how will compliance across multiple iterations of the Plan be tracked? If Plan compliance is used as a metric for future funding, the Policy should fully articulate the criteria for success for compliance and make those criteria transparent to researchers in advance of their proposal submission. It would also be beneficial to clearly articulate the NIH staff who are responsible for the initial review and ongoing compliance of the funding.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:**Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:****Other Considerations Relevant to this DRAFT Policy Proposal:**

Thank you for the opportunity to comment on the draft of the NIH Policy for Data Management and Sharing and supplemental drafts. I submit these remarks on behalf of my colleagues at the Texas A&M University Libraries: Stephanie Fulton, Sheila Green, Carolyn Jackson, Laura Sare, Christina Seeger, Robin Sewell.

Attachment:**Description:**

Submission ID: 1413

Date: 1/10/2020

Name: Dennis Dean

Name of Organization: Seven Bridges

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Other

Type of Organization - Other: Biomedical Analysis

Role: Other

Role - Other: Director, Scientific Operations

Domain of Research Most Important to You or Your Organization:

All Genomics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The data management and data sharing document adequately addresses much of the scope one might expect in a Data Management and Sharing Document. However, we at Seven Bridges do believe that there is an opportunity to extend the data management and data sharing plan to include concepts that will strengthen support for future data integration across multiple types of research structures that include research teams, consortia, and national research efforts. We believe that this is important due to ongoing trends in using large datasets generated by the collection of smaller datasets and the growing demand for datasets to support machine learning and artificial intelligence initiatives. Specifically, we suggest that the data management and sharing plans require the inclusion of additional information that defines the larger context by which data collections are completed (i.e, research is part of larger consortia effort). We expect that this information will help to guide dataset harmonization. It is imperative to collect this information prior to the start of the grant so the harmonization efforts can begin prior to the completion of data collection and so that the larger context by which data is collected is available well after the completion of the research effort.

Section II: Definitions:

The definitions provided in the policy leads to a classic data management framework that tacitly assumes data generated from a single study and generated by a single individual investigator. The ecosystem and context by which data generation occurs in practice are more

complex with the context of data collection influence by data previously collected by the investigator, as well as data previously collected by the wider research community. The motivation and source of data collection are important because it provides a framework by which data might be integrated and harmonized.

We would argue that the importance of a dataset will be increasingly driven by the features that can drive the harmonization of smaller datasets into "super-datasets." Such super-datasets could be defined as those datasets which are an amalgam of smaller datasets and could include information on how the data was collected and the research context by which the data was collected. If there are indications of where the data could be integrated into other datasets in the future, these insights could be incorporated into the construction of super-datasets. As they become more widely shared, the super-datasets then lend themselves to analysis by automated tools (namely Artificial Intelligence) to make data harmonization faster and more efficient.

Consequently, we believe there is a need for the data management policy to reflect and explicitly include indicators by which the wider context that data collection occurs. We specifically propose that data management plan include components that facilitate the communication of how data collected for a particular study might be related to research teams within and across institutions, part of a larger institute-wide effort, and/or part of a larger national or consortia effort. Furthermore, explicitly stating the 'research tier' by which data collection occurs provides an opportunity to include policy components that address data management challenges that may occur at different research tiers. For example, data generated by a laboratory trainee may reflect a specific investigation with limited potential for integration with other data sets. Contrastingly, harmonizing data collected as part of a consortium is likely to lead to powerful data sets that will lead to super-datasets that may be queried to answer questions beyond the original goals of the consortia.

To summarize: we believe that an explicit statement of the policies and responsibilities of individuals at each tier of research group will be important going forward so that data management policy requirements (1) will evolve to include each data management requirements for each tier, and (2) ensure the data management policy supports integration of individually contributed data sets into larger super-data sets that we expect will support research beyond original goals/context by which data was collected.

Section III: Scope:

No comment on Scope

Section IV: Effective Date(s):

We believe it is important to have specific dates by which data will be made available. Whereas the Cancer Moonshot committee includes aggressive timelines to make sequencing data available, we believe the community needs to go further in setting aggressive deadlines for making data available. For example, the cancer moonshot sharing plan recommends a two year embargo period. We propose ensuring a specific timeline; potentially an even more aggressive timeline especially for data that is not controlled.

We suggest that the first six (6) to twelve (12) months serves a guidance period, wherein the NIH states what the final policy will be, and gives enough time to that users can reply to the NIH, state concerns, and/or make adjustments to their current practices. After the end of the first year up until the deadline, portions of the policy should be non-negotiable and data-sharing stipulations should begin to be implemented. Some of the steps for policy implementation may require technological advancement, and these aspects of the policy should be planned to be implemented in the latter half of the two-year window. If an element of policy would require any more lead time to develop the required technology, it is not ready to be implemented in the policy. Additionally, a two year period before the policy would be implemented could have effects on the grant submission cycle for investigators.

Section V: Requirements:

Reporting of secondary data, as is commonly done in NGS analysis, requires a deep understanding of the tools/workflows required to process and generate data. Consequently, data management and data sharing policy explicitly requires this information to be shared and made available publically. Furthermore, this process of sharing workflow information will be facilitated by using emerging standards for workflow communication.

Closing the gap between practice and regulatory science is of key importance to ensuring regulatory frameworks can be grounded in reality, implemented in practice, are accessible and practical for users, and meet the intended objective of ensuring safety whilst encouraging innovation. The NIH and its advisory committees need to be involved and ever-evolving to meet the challenges of providing suitable regulatory frameworks, particularly for products that present challenging issues and new technology such as next-generation sequencing so that data collected early in the research story can be used potentially through FDA regulatory submissions. The NIH can take a leading role in helping users be confident in their validation processes and workflows. The goal is to ensure accurate diagnoses, advance drug development, ensure patient safety and appropriate care.

The setting of benchmark values and metrics by the NIH and the sharing of the rationale behind their selection would be of great benefit to the research community. Benchmarking enables the exploration of emerging applications of genomics data and sharing that are expected to impact the regulatory review process in the future. This will expand the focus of the data sharing to include more complicated workflows adapted for specific datasets, and for integration with larger datasets and the specific challenges they bring.

One suggestion of a requirement for future implementation would be for the community to adopt workflow standards, such as Common Workflow Language (CWL), and develop them further to support the unique needs of the user community. Integrating standards into research community activities will advance the goal of accelerating the adoption of standards and implementation expertise. If the NIH could provide a platform for distributing the information on emerging standards and working groups so that it was accessible to researchers, it would serve to increase the rate of which such standards are adopted and to promote collaboration within the community.

Overall, there is a need for more stringent data sharing requirements in the research community, but doing so incurs increasingly higher costs. It should be noted that depending on where a user resides on the research hierarchy (individual, research group, institution, or consortium), such high costs of sharing compliance can have varying levels of impact. For larger research groups or institutions, finances may be less of an issue, but at the level of an individual researcher or research team, such costs may be prohibitively expensive and pose a barrier to research progress. We propose separate mechanisms to facilitate data sharing among smaller research teams, and that the NIH may be required to provide funding to cover these costs. Smaller teams especially need additional funding available to make data query-accessible, for data integration, storage costs, etc, in order to comply with NIH policy moving forward. In cases where resources are scarce, there is the "throw it over the fence" approach of uploading unannotated data and not maintaining it, where it adds minimal value to the community. Data should be available and actionable, requiring proper annotation, QC, and support. Furthermore, the more stringent the policies to be implemented, the more likely it should be that the NIH should provide funding or support for these costs to researchers through grants or other funding opportunities.

Alternatively and over time, it will be incumbent of the NIH to support the development of tools integrated within the national data sharing and management infrastructure that supports and eases the burden of making data available. For example, a smaller research group may not have the resources to map their dataset to an ontology; which can be a powerful data organization tool that automatically supports advanced features such as data integration.

However, many individuals could be trained to use the next-generation data integration and harmonization tools that will both decrease the cost and time required for sharing data.

Section VI: Data Management and Sharing Plans:

An initial requirement for a Data Management and Sharing Plan should be that investigators provide a paragraph that defines how the data could be integrated with other data in the field. This would not only engage the investigator to examine how their own current research could best be integrated with other datasets but also serves as an indicator to the NIH as to how an investigator foresees their future work being shared even as policies and datasets change.

From the Seven Bridges user survey, users report they would greatly benefit from having some form of documentation of current best practices, as an alternative to the slow and inefficient process of sorting through published literature. Towards the goal of user education and guidance, data sharing plans for completed grants might be made publically available by the NIH as a way for users to understand what data is available for a completed study, to encourage sharing standardization between communities, and to begin the process of planning for integrating smaller datasets into larger datasets.

Furthermore, NIH-provided benchmarking could provide early insight into emerging areas, allowing regulators to identify emerging trends, new applications and bringing possible issues to attention. Beyond providing guidelines for data sharing and management, these benchmarks could also cover other current best practices in the field such as to promote the development of methods that characterize workflow robustness, promote approaches for verifying workflow accuracy, and to expand access to novel computational approaches for validating workflows.

Verification and validation of the data submitted by users are also of paramount importance for the good of the research community. There are tools available for quality control of the data, and we at Seven Bridges believe that the use of such tools should be another minimum requirement as mandated by this policy. QC workflows that extract quality features for NGS data and secondary analysis performed on these tools are examples. Beyond descriptions of these tools, sharing workflows in common workflow language (see above) could support replication of verification and validation methods used by individual or consortia researchers. Such tools should be made available to users by the NIH in order to promote compliance among the community.

Section VII: Compliance and Enforcement:

For the general community, dataset access could be the first line of defense for quality control, and access control is vital to protect the identity of individuals that contribute their data.

Dataset access could be controlled through a common authentication and authorization mechanism that secures the data. Seven Bridges' BioData Catalyst ecosystem, for example, manages user access to the hosted controlled data using data access approvals from the NIH Database of Genotypes and Phenotypes (dbGaP). Therefore, users who want to access one or more of the hosted controlled studies on the ecosystem must be approved for access to that study in dbGaP. Principal Investigators who have approved Data Access Requests on dbGaP for the BioData Catalyst datasets will be able to programmatically access those data on the platforms and services within the BioData Catalyst ecosystem.

This proposed policy could also be served by the NIH providing the current best practice security protocols and information to the research community. We at Seven Bridges rely on advanced security protocols and employ frequent scanning to ensure end-to-end reliability and security of all data. Users connect to the front end server(s) using a secure connection via the latest TLS. All requests between any back-end services are established using TLS again. All database connections also use SSL/TLS for encrypting the data to and from the databases. Data is encrypted throughout the complete lifecycle. Security, compliance, and consent controls have been incorporated throughout the Seven Bridges platform to support researchers effectively learning from genomic data. Because appropriate controls are enabled by default, researchers don't need to worry about configuring systems or managing consents. Data is available only to approved researchers and data use policies allow enforcement of geographical or combinatorial analysis restrictions. Finally, all actions are logged to facilitate auditing, while monitoring and alerting procedures allow for the rapid detection and resolution of any abnormal events.

Other technologies that facilitate access control such as client-encryption, two-factor authentication, integration with a client's single sign-on solution, and integration of external key management can all be configured according to the needs of the client organization to promote security and compliance. We suggest that similar mechanisms be set in place as a requirement to promote data security within the user community. The NIH could also facilitate compliance by publishing links to data or metadata repositories that facilitate access and compliance checks. The policy should also provide contingencies for when data are not made available at the end of the granting period, specific procedures and funding opportunities for getting research groups back to compliance.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Based on results of our NCI funded Cancer Genomics Cloud users survey, we were surprised to find the majority of users are not aware of or planning for the costs of cloud computing. The most valuable improvement would be usage and cost monitoring for their projects. Specifically, oft-requested features include email warnings of data limits, ability to view costs per project,

and the ability to view costs per file. As such, without the proper knowledge of such costs beforehand, users may have a difficult time estimating budgets or funding requirements for their projects. A related suggestion for allowable costs for this proposal would be for the creation of data sharing standards across multiple institutions based on emerging needs. Having the typical values for project and data costs across the community would benefit individual and/or new researchers who have no basis for comparison of such costs for their own projects.

Another suggestion for allowable costs would be for creating a "super-dataset" (defined above) that integrates data across multiple investigators. Such a dataset would have obvious value to the research community, but the financial and time costs of such an endeavor would be non-trivial. Another suggestion to mitigate this issue would be to amortize this cost across multiple investigators contributing to the dataset. Lastly, a final suggestion for allowable costs would be to include opportunities for additional funding for researchers to make their data query accessible, for integration, for storage costs, etc. Implementing "one-click" sharing, possibly as a part of data submission or dataset hosting, would ease the burden of effort on the researcher while also promoting compliance. Developing a national data repository infrastructure likewise would assist in their endeavor. The success of individual researchers in sharing their data is incumbent on the national infrastructure developed by government agencies and potentially industry will need to collaborate on mechanisms that reduce the overhead required to make data publically available.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

The current open-ended nature of the proposed policy leaves it up to the investigators to determine their own timeframe and methodologies for data management and sharing. While this provides individuals with a great degree of agency over their data, there is a concern that if there is not a movement towards a unified set of standards presently, the time and monetary costs of standardization, harmonization, and integration could increase as time goes on, as datasets increase in size and complexity, etc. Consequently, we specifically propose that data sharing plans include suggested time frames for making data available should be provided within the policy (i.e, data will be made publicly available within two years from collection or end of grant).

We propose that this policy should require standards to ensure data can be harmonized efficiently. There need to be incentives for using policy-established standards and providing guidance for integrating submitted data with other and larger datasets. These requirements should also reflect the diversity in research groups and their needs, ranging from individual researchers to institutions to consortia. Data management may be a larger load for individual researchers and small lab groups who lack the hardware or financial resources of major

institutions. Institutional and funding agency support may be required to release high-quality datasets from these smaller groups.

Towards the aim of standardization, we would propose supporting specific standard recommendations when standards are well established. For example, the NCI efforts resulted in the development of at least two workflow languages (WDL and CWL) that are now commonly used. Consequently, for the documentation of workflow language, the NCI might recommend using these two languages when representing workflows, when possible.

Beyond the policy, developing the infrastructure that supports the sharing of data and the tools required to process them is also vital. The cloud pilots funded by the NCI are a good example of an ecosystem designed to promote data sharing and analysis. For example, Seven Bridges has developed an extensive Software Development Kit (SDK) that allows us to wrap tools and create workflows with a graphical user interface. All tools are installed and executed inside Docker containers. This allows flexible porting and re-use of workflows or workflow components, something that is not possible with monolithic virtual machines. For working with Common Workflow Language (CWL), we have also created an open-source and locally installed SDK, called Rabix. Rabix allows users to design and execute CWL locally in a scalable and portable way that enables reliable and reproducible analysis across a wide variety of environments, ranging from laptops to cloud infrastructures to high-performance computing (HPC) clusters. Other workflow description languages that have been widely implemented by the research community such as Workflow Description Language (WDL) or Nextflow are also potential candidates for a standardized language set by the NIH.

In addition to standard guidelines for data management, the policy would also benefit by making default timelines for making data accessible. By setting a default, this hopefully will provide a benchmark value for the research community, against which research groups to compare their own workflows and gain better insight on what is considered a reasonable timeframe for sharing.

Lastly, the development of an "External Data Management Oversight Committee" would provide a much-needed service to ensure best practices for data management and sharing are being met by investigators, institutions, and consortia. This committee would ensure that someone is checking that the data developed with government funding is actually available to the research community. There are many examples of oversight programs in place currently, such as database tracking to manage intellectual property and to ensure research compliance with standards such and those set by the Institutional Animal Care and Use Committee (IACUC),

and we feel that the creation of such an External Data Management Oversight Committee would be a reasonable addition.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1414

Date: 1/10/2020

Name: Duke Office of Scientific Integrity

Name of Organization: Duke University

Type of Data of Primary Interest:

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Institutional Office

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Comment 1: The purpose section conveys a strong statement of NIH interest in promoting data sharing with the goal of improving reproducibility, the validity of research findings, strengthening analyses, etc. and, therefore, increasing the integrity of scientific research. We suggest adding a brief statement that also conveys the second, perhaps more implicit goal of sharing data, to provide reliable information and data to the greater public which in turn builds public trust in science and its service to society.

Comment 2: The second paragraph in the purpose section states the expectation that researchers and research entities will provide a Data Management and Sharing Plan (DMP). Under this DMP, "shared data will be made accessible in a timely manner for use by the research community and the broader public". Although having non-specific, broad language in this introduction is understandable, we suggest adding "timely manner" to the list of "Definitions" in Section II to allow for a more detailed and applicable definition of the term.

Section II: Definitions:

Comment 1: Under the "Scientific Data" definition, the policy states: "NIH expects that reasonable efforts will be made to digitize all scientific data." It would be helpful if specific examples could be provided to better explain what is meant by "reasonable effort." Will an investigator's management plan be scored less favorably because their research group maintains paper records, if preferred or needed?

Comment 2: Under the Data Management definition, the policy states: "A plan describing how scientific data will be managed, preserved, and shared with others." We suggest that the policy on data management also includes "documentation," such that a plan describes how both the data and the records/notes/documentation are managed, preserved, and shared. We also suggest that the policy recommends including documentation of how data are transformed, cleaned, and analyzed in the plans.

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Comment 1: We are concerned about ensuring that there are sufficient funds for both the research and the materials and costs required for proper data curation and preservation.

With the Data Management and Sharing Plan being submitted as part of the "Just-in-Time for extramural awards, contracts, etc," will items budgeted for data management and sharing come out of the total grant budget? Will the NIH add additional funds to awards to supplement the additional costs incurred by specific data management and sharing plans?

Comment 2: "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." Because this length of time is almost impossible to accurately assess, does the NIH want all investigators to deposit their "scientific data" into an institutional or organizational repository so that data stewardship may be maintained long after the investigator has completed the data collection and analysis? For example, would an investigator be penalized for a guarantee that their data be shared for 5 years, but not afterward? We suggest that this policy include an annex which will explain how investigators will be evaluated on the length of time they are able to commit to making their data available, if not stored in an institutional repository.

Comment 3: "NIH may make Plans publicly available." How would this process be implemented securely? Would the plans be individually attached to the grants through <https://projectreporter.nih.gov/>? Plans may include specific network names or algorithms that, if shared openly, could compromise the security of the data.

Comment 4: Will the NIH make specific mention of exemptions of data sharing requirements when there is combined private/public funding mechanisms? (i.e. possible intellectual property related data that the private funder would not want shared to the public)

Section VII: Compliance and Enforcement:

Comment 1: How will the NIH ICO assess the data management and sharing plans? Will the investigator be given an opportunity to submit a Plan more than once for review or appeal decisions and ask for additional considerations based on specific circumstances? The NIH ICO should clearly explain their review process and whether the investigator will have an opportunity to re-write or amend their Plan in order to meet the reviewers' criteria, especially in the beginning of the implementation of this new policy.

Comment 2: Will the NIH audit implementation of data management at the organizational level of each grantee (similar to current financial audit requirements)?

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Comment 1: Section 1, "Curating data and developing supporting documentation". In Section 1, it will be important to mention "containerization" as part of the data curation process. Containerization is an important practice for encapsulating or packaging up software code/versions in a way that allows the program to be run on any infrastructure.

Comment 2: Section 3, "Local data management considerations". In reference to Section 3, how should an investigator differentiate between "local data management considerations" and "regular costs of conducting research"? For example, would the cost of an electronic research notebook license be allowed in this budget as a local data management requirement? Or would the cost of data storage (hard drive/cloud) for each key personnel on the grant be allowable as a local data management requirement? We ask these questions, because in order to maintain electronic records, access to a computer will be necessary for every key personnel involved in data collection and analysis.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Comment 1: In the introductory paragraph, it is important that the NIH notes that it "does not expect researchers to share all scientific data generated in a study." However, this then leads to the question of how the NIH CIO will be "evaluating" an investigator's Data Management Plan and Sharing Plan. Will they be evaluated on the intentions and justifications for the Plan? (i.e. If the NIH CIO infers from the Plan that the investigator is making a good faith effort to share and manage as much meaningful scientific data as possible in a responsible manner, and adequately justifies what cannot be reasonably shared, and can report adherence to the Plan over time, then that is considered a sufficient Plan?)

Comment 2: Under either Element 1 (data type) or Element 2 (related tools, software, code), we suggest that this supplemental guidance provide an organizational data table example that includes data type, data size, volume of files, data source, related software, etc as a useful template for organizing and documenting this information as a list.

Comment 3: Under Element 6 (oversight of data management), it would be useful to offer some specific guidance about how to document the specific roles and responsibilities of the different members of the research team. In our institution's Data Management guidance document, we included an example of a RACI chart (accountability matrix) where each team member is listed as responsible, accountable, consulted, or informed on the various data management tasks.

Other Considerations Relevant to this DRAFT Policy Proposal:

Comment 1: In general, the proposed policy will promote the documentation of - and justification for - an investigator's current data management and sharing processes and requires self-reporting of an investigator's adherence to those documented intentions. Is it accurate to summarize (i.e. are we understanding this document correctly) the contents of this policy as follows: this policy establishes clear definitions for the elements of a data management plan and important considerations in data sharing, but allows the investigator a fair amount of individual and discipline-specific discretion and best judgement to discern which data will be shared, when the data will be shared, and how the data will be shared, if the generated data are outside of the data-sets already required to be shared by NIH's current data sharing policies (see comment 2 below, with three examples of required data sharing policies from the NIH)?

Comment 2: Within these three draft documents, there was no mention of how this policy will affect/integrate with other NIH sharing policies, such as Genomics data (<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html>) , transgenic model organisms (<https://grants.nih.gov/grants/guide/notice-files/not-od-04-042.html>), and clinical trial data (<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-149.html>) - except for the explicit mention that this policy will replace the 2003 NIH Sharing Policy. We suggest the NIH try and combine some of these elements into one policy with different subsections or at least reference the other policies with links so that it's more clearly understood that all policies are still in place, and their implementation should be integrated into one coherent data management and sharing plan for each NIH funded project.

Comment 3: Overall, the draft policy and supplemental guidance are well written and will help to guide investigators in terms of the scope of the data management and sharing plan. However, providing additional discipline specific (or data-type specific) examples of acceptable and unacceptable plans will be very helpful to clarify NIH expectations and will ensure appropriate levels of detail and selection of data sharing modalities within the 2-page space allotted.

Attachment:

Description:

Submission ID: 1415

Date: 1/10/2020

Name: Anthony Carvalloza

Name of Organization: The Rockefeller University

Type of Data of Primary Interest: Basic Biomedical (e.g. biochemistry)

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Chief Information Officer

Domain of Research Most Important to You or Your Organization:

Chemical and Structural Biology, Immunology, Genomics, Neurosciences, Regenerative Medicine and Aging

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

The NIH's draft guidance proposes that the submission of a Data Management Plan would be during the JIT (Just-In-Time) phase of the application. We believe that the inclusion of this step at an earlier stage in the process would allow concurrent planning with experimental design and facilitate better time management in the process of planning. Especially important is that the Plan be completed in tandem with budgeting rather than after, as the process is proposed in the draft guidance. Internally run at RU is the Rockefeller University Press (RUP), which publishes Journal of Cell Biology (JCB), Journal of Experimental Medicine (JEM), and Journal of General Physiology (JGP) and co-publishes Life Science Alliance (LSA). RUP has already begun adopting many of the proposed measures of the draft policy, particularly a push towards requirement for relevant data and supplemental information to be made publicly accessible simultaneously with publications. This is being done through the actions of requiring the source

code for all custom computational methods, the inclusion of accession numbers in manuscripts, and inclusion of robust linking to data available in public datasets. This will be further supported internally at RU via the library's forthcoming launch of DMPTool services as well as DOI minting services. Further specification is needed in the forthcoming policy regarding where NIH funded data will be housed and what will be done to facilitate or support connecting researchers to repositories. We recommend that NIH create and maintain a meta index of domain specific data repositories to act as a reference guide, better directing researchers to where their data would be most appropriately housed. We also request more explicit language in the official policy regarding publication of plans, with clear parameters of what will or will not be published along with the recommendation that consideration be given to modifiability, minability, and public accessibility of data management plans, and allowing the opportunity to link published articles to data management plans, which RUP stands ready to implement.

Section VII: Compliance and Enforcement:

We propose designation of an individual available at each institution who is trained in how to structure and maintain adherence to good data management under the new policy in addition to the proposed guidance by NIH provided during the regular reporting intervals (e.g. RPPR). This role would be given within an institution to an individual able to coordinate data management planning between PIs, Grants Offices, and NIH representatives. We propose that this individual be designated by name in the grant application. Additionally, the forthcoming policy needs greater clarification of what qualifies as non-compliance.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We propose that the allowed costs be more flexible. As proposed, they will cover some established repositories, but internal infrastructural support is not adequately included. As we push for better data management practices, the cost behind these supported procedures must gain transparency as well, and the early consideration of infrastructural support – as well as acknowledgement of the importance and necessity of said support – needs to be considered fundamental to the formation of data management plans in grant applications. These are costs that exist regardless of how they are considered in the allocation of funding from a grant, making the inclusion of and transparency of infrastructural support costs in tandem with data management planning essential. Also, it is requested for consideration that a minimum cost towards publishing be provided to ensure this need be met by researchers without sacrificing quality of work.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Data management plans will still be largely freeform under the proposed policy. It is our suggestion that additional guidance be provided so that applicants and institutions have a clear sense of what is required from them in constructing a plan and what is useful to consider for good data management. An explicit template of required elements would be a useful resource in addition to the suggested elements provided in the supplemental material of the draft policy.

While there are many advantages to a freeform approach, a more structured approach would help guide applicants towards creating more useful, sustainable, and consistent Data Management Plans. At RU, we are in the process of preparing for launch our own developed templates and guidelines via DMPTool and are prepared to coordinate these developments with the official policy from NIH and any associated materials for data management that may be provided. It should be noted that such tools have been developed by numerous institutions and are becoming increasingly standard, signifying that the trend should be met at NIH and the current materials from existing sources (e.g. Alfred P. Sloan Grant Application Guidelines) could be useful for consideration in the formation of similar resources. If provided by NIH, such resources could be more widely accessible than institution-specific guidelines. In the event that a researcher is performing research under multiple grants, it should be recommended that whichever data management plan required for each grant is most detailed be the default chosen and be accepted as the plan for the NIH application as well, rather than potentially requiring multiple formats of a data management plan to comply with varying requirements.

Other Considerations Relevant to this DRAFT Policy Proposal:

We consider it an oversight in the proposed draft that there is a focus on data without consideration of methodology. We believe methodology should be treated as data, and as such be standardized accordingly with the purpose of improving reproducibility. For consideration we offer the practice of Rockefeller University Press and their publications that there is no imposed limit on the length of materials and methods sections, supporting the underlying goal of optimizing reproducibility, in accordance with NIH's Principles and Guidelines for Reporting Preclinical Research. RUP journals are signatories of NIH's Principles and Guidelines for Reporting Preclinical Research. We also recommend the publication of methods in a ubiquitous and easily reproduced format, such as with Docker containers.

Attachment:

Description:

Submission ID: 1416

Date: 1/10/2020

Name: Molly Timko

Name of Organization: Hugo W. Moser Research Institute at Kennedy Krieger, Inc.

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of the above

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

cognitive neuroscience

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Kennedy Krieger Institute (Kennedy Krieger) is an internationally recognized institution dedicated to improving the lives of children and young adults with pediatric developmental disabilities and disorders of the brain, spinal cord and musculoskeletal system, through patient care, special education, research, and professional training.

Kennedy Krieger supports the NIH's data sharing initiatives and agrees that shared data should be made accessible in a timely manner for use by the research community and the broader public, as appropriate.

Section II: Definitions:

Section III: Scope:

The Draft Policy "applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data."

It is unclear whether this Policy would apply to new data that is generated after the NIH funding period has ended. It is also unclear whether this Policy applies to sharing data with for-profit and foreign entities. Kennedy Krieger recommends that the NIH provide additional clarification in these areas.

Section IV: Effective Date(s):**Section V: Requirements:**

It is unclear whether there is a required "minimum amount of data" to be shared in a Data Management and Sharing Plan. Kennedy Krieger recommends that the NIH indicate whether there a minimum amount of data must be shared.

Section VI: Data Management and Sharing Plans:

The Draft NIH Policy provides that researchers with NIH-funded or conducted research projects resulting in the generation of scientific data are required to submit a Plan to the funding NIH Institute and Center Operations (ICO) as part of Just-in-Time for extramural awards, as part of the technical evaluation for contracts, as part of the NIH Intramural Annual Report, or prior to release of funds for other funding agreements.

Kennedy Krieger recommends that the NIH provide clarity about the timing of any executed data sharing agreement in a multicenter study. Applicants may receive as little as 10 days to respond to a Just-in-Time request. If a fully executed agreement is due at Just-in-Time, this would likely be an insufficient period of time to negotiate and finalize such an agreement. Alternatively, Kennedy Krieger suggests that the data sharing agreement could be required at time of application, or that it be a required task to complete following the issuance of the Notice of Award.

Section VII: Compliance and Enforcement:

Advances in technology could alter the ability to effectively and/or permanently de-identify data.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

This Draft Guidance outlines potential categories of allowable costs. However, applicants may have difficulty with respect to accurately predicting such costs. For example, long term data preservation costs and licensing or repository fees may increase over time. Advances in technology could alter the originally contemplated Data Sharing and Management Plan and costs associated with that Plan.

In addition, applicants may have difficulty predicting the appropriate amount of resources that would be required to facilitate responsible data sharing as well as selecting the appropriate technology to facilitate responsible and ethical data sharing.

Kennedy Krieger recommends that the NIH provide guidance with respect to best practices for how data should be managed and stored and that such guidance be updated regularly based on current technology.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Under Data Sharing Agreements: Licenses, and Other Use Limitations (Clause 5, page 3), in describing proposed plans for managing data sharing agreements, the NIH encourages applicants to consider "[h]ow relevant limitations to sharing that are consistent with community expectations." However, Kennedy Krieger notes that "community expectations" is a vague term, which should be clearly defined in the Policy to avoid ambiguity.

Other Considerations Relevant to this DRAFT Policy Proposal:

Kennedy Krieger recommends that the NIH provide guidance on best practices for responsible data sharing in accordance with applicable privacy, confidentiality, and export control standards (e.g., mitigating risk of foreign influence on research integrity including risk of data misappropriation and diversion of intellectual property).

Attachment:

Description:

Submission ID: 1417

Date: 1/10/2020

Name: Damien Croteau-Chonka

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

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Genetics

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I agree with the comments of Michael Hoffman .

Section II: Definitions:

I agree with the comments of Michael Hoffman .

Section III: Scope:

I agree with the comments of Michael Hoffman .

Section IV: Effective Date(s):

I agree with the comments of Michael Hoffman .

Section V: Requirements:

I agree with the comments of Michael Hoffman .

Section VI: Data Management and Sharing Plans:

I agree with the comments of Michael Hoffman .

Section VII: Compliance and Enforcement:

I agree with the comments of Michael Hoffman .

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

I agree with the comments of Michael Hoffman .

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

I agree with the comments of Michael Hoffman .

Other Considerations Relevant to this DRAFT Policy Proposal:

I agree with the comments of Michael Hoffman .

Attachment:**Description:**

Submission ID: 1418

Date: 1/10/2020

Name: Amonida Zadissa

Name of Organization: European Bioinformatics Institute (EMBL-EBI)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All of the specified fields

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Other

Role - Other: Senior Scientific Services Officer

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

EMBL-EBI_response_to_DRAFT_NIH_Policy_for_Data_Management_and_Sharing.pdf

Description:

EMBL-EBI response to the DRAFT NIH Data Management Policy

EMBL-EBI response to DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

The European Bioinformatics Institute (EMBL-EBI) is the premier European centre for services and research in bioinformatics. EMBL-EBI data resources cover the entire range of biological sciences from raw DNA sequences to curated proteins, chemicals, structures, systems, pathways, ontologies and literature. EMBL-EBI's mission and its mandate is to make its tools and infrastructure freely available to the global scientific community.

One of the primary missions of EMBL-EBI is to collect, organise, add value, and make available biomolecular science data to the global life sciences community. This includes data resources that receive primary data depositions as well as curated-knowledge resources. A fundamental tenet of this mission is that all hosted data, tools and infrastructure are freely available worldwide, and that data is represented in and shared in a variety of structured and standard formats for consumption by both people and machines.

Section I: Purpose (limit: 8000 characters):

EMBL-EBI works continuously to make its data resources FAIR and we strongly support the NIH's leading strategic planning for data management and data sharing and welcome this effort to promote sustained and systematic use, management and sharing of research data, specifically following the FAIR principles.

The scientific, economic, and societal value of scientific research is maximised if the outputs of that research are preserved and made available for reuse by other researchers. All data generated by public funding should, in principle, be preserved and made publicly available and open access for future use. Without EMBL-EBI data resources are typically delivered in global partnership and with mandates from research communities. Indeed, NIH is a frequent partner in building and funding these resources. Public repositories will promote open data, provide reference data for the research community and allow for new scientific discoveries to be made with existing data.

Section II: Definitions (limit: 8000 characters):

EMBL-EBI fully supports the listed definitions.

EMBL-EBI support the proposed broad scope of the policy in applying to all research, funded by NIH, that results in the generation of scientific data.

The ELIXIR infrastructure in Europe (of which EMBL-EBI is a founding member) has, through a careful evaluation process, identified many of the EMBL-EBI data resources¹ including, for example, European Nucleotide Archive², PRIDE³ and EGA⁴, as the required deposition databases for different research communities. These data resources have been identified to have fundamental importance to the scientific community. EMBL-EBI strongly supports this effort of formalisation and identification of publicly available resources and encourages NIH to promote such data resources.

Data resources for emerging data types, such as biological images, are also arising with the aim of serving that particular community. The mission of the newly launched BioImage Archive⁵ at EMBL-EBI is to make available biological image data, of all scales, from molecules to entire organisms.

However, if no public data resource is available for deposition other provision should be made. For example, the BioStudies⁶ data resource at EMBL-EBI hosts data for which there is no established public repository. NIH should encourage researchers to deposit as much research data as possible in public repositories rather than creating bespoke local solutions within their institutions. BioStudies is in addition a recognised ELIXIR deposition database.

Section IV: Effective Date(s) (limit: 8000 characters):

We would recommend that relevant NIH stakeholders and funding recipients are made aware of the policy, along with any updates to the policy, as early as possible to support the funding applicants in their data management policy compliance.

Section V: Requirements (limit: 8000 characters):

EMBL-EBI is one of the key providers of well-established, open community repositories following FAIR principles (for example, providing unique identifiers and specifying metadata) to maximise data computation and reuse. The importance of these databases is highlighted by leading journal data deposition policies that specify in which databases certain data types should be deposited, in order for the publication to be accepted. Scientists submitting data to these resources are supported in making their data FAIR by the submission requirements of the database. Following the FAIR principles, this makes data from publicly funded research more discoverable and reusable, allowing easier

¹ <https://elixir-europe.org/platforms/data/elixir-deposition-databases>

² <https://www.ebi.ac.uk/ena>

³ <https://www.ebi.ac.uk/pride/archive/>

⁴ <https://ega-archive.org/>

⁵ <https://www.ebi.ac.uk/bioimage-archive/>

⁶ <https://www.ebi.ac.uk/biostudies/studies>

Section VI: Data Management and Sharing Plans (limit: 8000 characters):

We fully support the NIH encouragement that scientific data be made available to the research community and the public for as long as it is deemed useful by the scientific community. EMBL-EBI is committed to long-term sustainability of its data resources through a process of life cycle management of each individual data resource.

In addition, sharing and re-use of valuable scientific datasets, as defined by the FAIR principles, fundamentally advances and enables scientific discovery. Therefore, we recommend that researchers that publish the results of their scientific work as research papers are also encouraged to submit the complete associated datasets, and not just the subset of the data analysed in the publication, to the appropriate repositories.

We also welcome the prioritisation of human data and the emphasis on securing all identifiable data. As long as human identifiable data is protected, sharing the generated scientific data should not be hindered solely because they were derived from human participants. The European Genome-phenome Archive (EGA) provides the necessary security required to control access, and maintain patient confidentiality, while providing access to those researchers and clinicians authorised to view the data.

Section VII: Compliance and Enforcement (limit: 8000 characters):

Although there are references to other NIH policies, there is no explicit information about potential consequences if the policy is not implemented, or how the efficacy of the policy will be monitored.

Additionally, to support the researchers complying with the policy, NIH may wish to consider funding data scientists who can advise the submitting researchers on how to manage and publish their data, as part of the data management plan. Such support will also encourage data sharing within the scientific community. This is especially crucial in projects with a focus on the discovery of rare events, such as rare diseases. In the same way for example, clinical metabolomics only becomes most powerful and useful when the complete large datasets are shared so that they may be reused by the wider metabolomics community.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing (limit: 8000 characters):

We welcome this supplementary guidance regarding allowable costs; however we are concerned that this initiative may divert researchers away from using publicly funded

Curating data and developing supporting documentation would be best achieved by a direct collaboration between the submitting scientist(s) and the deposition data resource that will have the necessary expertise specifically associated with the data.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (limit: 8000 characters):

Please see comments to specific sections below.

2. Related Tools, Software and/or Code - Consideration should be given that if very specific tools or software packages are required there may be a risk that data will not be shared, especially if the tools are only commercially available.

3. Standards - As science evolves so do the data standards, potentially making a policy out of date if it lists specific standards or versions of standards. We would recommend that instead, the document contains a list of key community data resources, which are usually at the hub of implementing community standards. Providing guidance around key data resources may be a better way to make the policy more robust for the future.

4. Data Preservation, Access, and Associated Timelines - Life cycle management of data preservation is crucial and there are many established deposition repositories that are expert in making data available over time, so their usage should be encouraged by the policy. The way in which repositories hosted at EMBL-EBI are managed ensures that the deposited data are available as long as they are needed and in appropriate formats so the data can be freely reused.

Within the greater scientific community, it is now widely accepted that established scientific journals request that all published data is deposited with the appropriate repository at the time on paper submission. Many will not even consider publications that have no associated recognised identifiers from a public repository attached to the data. In the same way, as a funding agency NIH may consider putting similar constraints on researchers applying for NIH funded grants.

Other Considerations Relevant to this DRAFT Policy Proposal (limit: 8000 characters):

All data generated by public funding should, in principle, be preserved and made publicly available and open access for future use. Open access data resources provide tremendous value to scientists, to funders, and to the public by making the deposited data available for re-use and analysis by the scientists worldwide. Free availability, ease of access, and multiple access points are all critically important to maximise the utility and re-use of scientific data. Free and open access to scientific data produced by NIH

Submission ID: 1419

Date: 1/10/2020

Name: Sarah Nelson, on behalf of the UW Genetic Analysis Center

Name of Organization: University of Washington Genetic Analysis Center

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

genetic epidemiology, biomedical research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We, the Genetic Analysis Center at the University of Washington Biostatistics Department, commend NIH for drafting this Policy that will require awardees to prospectively articulate, plan, and budget for the specific steps to make their research products as broadly available as possible. As the draft Policy states, this will serve the purposes of promoting transparent, reproducible, and responsible research and adherence to existing NIH sharing policies and guidance (<https://grants.nih.gov/policy/sharing.htm>).

An important aspect of good data management and stewardship practices that is not currently addressed in the draft Policy is respecting research participants' consents. We recommend adding on to the sentence starting "NIH emphasizes the importance of good data management practices, which provide..." the following: "and encourage data stewardship and governance practices that respect participant consent and enhance researchers' relationships with participants and communities."

We also recommend that this opening Purpose section clearly place data processing methods (software and code) in the purview of this Policy. While software and code developed in the course of research are "outputs" of research, and thus broadly referenced in the opening

paragraph, the second paragraph's focus on the management and sharing of "scientific data" does not appear to encompass the software and code developed (and perhaps necessary) for operating on said data.

Section II: Definitions:

We recognize the conceptions and definitions of "scientific data" and "metadata" vary across scientific fields and contexts, and therefore across different types of research conducted and funded by NIH. This clearly makes it difficult to craft comprehensive and satisfying definitions for the purposes of this draft Policy. However, further clarifying these definitions seems crucial for stipulating the types of data/products/materials that are to be covered in Data Management Plans.

In our experience working with biomedical human genetic studies, we typically consider "data" to be any observation or measurement at the level of individual participants (e.g. a matrix of genotypes by participant, or phenotypic variables by participant), whereas metadata is that which describes or further documents the individual-level data (e.g., a data dictionary for those matrices, or a survey instrument used to collect those phenotypic observations). This distinction aligns nicely with levels of access --- i.e., in dbGaP, metadata as we have described is open access/publicly available whereas the individual-level "data" is controlled access (<https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html>).

The definitions in the draft Policy therefore raised the following specific questions for us: (1) Why are "outcome measures" and "phenotype observational variables" under the definition for "metadata"? (2) Are software or code written for analyzing or processing data considered either "metadata" or "scientific data" under these definitions? (3) Is "scientific data" meant to encompass "metadata"? If the former is what is "necessary to validate and replicate research findings," then presumably metadata should meet that definition. Note this is inconsistent with the definition of metadata as written: "data describing scientific data." For example, to reproduce or extend an analysis, you need definitions of the variables contained in the individual-level data, definitions such as would be found in a data dictionary and thus what we'd consider to be "metadata."

We also recommend adding to the definition of "Data Management and Sharing Plan (Plan)" in this section the detail that Plans are to be two pages or less. Currently this information seems to only appear in the "Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)."

Section III: Scope:

Section IV: Effective Date(s):**Section V: Requirements:****Section VI: Data Management and Sharing Plans:****Section VII: Compliance and Enforcement:****Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:**

As a coordinating and analysis center for several large-scale, NIH-funded human genetic consortia, we have led the cleaning, curation, and deposition of more than 100 studies into dbGaP. Managing, documenting, and providing high-quality data to repositories such as dbGaP requires considerable time and effort from investigators. Therefore, we are encouraged to see these activities included in allowable costs under this Policy.

One question raised for us in this supplemental guidance is how a proposed cost will be determined as "reasonable." The scale and complexity of a study, along with the extent of researchers' prior experiences with these activities, will determine the level of effort involved. Based on our experience, we estimate that \$60,000 is a lower bound cost of cleaning, curating, documenting, and depositing one study into dbGaP.

A category of allowable costs that we do not see explicitly mentioned is that of local storage and preservation, i.e. when some or all of the data cannot be transferred to an external repository. Will such local storage costs be allowed, and for what duration? It is not unusual for a research group to need access to data beyond the generative funding award, i.e. to support publications and/or analyses that are still ongoing at other institutions.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

As in the draft policy document, privacy and confidentiality are mentioned, but the word "consent" is not. For research involving human participants, the Plan should explicitly discuss respecting and managing participant consents in the context of data management and data sharing. Relevant questions that need to be considered include: (1) Who decides appropriate use(s) of participant data? (2) Who controls access? (3) How are updates to consent managed and propagated to potential downstream secondary users?

We strongly recommend that software and code be added to section 4, regarding long-term preservation. Software and code may be both necessary to use the data but also represent, on their own, a valuable research product for which sharing and management plans should also be articulated.

This supplemental guidance also raised a question for us whether NIH wishes to provide a Plan template. Given the current guidance document, many may choose to structure their Plan according to the same six headings. Absent a requirement or suggestion to do that, however, others will choose any variety of structure or format. To make these Plans easier to write and review, we suggest NIH recommend a structure and/or provide a template. Minimally clarifying whether the headers of the guidance document are meant to suggest a structure would be helpful.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1420

Date: 1/10/2020

Name: Abigail Goben

Name of Organization: University of Illinois at Chicago University Library

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All aspects of Biomedical research

Type of Organization: University

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

Biomedical

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

As datasets have become more common place as scholarly objects, recognized for their own importance and allowed as progress materials for grant submissions, in Biosketches, and for progress reports, we are very supportive of expanding the requirement of a data management plan for all NIH funded research which will generate data. As the policy is implemented and progress reports are received with datasets, we encourage the requirement of fixed identifiers such as ARKs/URIs/DOIs in order to facilitate access and discovery.

- One concern in the Purpose section is the inclusion of the FAIR Principles. While we agree with the principles as they stand in 2020, we recognize that they are driven by one specific community and may evolve over time. These principles, additionally, are not standards and this creates potential confusion in the policy. We recommend that the NIH pursue collaborations to develop standards which will detail the need for interoperability and reusability of data, working with known non-profits such as the W3C Consortium, FairSharing.org, and disciplinary organizations. Creating these standards may increase machine discoverability of data for re-analysis and reuse as well as application of additional semantic web techniques for computational access.

Section II: Definitions:

- A current specification of the lab notebook as "not data" is problematic – particularly as we continue to see the increased adoption of electronic solutions both in industry and academia. The rise of proprietary software that are storing raw data and which often do not provide a clear or standard mechanism for migration to another software solution without complete re-entry or loss of structure is a growing challenge. While it may not be appropriate to address in this policy, it is likely to prove an important issue in the future and we recommend that the NIH investigate the enhanced standards and preservation guidance for physical and digital lab notebooks.

Section III: Scope:

- It is unclear how or if this will relate to training grants. We would appreciate clarification.

Section IV: Effective Date(s):

- We appreciate a prospective date and implementation plan. We encourage the NIH to not delay on this date, but to make it effective within a year following the adoption of this policy.

Section V: Requirements:

- While we recognize that human subject data has particular sensitivity, particularly for minority and vulnerable populations or rare diseases, we are concerned that the requirements not be abused by researchers unwilling to share in order to mask insufficient data management practices or to prevent other researchers reusing their data. A default statement of that research is based on human subjects and therefore cannot be shared should not be allowed. Similarly, researchers should adopt informed consent documentation that allows for de-identified sharing explicitly. Researchers should be actively discouraged from defaulting to "only the immediate research team can have access to the data" for human subject research. The default expectation of researchers going forward should to provide a mechanism for sharing and reuse with detailed justification required in order to restrict data access.
- Where data cannot be openly shared, for which there are several security and privacy reasons, researchers should be encouraged to provide a metadata record indicating where the data is stored and the responsible parties to arrange access. This would first and foremost respect the human subjects or other sensitive data which may be included in the dataset but allows for a mechanism of discovery where data could be reused through appropriate review channels.

Section VI: Data Management and Sharing Plans:

- Within this section, the policy is written with great amounts of flexibility. While the intention is to avoid proscription, it instead will inhibit data sharing as researchers are left without specific and ongoing expectations. Examples include "when justified" and "as long as it is

deemed useful." This should not be determined by an individual researcher but should be developed and more clearly addressed by the ICOs in collaboration with disciplines. It will prove difficult for researchers to appropriately plan for long term data sharing without a definition of long-term to assist them in format and budgeting requirements.

- Due to the volume of identifiable data being aggregated through NIH research, it is important that the NIH invest in software, education, and mechanisms for a more robust de-identification pipeline. While researchers face the obligation to share and reuse the precious data shared with us by human subjects, current tools and practices do not allow for comprehensive or adequate de-identification. Further, individual research teams attempting to navigate de-identification would create a significant burden. Centralized guidance, tools, and instruction would improve best practices and allow for more comprehensive data storage and sharing. The NIH is encouraged to call upon scholars in the field of digital forensics and anonymity experts such as Dr. Latanya Sweeney.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

- The draft NIH Policy suggests that the timeline for submission of the data management plan align with the Just in Time period. This is likely to prove problematic for both researchers and their institutions in being prepared to comply with requirements and intentions.
 - o If researchers are not asked to write a data management until the Just In Time portion, it is unlikely that they will have appropriately addressed the financial and personnel costs required. This will then either require a significant budget adjustment or will lead to data management being under or un-funded; either of which will have significant impact on whether researchers are able to comply with requirements. We would encourage having a financial officer review the DMPs to estimate the cost for data reuse and the need for clinical models, etc.
 - o Additionally, by waiting until the Just In Time period, the data management plan –rather than being perceived as an active guide to assist throughout the stages of the data life cycle, will instead be only an administrative document. As such, researchers are likely to default to boilerplate templates which do not accurately reflect the ongoing data management obligations they will likely face and this will not serve the agency in promoting the improvement in data management practices.
 - o Further, this is not in alignment with other federal agency requirements and raises the potential for confusion or inconsistency. The National Science Foundation’s requirement for a data management plan comes with the preliminary grant submission, and this has been successfully implemented across other federal agencies over the past decade. Establishing consistency across the federal agencies for data management plan submission will reduce

challenges for researchers and broadly promote the consideration of data management requirements prior to requesting research funding.

- o The addition of the data management plan with the initial submission is not an unmanageable burden for researchers. By moving it to the point of initial application, it signifies to researchers the NIH commitment to improved data management and interest in data as a scholarly object.
- o Should the NIH maintain the Just in Time submission for the data management plan, it will be crucial to permit budget revisions to adequately address the costs of data management, sharing, preservation and access.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

- Clearer guidance on the use of repositories, the acceptability of disciplinary, generic, or institutional repositories, and expectations for use of them is needed. Without further clarification, institutions and researchers are at risk of the need to indefinitely retain large datasets for which they may not have adequate infrastructure.
- o Additionally, many researchers will need to work with multiple repositories for appropriate data storage for a single project, such as current work involving data deposited in both ProteinDB and GenBank. As the scientific communities continue to develop data type specific repositories (such as microarray data), NIH should encourage collaboration with non-profit organizations such as FairSharing.org
- o While we anticipate publishing partnerships for some aspects of repository storage and management, it is imperative that the NIH provide guidance which prohibits the introduction of financial barriers for access to tax-payer funded research data. Data restrictions within repositories should be focused on protecting participant data and vulnerable populations.
- o We encourage the NIH to prompt the various ICOs to identify more comprehensive recommendations and requirements for appropriate repositories.
- While investigators who are submitting grants to NIH will have a specific obligation to data management, we are concerned with the vagueness of the language used for oversight of data management. We appreciate that an individual is required to be named, however, in many instances this may be assigned to the primary investigator rather than accurately identifying the person or role who will be tasked with data management oversight. This also does not reflect the movement by many academic institutions to provide more centralized data management support. We encourage more distinct language in this section to clarify roles and responsibilities.

Other Considerations Relevant to this DRAFT Policy Proposal:

We are pleased to see the continued efforts of the NIH to promote Open Science and Open Data. As you continue to develop and implement policy, guidance, and education, we

recommend collaborating with professional organizations such as the Research Data Access Preservation Association; the Medical Library Association; Force11; and the Research Data Alliance. These organizations are already engaged in several aspects that this policy is addressing and can provide assistance and expertise.

Attachment:

Description:

Submission ID: 1421

Date: 1/10/2020

Name: Katie Steen

Name of Organization: Association of American Universities

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All extramural research at NIH

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Other

Role - Other: University Association

Domain of Research Most Important to You or Your Organization:

Association of American Universities

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Definition of Scientific Data

The proposed definition of scientific data is too broad. This is a departure from the definition of scientific data in NIH's 2018 Request for Information on Proposed Data Policy Provisions in that it now includes data "regardless of whether the data are used to support scholarly publications." This expanded definition will be difficult for universities and researchers to interpret and comply with because it requires extensive time and technical data expertise to assess the endless amounts of data that may be generated over the life of a grant.

The amount and type of data necessary to validate and replicate research findings is, in many cases, subjective and varies widely across disciplines. Without appropriate guard rails in place, this definition may result in the sharing of large swaths of data that are unnecessary, costly, and burdensome to manage and share. Additionally, maintaining quality and ensuring FAIR principles will be difficult, if not impossible, if researchers are required to share any and all data used in research findings throughout the project. To ensure the data researchers share is usable

by the research community and broader public, we suggest the definition of scientific data only include data underlying scholarly publications.

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

We support the statement that NIH may request additional information be included with the Data Management Plan. We interpret this to mean NIH will make it clear to the grantee if other information is expected, particularly for compliance with the Policy. This approach supports a collaborative relationship between the program officer and the grantee which is essential throughout the life of the grant. An expectation that NIH will indicate the need for more information ensures researchers will be selective and thoughtful about their Plans versus submitting an array of unnecessary information.

Section VI: Data Management and Sharing Plans:

It is helpful for NIH to allow for updates to be made to Plans at regular reporting intervals. This is critical as research projects often change throughout the grant. Researchers should not be expected to adhere to the initial Data Management Plan elements if changes occur during the research project that require new or different approaches to the scientific data produced. Furthermore, we are concerned the proposed Policy does not clearly state researchers may submit costs estimates after initial submission of the Plan. To account for requirements instituted by the program officer, we suggest NIH allow for additional direct costs to be submitted after the negotiated Data Management Plan is final.

More clarity is needed around the data NIH encourages grantees to make available. Asking researchers to make data public "as long as it is deemed useful" is not specific enough and will be confusing to researchers. The research community varies on what it deems "useful" data and researchers cannot be expected to know or understand NIH's view of what is useful. NIH should provide additional guidance as to what this language means or remove it completely.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The Guidance is confusing in that it allows "local data management considerations" as a direct cost, but then states "costs associated with collecting or otherwise gaining access to research data (e.g., data access fees)" are part of the costs of doing research and therefore not allowed. In addition to determining what data should be shared and the metadata needed to adhere to FAIR principles, appropriate local data management, curation, and access on university campuses is essential. Establishing appropriate data storage and sharing infrastructure is one of

the most difficult challenges facing research universities in their efforts to share data. Given the volume of platforms, repositories, referatories, persistent identifier (PID) generators, etc. used in the research data community, NIH should clarify what this language means.

We would suggest that the use of services and tools like DataCite, ORCID, CrossRef, figshare, and others be allowed as a direct cost in the grant proposal. Many of these tools require membership fees to participate or charge fees for additional services. These entities are critical to local data management on campus and may require significant campus investment through direct fees or human capital.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We appreciate NIH's proposal to limit the length of Data Management Plans. However, it may be difficult to address all of NIH's suggested elements in the proposed Plan. We welcome the opportunity to work with NIH, researchers, and other campus stakeholders to develop more guidance for Data Management Plans which may include allowing supplemental materials that augment the Plan.

We agree the rationale for decisions about which data should be made publicly available is important in both the research process and as a compliance mechanism. However, this requires a broad understanding of available data services and community standards. The specific rationale may be very difficult for an individual researcher or university to determine and describe accurately. To improve understanding within disciplines and across universities, NIH should consider providing its own guidance and rationale with specific reference to costs, security, and privacy. NIH is better positioned to indicate the appropriate balance between public access and associated costs, security, and privacy. To facilitate compliance and enhance research quality, we recommend NIH provide specific rationale on the balance between these priorities.

Other Considerations Relevant to this DRAFT Policy Proposal:

Introduction:

On behalf of the over 200 universities we represent, the Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU) greatly appreciate the National Institutes of Health's (NIH) efforts to seek public comment on the proposed Draft NIH Policy for Data Management and Sharing. Consultation and engagement with university stakeholders is critical to developing and implementing successful data sharing policies.

Our associations agree it is beneficial to make data from federally funded research available to the public to accelerate scientific discovery and ensure research integrity through robust replication and re-analysis. At the same time, it is imperative we achieve the appropriate

balance between public access and privacy to support and enable scientific inquiry. Appropriate data sharing and access requires significant consultation, collaboration, and investment by federal agencies, universities, scholars, and the research community more broadly. Policies should not only support access to data but enable reuse through adherence to FAIR data principles. AAU and APLU are actively working with our member campuses to develop appropriate campus policies, practices, and guidance to enable public access to research data.

To support public access to data on our campuses, we hosted an NSF-funded workshop in October 2018 on Accelerating Public Access to Research Data that brought together federal agency representatives and 30 institutional teams comprised of senior research officers, data librarians, general counsels, information technology specialists, faculty members, and other university administrators. The workshop identified challenges and opportunities for collaboration in data sharing through the development of campus action plans. In 2020 we will host a follow-up convening and two National Summits, funded by NSF with additional support from NIH, to continue this work and create a Guide to assist institutions in implementing appropriate data policies and practices.

Conclusion:

In summary, it is helpful NIH has acknowledged the variation across disciplines in data standards and the lack of standards in some cases. In addition, the examples of metadata standards and reference to NIH's Common Data Element Resource Portal in the guidance is informative for our researchers and university staff. However, more guidance is needed around allowable costs and "useful" data to ensure full compliance. We also hope NIH will consider changing the definition of "scientific data" to only include data underlying scholarly publications. Finally, we encourage NIH to harmonize Data Management Plan formats and submission processes across the institutes and centers to streamline compliance and accelerate scientific discovery.

We appreciate NIH's dedicated work with the research community to solicit feedback on potential data sharing and management policies. A collaborative approach with stakeholders is imperative to ensure public access to federally funded research outputs and compliance with associated agency policies.

Attachment:

AAU-APLU-NIH-DRAFT-Data-Plan.pdf

Description:

AAU & APLU Comments on NIH's Draft Data Management and Sharing Policy



MEMORANDUM

TO: Office of Science Policy, National Institutes of Health

FROM: Association of American Universities
 Contact: Katie Steen, katie.steen@aau.edu; (202) 789-5377

Association of Public and Land-grant Universities
 Contact: Kacy Redd, kredd@aplu.org; (202) 478-6022

DATE: January 10, 2020

Re: NOT-OD-20-013: “Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance”

On behalf of the over 200 universities we represent, the Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU) greatly appreciate the National Institutes of Health’s (NIH) efforts to seek public comment on the proposed Draft NIH Policy for Data Management and Sharing. Consultation and engagement with university stakeholders is critical to developing and implementing successful data sharing policies.

Our associations agree it is beneficial to make data from federally funded research available to the public to accelerate scientific discovery and ensure research integrity through robust replication and re-analysis. At the same time, it is imperative we achieve the appropriate balance between public access and privacy to support and enable scientific inquiry. Appropriate data sharing and access requires significant consultation, collaboration, and investment by federal agencies, universities, scholars, and the research community more broadly. Policies should not only support access to data but enable reuse through adherence to FAIR data principles. AAU and APLU are actively working with our member campuses to develop appropriate campus policies, practices, and guidance to enable public access to research data.

To support public access to data on our campuses, we hosted an NSF-funded workshop in October 2018 on [Accelerating Public Access to Research Data](#) that brought together federal agency representatives and 30 institutional teams comprised of senior research officers, data librarians, general counsels, information technology specialists, faculty members, and other university administrators. The workshop identified challenges and opportunities for collaboration in data sharing through the development of campus action plans. In 2020 we will host a follow-up convening and two [National Summits](#), funded by NSF with additional support from NIH, to continue this work and create a *Guide* to assist institutions in implementing appropriate data policies and practices.

General Policy Definitions and Requirements

Definition of Scientific Data

The proposed definition of scientific data is too broad. This is a departure from the definition of scientific data in NIH's 2018 Request for Information on Proposed Data Policy [Provisions](#) in that it now includes data "regardless of whether the data are used to support scholarly publications." This expanded definition will be difficult for universities and researchers to interpret and comply with because it requires extensive time and technical data expertise to assess the endless amounts of data that may be generated over the life of a grant.

The amount and type of data necessary to validate and replicate research findings is, in many cases, subjective and varies widely across disciplines. Without appropriate guard rails in place, this definition may result in the sharing of large swaths of data that are unnecessary, costly, and burdensome to manage and share. Additionally, maintaining quality and ensuring FAIR principles will be difficult, if not impossible, if researchers are required to share any and all data used in research findings throughout the project. **To ensure the data researchers share is useable by the research community and broader public, we suggest the definition of scientific data only include data underlying scholarly publications.**

Requirements for Data Management and Sharing Plans

We support the statement that NIH may request additional information be included with the Data Management Plan. We interpret this to mean NIH will make it clear to the grantee if other information is expected, particularly for compliance with the Policy. This approach supports a collaborative relationship between the program officer and the grantee which is essential throughout the life of the grant. An expectation that NIH will indicate the need for more information ensures researchers will be selective and thoughtful about their Plans versus submitting an array of unnecessary information.

Data Management and Sharing Plans

It is helpful for NIH to allow for updates to be made to Plans at regular reporting intervals. This is critical as research projects often change throughout the grant. Researchers should not be expected to adhere to the initial Data Management Plan elements if changes occur during the research project that require new or different approaches to the scientific data produced. Furthermore, we are concerned the proposed Policy does not clearly state researchers may submit costs estimates after initial submission of the Plan. **To account for requirements instituted by the program officer, we suggest NIH allow for additional direct costs to be submitted after the negotiated Data Management Plan is final.**

More clarity is needed around the data NIH encourages grantees to make available. Asking researchers to make data public "as long as it is deemed useful" is not specific enough and will be confusing to researchers. The research community varies on what it deems "useful" data and researchers cannot be expected to know or understand NIH's view of what is useful. **NIH should provide additional guidance as to what this language means or remove it completely.**

Allowable Costs for Data Management and Sharing

The Guidance is confusing in that it allows "local data management considerations" as a direct cost, but then states "costs associated with collecting or otherwise gaining access to research data (e.g., data access fees)" are part of the costs of doing research and therefore not allowed. In addition to

determining what data should be shared and the metadata needed to adhere to FAIR principles, appropriate local data management, curation, and access on university campuses is essential. Establishing appropriate data storage and sharing infrastructure is one of the most difficult challenges facing research universities in their efforts to share data. Given the volume of platforms, repositories, referatories, persistent identifier (PID) generators, etc. used in the research data community, NIH should clarify what this language means.

We would suggest that the use of services and tools like DataCite, ORCID, CrossRef, figshare, and others be allowed as a direct cost in the grant proposal. Many of these tools require membership fees to participate or charge fees for additional services. These entities are critical to local data management on campus and may require significant campus investment through direct fees or human capital.

Elements of a Data Management Plan

We appreciate NIH's proposal to limit the length of Data Management Plans. However, it may be difficult to address all of NIH's suggested elements in the proposed Plan. We welcome the opportunity to work with NIH, researchers, and other campus stakeholders to develop more guidance for Data Management Plans which may include allowing supplemental materials that augment the Plan.

We agree the rationale for decisions about which data should be made publicly available is important in both the research process and as a compliance mechanism. However, this requires a broad understanding of available data services and community standards. The specific rationale may be very difficult for an individual researcher or university to determine and describe accurately. To improve understanding within disciplines and across universities, NIH should consider providing its own guidance and rationale with specific reference to costs, security, and privacy. NIH is better positioned to indicate the appropriate balance between public access and associated costs, security, and privacy. **To facilitate compliance and enhance research quality, we recommend NIH provide specific rationale on the balance between these priorities.**

Conclusion

In summary, it is helpful NIH has acknowledged the variation across disciplines in data standards and the lack of standards in some cases. In addition, the examples of metadata standards and reference to NIH's Common Data Element Resource Portal in the guidance is informative for our researchers and university staff. However, more guidance is needed around allowable costs and "useful" data to ensure full compliance. We also hope NIH will consider changing the definition of "scientific data" to only include data underlying scholarly publications. Finally, we encourage NIH to harmonize Data Management Plan formats and submission processes across the institutes and centers to streamline compliance and accelerate scientific discovery.

We appreciate NIH's dedicated work with the research community to solicit feedback on potential data sharing and management policies. A collaborative approach with stakeholders is imperative to ensure public access to federally funded research outputs and compliance with associated agency policies.

Submission ID: 1422

Date: 1/10/2020

Name: Jacob Carlson

Name of Organization: University of Michigan Library

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: As a provider of research data services the library is interested in multiple data types.

Type of Organization: University

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

The NIH must make a solid case that the costs required to manage and share data will produce benefits for the public, the research community, the university and the researcher themselves. Without making this case it is difficult to envision the requirement to manage and share data being anything more than a perfunctory exercise performed with minimal investment. This policy document should reinforce arguments and statements made elsewhere by the NIH in support of data management and sharing, with references where appropriate.

Furthermore, the policy is written with the researcher as its sole focus. In reality, enabling good data management and sharing requires a broader set of stakeholders to provide needed expertise, resources and support. Acknowledging, addressing and supporting the larger network of stakeholders in making data management and sharing happen would go a long way in strengthening the policy. Much of the work that is done in supporting data management and sharing is invisible labor and often goes unrecognized despite its importance. Shining a light on the importance of this work and the people who do it would encourage broader discussions, interactions and collaborations around making data management and sharing more effective and efficient.

The NIH should be more clear and consistent in stating its expectations for sharing data. The purpose section of the policy states "improving the reproducibility and reliability of research findings" as an outcome of good management and sharing practices. However, it is not clear that the NIH expects researchers to share their data in ways that enable their work to be reproducible. Reproducibility is a high bar for researchers to attain, so if this is an expectation, it needs to be clearly stated as such throughout the policy. In addition, researchers may not have the tools, personnel or training to reasonably manage and share their data in ways that enable the reproducibility of their work. All of these factors will need to be accounted for by the NIH in setting expectations for managing and sharing data. Setting clear expectations as to what the intended outcomes of data sharing are, either in the data management and sharing policy or within individual ICOs or programs, will help researchers better understand and address expectations.

Section II: Definitions:

The wording used in these definitions is not as clear or direct as it could be in order to convey the explicit expectations that the NIH has of researchers. This is true of the definitions provided here as well as the policy document as a whole. The statement that "NIH expects that reasonable efforts will be made to digitize all scientific data" is an example of the broad and vague language that permeates the draft policy document. How will reasonable efforts be defined in practice? Individual researchers will likely determine what constitutes a reasonable effort based on their access to resources, tools, money, expertise and institutional support. Their determinations may or may not align with the expectations of the NIH, which may cause problems down the line, particularly for researchers at institutions with minimal resources. As sharing research data becomes a normative component of research and scholarship, the NIH and others will need to look beyond the availability of data and towards developing expectations around data quality and utility.

Section III: Scope:

Section IV: Effective Date(s):

We would encourage the NIH to not to wait too long to implement the new data management and sharing policy as there is a danger that people will put off preparing for its requirements if given leeway. However, the NIH should provide some time for researchers, NIH staff, and other stakeholders to adjust to what may be new and unfamiliar decisions and actions that they will need to take to comply with this policy. We would encourage the NIH consider a gradual implementation of the provisions contained in the policy and their enforcement rather than listing a single fixed date when the policy would go into effect.

We believe that a gradual implementation, if done well, would encourage researchers impacted by the new policy to make the necessary changes into their processes and practices with

deliberate consideration, rather than feeling rushed and overwhelmed by the policy. However, in order for this approach to be effective the NIH should make it clear what is expected from researchers and when. The NIH should also provide, sponsor or connect to training programs and have these programs available as a part of the implementation schedule. Researchers will naturally have questions not just on what the requirements of the policy are, but what decisions or actions they will need to take to comply with these requirements. Connecting researchers to training programs and other resources as a part of implementing effective dates will help researchers answer these questions. Finally, the NIH needs to consider and develop training programs for program officers and other key personnel involved in applying or overseeing the provisions of the data management and sharing policies. Training programs will help NIH personnel create more uniformity in how the policies are understood and enforced.

Section V: Requirements:

The NIH requiring that individual Institutes, Centers and Offices (ICO) review and approve data management plans as a part of awards is an important component of the data management and sharing policy. We applaud the NIH for including a statement indicating to researchers that the funding ICO may request that additional funds be allocated or that more details be provided by the researcher on how they will manage and share their data. This statement indicates the importance the NIH ascribes to managing and sharing data, and that this policy should be taken seriously by grant awardees.

However, just as researchers will need training in order to be able to respond effectively to the data management and sharing requirements of the NIH, so too will program officers and other NIH personnel need training in reviewing and overseeing data management plans. We recommend that each of the NIH ICOs develop clear expectations for their data management and sharing requirements and define expected outcomes for researchers. In other words, what would "effective management and timely sharing of scientific data" look like in practical terms? ICOs should then consider what knowledge, expertise and resources program officers will need to evaluate DMP statements to ascertain if expectations are met and the desired outcomes are likely to be realized. Expectations and outcomes should be communicated to researchers applying for funding early on in the process and then reinforced across the life of the award.

Allowing the costs of managing and sharing data as allowable expenses in a grant is also welcome, though it is unclear whether or not researchers would actually claim these expenses if they view doing so as taking money away from their research. We recommend the NIH consider allocating money explicitly and exclusively for data management and sharing purposes in their call for proposals.

We strongly recommend that the NIH expand its consideration of the support needed by researchers to successfully address data management and sharing requirements to include universities and other research institutions where the work will be done. DMPs are typically written to an audience of program officers based on the lofty and broadly defined provisions in data management and sharing policies. This can create a disconnect between the commitments in the DMP that have been reviewed and approved by the NIH and the capabilities and capacities of the units in the university tasked with supporting researchers in carrying out these commitments. As service providers, we in the U-M Library would welcome the opportunity to meet with program officers in the NIH to share our experiences and expertise in working with researchers directly to support their data management, sharing, curation and preservation needs. Regular communication between service providers, such as libraries, IT units and sponsored programs offices, and the NIH would promote better coordination, enable clearer guidance to researchers, and generate support for capacity building for data sharing by universities. This in turn would lead to better outcomes in satisfying the provisions of the NIH's data management and sharing policies. The Public Access to Research Data workshops (<https://www.aplu.org/projects-and-initiatives/research-science-and-technology/public-access/>) sponsored by the AAU-APLU could serve as a model for regular interaction between the NIH and support units within Universities on further defining the implementation of the NIH's data management and sharing policies.

Section VI: Data Management and Sharing Plans:

We appreciate the level of detail provided in the 2019 Policy for Data Management and Sharing in comparison to previous NIH policies, but feel that additional details are still needed for this policy to be effective. In particular, more information as to what constitutes an "established repository" is necessary to enable researchers to make informed decisions on how to share their data. Providing a definition of "established repositories" will also assist research communities, libraries and other agencies providing services to researchers in developing repositories and services to meet increasing demand. We recognize that different communities will have varying expectations for repositories in hosting and disseminating data generated through NIH funded research, but defining a common set of requirements will establish and communicate the baseline functionality needed for sharing, curating and preserving data effectively. These baselines would be valuable in informing researchers of their responsibilities for their data and in giving repository providers a basic framework for their systems. The NIH should consult with data sharing experts to appropriately select these requirements.

NIH ICOs should also work closely with the communities they serve and with supporting agencies such as libraries, IT units and IRBs to develop specific standards and guidelines for data repositories seeking to provide services to these communities. Although we are intrigued by the NIH Figshare pilot program, we believe that the scope and scale of the needs of the many communities served by the NIH will require a federated network of repositories rather than a

single centralized one. We strongly urge the NIH to consider and fund multiple approaches towards supporting the technical and human resources that will be needed to create sustainable infrastructures for managing, sharing, curating and preserving data. In particular, we encourage the NIH to provide support for developing local infrastructures at the institutional level that can be connected to larger systems. The direct support provided by researchers' institutions will be critical in carrying out the NIH's requirements and in normalizing data management and sharing practices.

We are glad to see the NIH prioritizing responsible management and sharing of data gathered from human subjects. In an era where data about our personal lives is routinely harvested, often without our knowledge or concern for our well being, the NIH and other funding agencies should take a strong stand to protect the privacy and dignity of individuals who agree to participate in research. Attention is particularly needed in safeguarding the data of populations that have traditionally been marginalized or who are vulnerable in the current political climate. Given that addressing data sensitivity and identifiability can be complex and require expertise beyond that of an individual researcher, the development and support of reliable technologies and appropriately educated data practitioners such as those mentioned above are particularly important. In addition, we recommend that the NIH go beyond developing means to protect researched populations and invest time, energy and resources towards considerations of how they could benefit from having a voice in access to their data and how it is used. One possibility would be for the NIH to work with marginalized communities to develop repositories that would host relevant data and tools for the community to use data about themselves in ways that would advance their interests.

One way to further support and promote DMPs that measure up to the standards NIH envisions is to promote the visibility of exemplars in this area; we are gratified to see that as part of this policy NIH may make DMPs public, and would encourage this especially where plans provide a useful model others could follow.

Section VII: Compliance and Enforcement:

We are pleased to see the NIH allowing for DMPs to be updated by the researcher or the sponsoring ICO over the course of the grant, and that the DMP will be reviewed as a part of the progress reporting done by the researcher. We recommend the NIH provide training for its program officers to know what expectations to set in order to promote the likelihood that resulting data will be usable by others and successfully transferred to a data repository or other platform for appropriate sharing and preservation. This training should be developed in conjunction with the library and data repository communities who have expertise and experience in working with researchers on these subjects.

We would also recommend that the NIH take this statement a step further and begin exploring the application of machine actionable Data Management Plans (maDMPs). One of the largest limitations of the current approach by funding agencies to make data publicly available is the static nature of the DMP. As a component of a grant application, a DMP is considered when the project is written up by the researcher, reviewed by the funding agency, and then largely forgotten. Under this draft policy, the DMP would be used in determining and enforcing compliance at the end of the project, which is a step forward. However, the DMP could be used as a means for stakeholders (the NIH, the institution, the repository slated to host the data, etc.) and the researcher to communicate progress, make adjustments based on findings and to connect the researcher to needed services and support at the point of need. A maDMP is embedded into the workflow of the project through the institution's IT systems for supporting grants. It captures the methods employed to manage, describe, store and prepare the data for sharing as they happen and communicates this information to the units and agencies who provide support. As a result, the maDMP more readily supports the data life cycle of the research rather than just the beginning and end stages. The recent PLoS article "Ten principles for machine-actionable data management plans" (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006750>) provides an excellent overview of maDMPs and how they could work .

The perceived lack of enforcement in data management and sharing policies has hindered the compliance of researchers with the data management and sharing policies of funding agencies. Including compliance and enforcement as a stand alone section in the policy sends the message that the NIH takes these requirements seriously. However, the language in this section is looser than it should be to achieve the desired effect. We encourage the NIH to update the policy's language to indicate that enforcement will occur 4-5 years after the policy has been implemented.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We appreciate the inclusion of detailed statements explaining which costs in managing and sharing data are allowable by the NIH. However, enabling researchers to include costs for managing, sharing, curating and preserving data in their grants is not likely to be sufficient without additional support and investments by the NIH. As an emerging expectation, data management and sharing may not yet be considered part of the costs of doing research, and researchers may be reluctant to include costs for managing and sharing data if they feel that they are taking away support from the more traditional aspects of doing research. The NIH may want to consider making additional money available in grants that could only be spent on activities or resources devoted to managing and sharing data.

In addition, the infrastructure available to support data sharing, curation and preservation is still developing in many fields and fragile in many others. The NIH and other funding agencies should explore ways that they can assist the research communities they support in developing sustainable infrastructure, both technical and human, to ensure that shared access to data collections of value becomes an established component of the research process. The NIH should also work with research communities to build financial models to support repositories in providing continued and ongoing access to research data. Charging data deposit fees is listed as an allowable expense in the new NIH policy, but this model may not be feasible for all repositories. Ongoing storage and curation costs are currently unaddressed and represent a concern for both researchers and repositories. An examination of ways the NIH could support ongoing storage and curation costs should be an important piece of any deeper exploration of sustainable infrastructure development.

We would also strongly recommend the NIH consider additional financial allowances under this policy. Managing, curating, sharing and preserving data require more than just access to technology and tools; they require people with the knowledge and skills to do the work effectively and efficiently. The personnel costs to do this work should be explicitly acknowledged by the NIH and allowable under the terms of this policy. Furthermore, the NIH should do more to support training in data management and sharing, and assist in promoting and supporting data management as a viable and visible career for talented individuals.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

These guidelines on the composition of a data management and sharing plan are a distinct advancement in comparison to previous instructions. However, there are still improvements that could be made. There are parts of the guidelines that are still not as clear or as detailed as they need to be. For example, the policy states that if certain elements of a DMP have not been determined by the researcher, they can list these parts as "to be determined" provided that a timeline or milestone for making this determination is listed. How will these elements be reviewed and approved by the NIH once determined? The NIH also states that researchers are not expected to share all of their data, but which data types, formats or components are exempt from sharing is not made clear. If these elements are to be determined by a specific IPO or program, this should be stated in the policy.

Although limiting the length of a DMP to two pages is understandable given that it is only one part of a grant application, covering all of the required elements in a DMP with the degree of detail needed for researchers to employ them actively over the course of the project is unlikely. We appreciate the emphasis on the enforcement of the DMP through annual reporting, as described in the new draft data management and sharing policy. However, we encourage the NIH to consider additional ways to make DMPs a more dynamic, integrated and useful part of

the grant award rather than limiting it to a static component of the grant application. More specifically, how can the NIH position the DMP as a tool that researchers can use to define the life cycle of their data and what needs to be done at each stage to manage, and eventually share and preserve, their data.

From our vantage point as providers of a data repository, one of the greatest barriers to successfully sharing data is the transfer of data from where it was actively developed and administered by a researcher to a third-party data repository for access and preservation. The draft NIH policy on DMPs alludes to the connection between researcher and repository by encouraging the researcher to consider if special considerations are needed for implementing the DMP such as getting permission from the repository. We would ask the NIH to consider going further in helping to forge a connection between researcher and repository, including creating a mechanism to alert the repository listed in the DMP when the award is granted of the anticipated data deposit. If the repository is made aware that they can expect a data set to be deposited to the repository at a certain time and receives a basic understanding of what the composition of the data set will be, the repository can reach out to and work with the researcher to help guide the preparation of the data, connect the research team with relevant and useful resources, and make the deposit process smoother and easier.

The description of a data set through its metadata and documentation is an essential element in enabling others to discover, understand, trust and use the data and enabling repositories to steward and preserve. This information is often underdeveloped or absent from data sets when deposited by researchers to repositories. Therefore more attention should be paid to raising awareness and providing training to researchers in developing metadata and documentation with the needs of end users and repositories in mind. Librarians and archivists have a deep understanding and a long history of connecting people to the information they need through the application of metadata and documentation. Researchers are also unfamiliar with data sharing agreements or licenses generally and so do not always consider the ramifications of choosing one license or approach over another. We recommend working with the library and archival communities to connect with and teach researchers what they will need to know to share their data successfully.

It is indicated that if data is put into a repository, data will be preserved long-term. What is the expectation from NIH of the length of time a data set will be preserved? Too often the assumption is made by researchers, administrators and funding agencies that preservation is forever. This assumption is neither desirable nor feasible. Preservation requires ongoing management of the content by the providers of preservation services, which in turn requires dedicated resources to support not only the expense of storing the content securely, but the

staff needed to carry out preservation work. The content being preserved may lose value over time to its designated community, particularly in light of the costs incurred by preservation. In addition, repositories cannot make open ended commitments to preserve content indefinitely. Some estimated duration of preservation, or at least a set timeline for reviewing the value of the content being preserved to determine if it is still needed, made in conjunction with the repository should be included in the DMP. However, we recognize that determining a duration for preserving a data set can be a challenging task for researchers, or even for experienced archivists. We would encourage the NIH to treat the estimated time for preservation provided in the DMP as a guideline to assist repositories in stewarding the data set rather than as something to be strictly enforced. Furthermore, we would ask that the NIH work with repositories, archivists and other relevant stakeholders to create reasonable criteria and guidance for researchers to consider in developing an estimated duration of preservation for their data. Finally, the NIH should include education on the basic preservation concepts and activities needed for researchers to understand and address data preservation requirements in any training programs developed to support the NIH data management and sharing policy.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1423

Date: 1/10/2020

Name: Glenn Dillon

Name of Organization: American Heart Association

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Other

Role - Other: Director of Research Operations

Domain of Research Most Important to You or Your Organization:

cardiovascular disease and stroke

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We understand this relates to data specifically, but nonetheless are concerned about potential action by OSTP to make all publications (and potentially data) immediately accessible, for a number of reasons that are being advanced via other mechanisms. We do feel there should be a presumption that all research data underlying a publication is shared in a timely manner, but mandating immediate access is problematic. The default should be that shared data must be made accessible (in a timely manner), except when justified by a small number of reasons, such as participant privacy concerns that cannot be overcome by protective measures, or studies on vulnerable populations.

Section II: Definitions:

The definition of Scientific Data includes all the factual data necessary to replicate the research, but in the Scope and Requirement sections (III and V) it only references sharing scientific data that "results in the generation of" or is "generated from" the research. This does not address the fact that in many cases a research question is assessed through further analysis of a previously generated data set (i.e., starting data) obtained from another investigator. It may or may not be the case that the starting data would have been generated in a manner such that it would have been subject to the proposed policy. To that end, we suggest the policy should encourage awardees, when applicable, to seek approval to share all data.

Section III: Scope:

Section IV: Effective Date(s):

The effective date of the policy and subsequent implementation should be more concrete. The final policy should have a "no later than" date for implementation, ideally 12 months after issuance of the final policy.

Section V: Requirements:

It is certainly important to protect human participant privacy, but it is also important that concerns about human participant privacy not be abused to eliminate appropriate data sharing. Many human participants expect that data from their participation will be shared with other qualified researchers. Ineffective sharing of the resulting data (assuming appropriate protective measures such as de-identification are in place) is unethical as it wastes human participants' contributions to research and may result in more patients being exposed to harm. Therefore; it should be an explicit goal of this policy and any submitted Data Management and Sharing Plans to maximize access to the data, subject to necessary restrictions.

Section VI: Data Management and Sharing Plans:

The draft states that NIH encourages scientific data to be made available. Instead, it should REQUIRE that scientific data are shared.

We suggest a data preservation and sharing policy that is modifiable for all types of data including but not limited to wearable device data, online application data, and social determinants of health data.

We suggest the NIH work with the electronic health record vendors on making this data more accessible to all researchers.

We suggest that NIH build in a timetable around GDPR.

We suggest a standard format and data dictionary for data and code sharing to improve the usefulness, especially as data is being shared in cloud-based platforms.

It is important that Data Management and Sharing Plans be provided to NIH peer reviewers and potentially NIH Institutes, Centers and Offices so they can consider the plan's effect on the application's overall impact, significance, and approach. Training reviewers to score review criteria should include review of the Data Management and Sharing Plan. Therefore, NIH

should require Data Management and Sharing Plans at the regular submission due date for an application, and not as a Just-in-Time submission.

The draft says that only data "deemed useful to the research community or the public" need be shared. An applicant should not solely decide what data is useful to others. Any exceptions to the general principle that scientific data must be shared must be justified and funding conditioned on prior approval by an NIH advisory committee of data management experts. These external groups should also be used to determine how long shared data should remain available.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

An entry of "to be determined" in a plan should not be allowed, as it is likely to result in subsequent development of unusable plans.

Section 1

The plan references providing information about "suitable" data repositories, but it doesn't define "suitable". Basic criteria should be established.

Describes "consistency with community practices" as a possible rationale for deciding which data are preserved and shared. In many scientific disciplines, community practices lag behind best practices for data management and sharing. This language allows certain communities to settle for mediocrity in data management and sharing, defeats the aim of this policy to improve data management and sharing.

Section 4 says that "if an existing data repository(ies) will not be used, consider indicating why not". This language is soft; the policy should require the use of established repositories, unless exceptions are justified and approved. It should not be up to applicants to decide not to use standard established repositories without clear justification.

Section 5 anticipates that applicants may have restrictions on sharing imposed by existing or future agreements. This provides a major loophole in the policy in that applicants may choose to enter into more restrictive agreements than necessary so that they can avoid data sharing. This can be overcome by (1) providing data sharing plans as part of initial peer-review so that peer reviewers can appropriately score any decrease in impact that may come about from restrictions on sharing, and (2) review by an NIH advisory committee that includes data scientists and librarians.

Other Considerations Relevant to this DRAFT Policy Proposal:

Use of both requirements and incentives to encourage high-quality data management would be desirable. We suggest consideration of an "Incentives for High-Quality Data Management and Sharing" section be added to the policy, including the following incentives, and others deemed appropriate:

1. Add to the NIH biosketch a section for key personnel to describe their most significant contributions to data management and resource sharing (including data, code, reagents, samples, and other materials). This should be separate from other contributions to avoid it getting short shrift due to lack of space. The past record of the principal investigator and other key personnel should be explicitly added to the scored review criteria.
2. We support NIH efforts to recognize and cultivate excellence in data management and resource sharing, both at the individual researcher and institutional levels, and suggest expanding those efforts.
3. Increased incentives for performing research on shared data from many different angles.

Attachment:

Description:

Submission ID: 1424

Date: 1/10/2020

Name: Pamela Webb and Lisa Johnston

Name of Organization: University of Minnesota

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All

Type of Organization: University

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

All

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Please see attached letter

Section II: Definitions:

Please see attached letter

Section III: Scope:

Please see attached letter

Section IV: Effective Date(s):

Please see attached letter

Section V: Requirements:

Please see attached letter

Section VI: Data Management and Sharing Plans:

Please see attached letter

Section VII: Compliance and Enforcement:

Please see attached letter

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Please see attached letter

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Please see attached letter

Other Considerations Relevant to this DRAFT Policy Proposal:

Please see attached letter

Attachment:

UMN RFI Response to NIH Draft Policy for Data Management and Sharing - Final.pdf

Description:

University of Minnesota RFI Response

Tovii Citie;, C1Imp11

Spons"red Projects 4d11i1istr11tio11

*450 McNamara A/11mni Center
200 Oak Stree1 S.E.
Minneapolis, MN 55455*

*O/Jiee: 612-624-5599
Fax: 6.12-624-4843*

January 10, 2020

Carrie D. Wolinetz, Ph.D.
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Rockville, MD 20892

Subject: Response Draft NIH Policy for Data Management and Sharing and Supplemental Draft Guidance (November 2019)

Dr. Wolinetz;

The University of Minnesota writes in response to the request for public comments listed above, published November 6, 2019 on the NIH OSP website, *As a public land grant institution, we strongly support federal agency policies ensuring public-access to scientific research data. Concurrently, we support a thoughtful approach to balancing the need for data access with the administrative burden and cost associated with data management planning, curating, and storing data.*

Any policy implemented by the NIH would have direct implications to the researchers we support. In fiscal year 2019, the Univer ity of Minnesota received 312.8 million dollars in funding from the NIH, accounting for 58.2% of our federal research funding.

Specifically, with input gathered from key research committees on campus as well as individual faculty, we would like to respond to the following proposed policy across several major themes:

- I. **Purpose (no comments)**
- U. **Definitions**
 - a. We appreciate that the definition of scientific data supports the inclusion of research findings even if the outcomes are negative or not selected for publication. This definition will help guide researchers in selecting appropriate research findings for sharing.
 - b. We recommend the following definition edit: Data Sharing: The act of making scientific data with adequate metadata available for use by others (e.g., researchers, institutions, the broader public),
 - c. If "NIH expects that reasonable efforts wrn be made to digitize all scientific data" then will digitization be an allowable cost to include in a grant? This was not mentioned in the allowable costs supplement. If yes, then we recommend mentioning it explicitly in

#1 (curating data) in the Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing.

- d. While we agree that new data generated over the course of a research process should be digitized whenever practical (or "reasonable", as stipulated by NIH), we do not feel that all historic data must be digital.

III. Scope

- a. We suggest that NIH clarify what funding mechanisms this applies to. For example, will training grants or career development grants require a data management and sharing plan?

IV. Effective Date(s)

- a. We appreciate that this policy is not being applied retroactively to ongoing or completed studies.
- b. If the intent is to make this policy applicable to Other Transactional Agreements (OTAs), this applicability should be specified as "Other Transactions" is not explicit as written here.
- c. Depending on the final disposition of the requirement, we recommend a minimum of 1-2 year implementation period.

V. Requirements

- a. We agree with NIH that Data Management and Sharing Plans should be submitted as Just-In-Time for extramural awards (as stated in Section VI), however, we suggest that NIH consider the impact this planning may have on the budget. We recommend that the budget for extra funds associated with these plans also be submitted at this time.

VI. Data Management and Sharing Plans

- a. We appreciate that NIH encourages researchers to submit to established data repositories, and encourage NIH to adopt community standards for vetting such repositories. For example, our institutional data repository, Data Repository for the University of Minnesota (DRUM), underwent a peer-review and vetting process to become a core trust seal certified repository in 2017. This could be a good standard to adopt.
- b. We appreciate the attention to tribal communities and tribal data and the ethical stewardship of those resources.
- c. Section VI could also be updated to include encouraging data to be shared with open licenses (such as Creative Commons Zero) and in repositories that do not charge access fees for using the data.
- d. While true, we find this statement difficult to interpret: "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." In order for institutions to decide what data are useful, we need data reuse metrics that allow for better tracking of data access and reuse. Data citation is a critical component to determining long term access and retention. NIH

should encourage proper data citation standards, with persistent IDs, for any data used in a project.

- e. Related to "Extramural Awards" we recommend that plans undergo a "programmatic assessment" in a way that is transparent to PIs. NIH should publicly post evaluation criteria and the process, including roles, responsibilities and timelines. We also recommend that the timeline for review be within 3 months of award acceptance.
- f. Detailing guidance on protecting the data during the research process would be of help for the local data storage and security providers.

VII. Compliance and Enforcement

- a. It is not reasonable to associate non-compliance by a given researcher with the NIH ICO-approved plan to be a factor in considering future funding for the entire recipient organization since adherence to a plan falls primarily to the individual investigator. Rather than non-compliance should be treated the same way as adherence to other Public Access requirements which impacts only the impacted named Principal Investigator(s). This is of particular importance for large institutions since there would be a large negative impact on scientific progress to halt all awards to an institution with hundreds or even thousands of (compliant) investigators. It would nonetheless be helpful for recipient organizations to be informed about instances or rates of non-compliance so that compliance issues can be tracked, analyzed and addressed.
- b. We appreciate that the Data Management and Sharing Plan review and update process was integrated into RPPRs. We agree that these Plans should be a living document and adherence to the Plan commensurate with the research as it evolves over the life of the project; good faith attempts to meet the principles of these requirements (even if occurring via alternative methodologies) should be permitted when appropriate; it may be advisable to have a data ombudsperson available at NIH or an appeals board to address disagreements about what is and what is not acceptable adherence to a Plan (particularly if consequences are contemplated).
- c. For ease of checking compliance with this policy, NIH should encourage a persistent identifier (PID) to the shared data (and metadata) be included in interim and final reporting.

VIII. Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

- a. We recommend that digitization of data be listed as potential allowable costs in category #1.
- b. The draft guidance about allowable costs indicates "*Budget estimates should not include infrastructure costs typically included in institutional overhead (e.g., Facilities and Administrative costs), nor costs associated with the routine conduct of research. Costs associated with collecting or otherwise gaining access to research data [e.g., data access fees] are considered costs of doing research and should not be included in budgets.*" While we concur that the costs of storing and managing data are not today typically charged as direct costs, if new federal requirements expand the obligations of grantees in this regard, care must be taken to ensure the new costs can be funded - either as direct costs

or as F&A costs. NIH should work with grantees to determine how additional requirements can best be funded; it should be noted that simply requiring grantees to cover these costs as a part of their F&A costs is not viable given that these funds are already oversubscribed for existing administrative costs. This is further discussed below in (c) and (d) below.

- c. If there is no stated retention period for data and a repository has a recurring fee for hosting the data, how should this be budgeted? Should NIH stipulate a maximum number of years or should the researcher be allowed to request supplemental funds for storage/hosting costs once the data has been created and can be appropriately assessed for long-term value? If costs for data preservation and sharing cannot be incurred after the grant end date or after the grant closeout date (e.g., more than 120 days after the project period end), how should that be budgeted /requested and accounted for? Stated more globally, we recommend that NIH develop a way to support legally proposing and charging long term storage, access, and preservation costs, as well as a mechanism to analyze and determine when the data no longer merit storage.
- d. If research storage is a common good that is considered part of the institutional overhead, and therefore costs cannot be recovered directly through the grant as a direct charge, an alternative could be cloud storage and access (not "local") could be deemed to be an F&A type cost. Most research institutions, including our own, already incur substantially more administrative costs than are able to be recovered under the 26% administrative cap in the F&A rate. It is therefore unrealistic to expect more costs to be added to the F&A portion of a grant without federal willingness for that cap (set in 1991) being upwardly adjusted. Nevertheless, data management and storage is critical and perhaps could be a catalyst for this cap to be adjusted.

IX. Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan

- a. We are unclear on how much of the plan can be "to be determined" in the Just-In-Time submission? Who reviews and assesses the updates to the plan? Do researchers need to provide rationale or justification for changes to the data management plans after IRB approval? Please provide a process.
- b. "NIH does not expect researchers to share all scientific data generated in a study" - these data should still be managed, but the existing guidance for the Data Management Plan in this section is focused on the data that will be both preserved and shared. Researchers should indicate to what extent they expect to manage, secure, store and protect the data they collect but which are not intended for sharing.
- c. Under data type, we recommend that NIH acknowledge the importance of appropriate consent that informs participants about data sharing. We encourage NIH highlight the existing IRB processes responsible for ensuring that participants are informed, explicitly, when their de-identified data will be shared in a public repository. Researchers should work with IRB and their data sharing venue to make sure consent/agreements align with plans to share.
- d. We appreciate the thoroughness of the data type section, specifically including the modality, aggregation, and level of processing done to the data.

- e. We would like to see language in the Related Tools, Software, and Code section that encourages researchers to share alternative versions of the data when access, cost, or proprietary nature of the tools may overly restrict reuse. For example, proprietary files created using Microsoft Excel can be exported as nonproprietary files such as .csv for sharing. This statement could be more enforceable if the word "consider" is removed: "If scientific data will be archived in an existing data repository(ies), consider providing the name and URL web address of the repository(ies). If an existing data repository(ies) will not be used, consider indicating why not and how scientific data will be preserved and shared."
- f. We recommend the following wording change to decisions on when data can be discarded will vary greatly. "If applicable, consider indicating the process that will be used to ascertain when scientific data will no longer be available to other users."
- g. In Section 5, NIH should encourage researchers to share data with open licenses that explicitly detail the conditions for reuse (such as Creative Commons) when possible.


X. Other Considerations Relevant to this DRAFT Policy Proposal (no comments)

Finally, we thank you for giving us this opportunity to provide input on this critical topic to the NIH. We look forward to continuing the dialogue on this topic.

Sincerely,



Pamela A. Webb
Associate Vice President for Research



Lisa Johnston
Director, Data Repository for the U of
Minnesota (DRUM), University Libraries

cc: Chris Cramer
Wendy Lougee
Jakub Tolar

Submission ID: 1425

Date: 1/10/2020

Name: Heidi Rehm

Name of Organization: Massachusetts General Hospital

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Health Care Delivery Organization

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

rare disease

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I agree with most of the comments of Michael Hoffman, restated here "There should be a presumption that all research data underlying a publication is shared at time of publication. The current language is weak and has statements such as "shared data should be made accessible" or "not all data generated in the course of research may be necessary to validate and replicate research findings." Instead the policy should say that shared data MUST be made accessible, except when justified by a small number of reasons, such as participant privacy concerns that cannot be overcome by protective measures, or studies on vulnerable populations.

The draft lists an expectation of "timely" data sharing. This should be defined as generally at the time of publication. Funding opportunities specifically designated to create a shared resource should specify a date by which data must be available even in the absence of a publication. This aspect is a step backwards from previous NIH policy which clearly defines "timely" as "no later than the acceptance for publication of the main findings from the final data set." The relaxation of this existing requirement is not justified." However, I would add that for many research programs, even sharing data only at the time of publication would not be acceptable and instead, grant funding mechanisms should request that the Data Management and Sharing Plan (Plan) should state explicitly how soon after the production of data, the data will be shared (e.g. 6 months or 1 year after production), regardless of the timeline of publications using the data.

Section II: Definitions:

I agree with the comments of Michael Hoffman, restated here: "Should include definitions of FAIR data and the 15 FAIR principles."

Section III: Scope:

I agree with the comments of Michael Hoffman, restated here: "Scope should make clear that the policy continues to apply for scientific data produced by funding in whole or in part from NIH after the NIH funding period is over."

Section IV: Effective Date(s):

I agree with the comments of Michael Hoffman, restated here: "The current absence of an effective data management and sharing policy and lack of enforcement causes a serious negative impact on health research and enables an ongoing waste of public funds. The noncommittal implementation date of the draft is unacceptable. The final policy should have a "no later than" date for implementation, ideally 12 months after issuance of the final policy."

Section V: Requirements:

I agree with the comments of Michael Hoffman, restated here: "To ensure good data management, any data described as collected in a progress report must be deposited independently and an accession code or digital object identifier (DOI) supplied. Except when specified by the funding opportunity announcement, researchers may embargo this data until publication. Grant opportunities specifically designated to create a shared resource should specify a date by which data must be available even in the absence of a publication."

It should be clear that these requirements apply not just to research project grants and contracts, but most other forms of requests for support that will lead to the creation of scientific data. This includes cooperative agreements, career grants, fellowships, scholarships, and training grants.

Absent a compelling reason otherwise, contract solicitations should specify that collected data is the property of NIH. They should also include specific requirements that data should be made publicly available in a third-party repository as a periodic deliverable, upon which further funding can be conditioned.

There are a large number of digital repositories with different policies. You should require that acceptable digital repositories must not allow recipients to unilaterally change or delete deposited data. The repositories, may, however, allow adding new versions of data advertised in metadata for the original dataset.

It is important to protect human participant privacy but it is also important that concerns about human participant privacy not be abused to eliminate appropriate data sharing. It is especially worth considering that many human participants expect that data from their participation will be shared with other qualified researchers. Ineffective sharing of the resulting data (assuming appropriate protective measures such as de-identification are in place) is unethical as it wastes

human participants' contributions to research and may result in more patients being exposed to harm. Therefore it should be an explicit goal of this policy and any submitted Data Management and Sharing Plans to maximize access subject to necessary restrictions."

Section VI: Data Management and Sharing Plans:

I agree with the comments of Michael Hoffman, restated here:"The draft states that NIH encourages scientific data to be made available. Instead, it should REQUIRE that scientific data are shared.

An effective Data Management and Sharing Plan should increase the overall impact of a grant and an ineffective one will decrease it. It is important that Data Management and Sharing Plans be provided to NIH peer reviewers and ICO advisory council review so they can consider the plan's effect on the application's overall impact, significance, and approach. Guidance to reviewers on how to score review criteria such as significance and approach should include review of the Data Management and Sharing Plan.

Therefore, NIH should require Data Management and Sharing Plans at the regular submission due date for an application, and not as a Just-in-Time submission. Overcoming deficiencies in the Data Management and Sharing Plan identified in summary statements could be provided as a Just-in-Time submission.

NIH should require that data management plans must describe how the researchers address each of the 15 FAIR Principles.

NIH should publish data management plans for funded grants and contracts alongside abstracts in public databases such as RePORTER. This will increase transparency and let other researchers and the public know what the grantees promised to NIH. This is the only thing that will make enforcement of individual plan items possible, given that NIH does not have the resources for exhaustive, systematic checks on compliance. Grantees knowing that their data management and sharing promises are readily available to the public will provide some measure of self-enforcement. Currently data sharing plans are available through Freedom of Information Act requests, and putting them on RePORTER will reduce the burden on data requesters.

The draft says that only data"deemed useful to the research community or the public" need be shared. It should be clear that applicants do not get to unilaterally decide what data is deemed useful. Any exceptions to the general principle that scientific data must be shared must be justified and funding conditioned on prior approval by an NIH advisory committee of data management experts that includes data scientists and librarians.

For intramural research, you should not give a single NIH official (such as Scientific Director or Clinical Director) the ability to assess Data Management and Sharing Plans without oversight. Data Management and Sharing Plans must be reviewed and approved by Boards of Scientific Counselors and ICO advisory councils during the existing periodic peer review and site visit process."

Section VII: Compliance and Enforcement:

I agree with the comments of Michael Hoffman, restated here: "It is currently unclear where to turn when NIH data sharing expectations and policies are not followed. To solve this, RePORTER should list, for each grant, contact information to request corrective action for violations of the Data Management and Sharing policy or published Data Management and Sharing plans. This should include contact email addresses for the principal investigators/project directors of the grant, contact email addresses for officials representing the grantee institution, and a contact email address at NIH. That will allow for solving issues at the most local level, when possible, and escalation when the previous proves ineffective. Similar information should be available for contracts and for intramural research projects.

In addition to reviewing progress reports and addressing complaints, NIH ICOs should also perform more thorough random audits to ensure grantees are performing data management as expected.

Current sanctions listed in the draft policy are incredibly weak and will have no deterrent effect. The policy should mention that failure to follow the Data Management and Sharing policy can be considered research misconduct by NIH. The policy should specify that violating the policy in place at the time of competing award at any time thereafter (including after the end of the award period) can result in sanctions. These sanctions can include publication of a notice describing the violation in the NIH Guide to Grants and Contracts, debarment and suspension from contracting, subcontracting, or financial assistance from the federal government, and prohibition of service to the Public Health Service on advisory committees, boards, or peer review committees, or as a consultant. Because it touches on potential research misconduct, this policy must be reviewed by the HHS Office of Research Integrity."

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

I agree with the comments of Michael Hoffman, restated here: "The guidance should specify that fees that preserve data beyond the funding period are allowed, as are personnel expenses related to data sharing."

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

I agree with the comments of Michael Hoffman, restated here: "An entry of "to be determined" in a Plan is not acceptable. This language will encourage useless Plans and should be removed.

Statements like "NIH does not expect researchers to share all scientific data generated in a study" defeat the purpose of this policy. Instead NIH should make clear that they do expect and require sharing of scientific data except in limited exceptions, justified by the applicant, and prior approval by peer reviewers, program staff, and an NIH advisory committee of data management experts that includes data scientists and librarians.

Section 1 describes "consistency with community practices" as a potential rationale for deciding which data are preserved and shared. In many scientific disciplines, community practices lag far behind general best practices and what the public expects for data management and sharing. This language allows certain communities to settle for mediocrity in data management and sharing, defeats the aim of this policy to improve data management and sharing. It should be removed. This also illustrates why decisions to withhold scientific data from sharing should not only be reviewed by study section members trained in the same discipline but also an NIH advisory committee of data management experts that includes data scientists and librarians.

Section 4 says that "if an existing data repository(ies) will not be used, consider indicating why not". This policy should require the use of established repositories, except when exceptions are justified and approved. It should not be up to applicants to unilaterally decide not to use standard established repositories and to not even justify the same.

Section 5 anticipates that applicants may have restrictions on sharing imposed by existing or future agreements. This provides a major loophole in the policy in that applicants may choose to enter into more restrictive agreements than necessary so that they can avoid data sharing. This can be overcome by (1) providing data sharing plans as part of initial peer-review so that peer reviewers can appropriately score any decrease in impact that may come about from restrictions on sharing, and (2) review by an NIH advisory committee that includes data scientists and librarians."

Other Considerations Relevant to this DRAFT Policy Proposal:

I agree with the comments of Michael Hoffman, restated here: "I applaud your efforts to establish an excellent research data management and sharing policy. As written, I do not think this policy will provide a substantive change in data sharing. To maximize the benefit to the public of providing research funds, it is essential that the policy and enforcement be strengthened as described in this response.

In general, the draft policy is overly cautious and fails to consider the burden an ineffective policy will place on researchers who seek to use shared NIH-funded scientific data. The current system is incredibly burdensome on those seeking to obtain shared data because when data are not available as per existing NIH expectations, investigators can stonewall requests. There is no enforcement and the way to request enforcement is unclear. My most serious concern about this policy is that it is too vague on requirements in some places and lacks sufficient detail on enforcement.

A policy with ineffective, vague requirements and no real enforcement will have a serious negative impact on researchers who seek to use scientific data produced with public funds. There is a huge waste of researcher time and money attempting to obtain data that is lost, improperly described, or withheld. Failure to follow good data management practices leads to great inefficiency and slows the work of many researchers. There is also a large impact on our

research communities, which lose opportunities to aggregate data and create a whole that is greater than the sum of its parts.

It is good to have both requirements and incentives to encourage high-quality data management. I suggest that an "Incentives for High-Quality Data Management and Sharing" section be added to the policy, including the following incentives:

1. Add to the NIH biosketch a section for key personnel to describe their most significant contributions to data management and resource sharing (including data, code, reagents, samples, and other materials). This should be separate from other contributions to avoid it getting short shrift due to lack of space. The past record of the principal investigator and other key personnel should be explicitly added to the scored review criteria.
2. NIH should create awards to recognize and cultivate excellence in data management and resource sharing, both at the individual researcher and institutional levels."

Attachment:

Description:

Submission ID: 1426

Date: 1/10/2020

Name: Jessica Chong

Name of Organization: University of Washington

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

genetic disorders (rare and common)

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section II: Definitions:

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section III: Scope:

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section IV: Effective Date(s):

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section V: Requirements:

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section VI: Data Management and Sharing Plans:

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Section VII: Compliance and Enforcement:

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Other Considerations Relevant to this DRAFT Policy Proposal:

I agree with the comments of Michael Hoffman here: <https://hoffman.bitbucket.io/2019/nih-data-management.html>

Attachment:

Description:

Submission ID: 1427

Date: 1/10/2020

Name: Elizabeth A. McGlynn

Name of Organization: Kaiser Permanente

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other:

Type of Organization: Other

Type of Organization - Other: Integrated Delivery System

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

Various

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

See attached comment letter

Section II: Definitions:

See attached comment letter

Section III: Scope:

See attached comment letter

Section IV: Effective Date(s):

See attached comment letter

Section V: Requirements:

See attached comment letter

Section VI: Data Management and Sharing Plans:

See attached comment letter

Section VII: Compliance and Enforcement:

See attached comment letter

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Comment Letter

Submission ID: 1428

Date: 1/10/2020

Name: Megan Potterbusch

Name of Organization: George Washington University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Support for all research data

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other: Librarians and a Director of Research Integrity

Domain of Research Most Important to You or Your Organization:

Genomics, epidemiology, clinical research

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

There is a disconnect between the new Policy for Data Management and Sharing and the Elements of a NIH DMP. There are no elements in the guidance that cover how data will be validated, secured, maintained, and processed; however, according to the definition of "Data Management" these activities are relevant to the policy. In particular, the language used indicates that they are only talking about "preserving and sharing" in the long-term/for use outside the initial research team's proposed project. I.e. the questions primarily concern the logistics of the data as it regards to sharing (long-term preservation). Is this your intent?

Section II: Definitions:

In our local conversations between departments, we noticed a need for NIH to define the term "preserve," because preserve can mean at least two different things to different audiences:

- 1-Preserve (maintain, manage, secure) integrity of active data while in use by PIs.
- 2-Preserve with a focus on long-term stewardship and future use by others

Section III: Scope:

The data life-cycle starts at the beginning and goes to the very end and includes the stewardship of data while they are in process. The policy from the NIH appears to be really

focused only on "What are you doing to the data to get it ready to be shared?" and not as much on the living management of data during the research project.

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

If a PI entered "to be determined" in the initial DMP, who is responsible for following up on this detail?

Section VII: Compliance and Enforcement:

What level of information are you looking for from the DMP? How much evolving do you expect for the plan over time? If the plan changes, when should they be notified of the changes? Only at specific intervals or whenever a change is needed? What scale of change would warrant contacting the GPO?

-- How will the NIH ensure compliance? Additional information about this will really help for having an understanding of how we should expect compliance to be enforced. We would appreciate documents that show how serious they should take the data management and sharing plan

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Point 2 (Preserving and sharing data through established repositories). The section indicates that budgets may include costs that would be incurred for preserving and sharing data. Do these costs include those that may be incurred after the period of performance has ended, like continuing annual fees charged by repositories for data preservation?

-- General comment for this supplemental document: Would recommend NIH seek additional guidance from Tribal Nations regarding the possibility to include as allowable costs those related to the engagement of tribal communities to ensure data are managed and shared in culturally and socially sensitive ways. This may include (but not be limited to) the costs incurred for travel, community meetings, etc. Recommendation that costs include those above and beyond the costs of doing research and gaining access to data.

-- Point 3: Is there a limit to allowable costs in the data management for allowable costs? What is considered "reasonable"? These costs can be quite expensive. Is this only about getting the data ready to be shared? (Documentation, etc.?) Or does it include other things?

-- Will the policy be dictating clearly what the costs can be used for? For example, would this cover specially designed data storage for organizing, and active sharing within a team? What if the research team has a need for a huge amount of storage that cannot be included in

overhead costs for the university, would this be considered a "facility or administrative cost" or would it be specialized information infrastructure?

-- Would an institution with more unique, sophisticated, and specialized infrastructure be judged differently from an institution still developing these facilities/capacities? Will data management plans impact funding score? And how will this be reviewed consistently by the grant program officers?

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Can supplements and addendum be added to the Data Management and Sharing Plan? (e.g. a job description for a data manager)

-- Point 2, page 2 (Related Tools, Software, and/or code). Would suggest NIH consider adding language to encourage researchers to provide relevant links to their code and/or scripts preserved in repositories to help those attempting to engage data sources and replicate published works. In particular, the research group may be developing their own software or analysis code and thus not know the specifics of all the pieces they may bring together nor the name of the software they develop.

-- This guidance indicates that the only data to be described in the data management plan are data that will be curated, preserved, and shared at the end of the project for future use and re-use. Is this correct?

Other Considerations Relevant to this DRAFT Policy Proposal:

NIH's federal registry publication (Background section, 2nd column) indicates that the organization intends to continue conversations with Tribal Nations to develop culturally sensitive data management and sharing resources for researchers seeking to collaborate with Tribal Nations. NIH encourages comments on specific strategies for promoting responsible data management and sharing in these types of research settings.

-- In light of this context, one recommendation for NIH to consider would be to extend this set of conversations to also include members of historically-marginalized communities, especially communities of color given the historical trauma experienced by members of these communities in health-related research.

-- NIH may also want to seek additional guidance from researchers who engage participatory research methods in examining topics of community health. My concern is that the policy of data sharing (as it is currently framed) presumes that data created in the research process are "owned" by a research team who can make these data available for sharing. While this is in one sense true, researchers engaging participatory research methods in collaboration with members of low-income or historically marginalized communities to understand topics like food insecurity or community health may have access to the data but the data itself may actually be "owned" by the research participants. In this case, the policy is unclear if it is sufficient that

data created in these types of engagements be shared with research participants or participating communities? Or must researchers negotiate with the community owners of these data to make these data available through additional channels to be in compliance with the policy?

Attachment:

Description:

Submission ID: 1429

Date: 1/11/2020

Name: Greg Janée

Name of Organization: University of California at Santa Barbara

Type of Data of Primary Interest: Imaging

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role:

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Data management plans are currently bundled with competitive proposals, and as a general rule are not made available by granting agencies. NIH is no exception in this regard. This lack of transparency creates an issue that confronts both researchers creating data management plans, and curators such as myself who assist researchers in creating plans. Since so few data management plans are public, there are few models to go by to understand what constitutes a "good" or "acceptable" plan, or whether a plan follows what most others are doing within a given community.

NIH's change to allow DMPs to be submitted just-in-time is a welcome change toward openness and transparency. While I would advocate that plans still need to be evaluated, I see benefits in separating the evaluation of a proposal on intellectual and competitive merits, from its

evaluation on compliance to data publication and preservation requirements. I encourage NIH to go a step farther and to make the data management plans of all funded proposals public.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1430

Date: 1/11/2020

Name: Ellen O'Meara

Name of Organization: Kaiser Permanente Washington Health Research Institute

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

cancer

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Re. item 4, last bullet, "In general, scientific data should be made available as soon as practicable, independent of award period and publication schedule." We appreciate the need for timely sharing of publicly funded data. But because we obtain the funding to do novel research, we expect to publish our findings before sharing the relevant data elements beyond our research team. We ask to make data available **after** papers are accepted for publication.

Other Considerations Relevant to this DRAFT Policy Proposal:

Submission ID: 1431

Date: 1/11/2020

Name: Agnes

Name of Organization: University of California

Type of Data of Primary Interest:

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

University of California Office of the President

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

UC appreciates the NIH's philosophy that underlies the draft policy as discussed in the Purpose section. It serves as a helpful reminder that investigators are not conducting their work within a vacuum. Proper research data management will inevitably include not just researchers, but also various institutional research administrative units and departments, along with research funders, regulators and journals equally invested in promoting the dissemination of knowledge within the constraints of legal, ethical, technical, security, and privacy considerations. We suggest that the NIH include in this section a recognition of data preservation and sharing as part of the entire research lifecycle.

Section II: Definitions:

The definition of "scientific data" requires clarification. According to the draft policy, scientific data is "the recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings regardless of whether the data are used to support scholarly publications." The definition is very broad, but at the same time, excludes specific sources of information. This ambiguity could lead to confusion about how best to interpret "scientific data" and implement proper data management and sharing plans. For example, the definition of scientific data incorporated into the draft policy explicitly excludes laboratory notebooks. Information contained in laboratory notebooks, however, might reasonably be seen as "necessary to validate and replicate research findings." Given its breadth,

the proposed definition of scientific data seems open to the interpretation that almost all data collected for a study counts, and therefore should be shared. UC asks that NIH provide clarification on this in addition to examples of accepted data standards.

Furthermore, the NIH's expectation that reasonable efforts will be made to digitize all scientific data be digitized raises concerns to UC. Not all scientific records must be digital to be useful. Digitizing also imposes administrative burden upon institutions, and creates risk, both in loss or error in translation, as well as in of the disclosure or use of sensitive material, including medical information. If external parties are necessary to digitize such data, the risk of loss, error or exposure becomes more pronounced. Because of the potential for heightened risk, the decision to digitize data should be left to the principal investigator's discretion. UC recommends that this expectation be removed.

Section III: Scope:

Section IV: Effective Date(s):

UC asks that the NIH choose a policy implementation date that will allow the research community to prepare sufficiently for the policy change and to consider a reasonable embargo period to advance intellectual property in its Supplemental Draft Guidance for data sharing plan elements (Section 4, Data preservation, Access and Associated Timelines). We recommend a minimum implementation date of one year after the release of the final policy, with a delay in enforcement actions for at least one year after the implementation deadline. Any determination of non-compliance should follow well-defined and transparent criteria.

Section V: Requirements:

While UC supports the requirement of a data management and sharing plan for NIH-funded or conducted research, we are concerned about varied supplementary information requirements requested by individual NIH Institutes, Centers, and Offices (ICOs). To minimize confusion and administrative burdens, we strongly urge trans-NIH coordination of these supplemental requests and listing ICO-specific requirements as part of centralized resources associated with the final data management and sharing policy.

Section VI: Data Management and Sharing Plans:

UC appreciates the proposal to collect data management and sharing plans as part of Just-in-Time (JIT) documentation for extramural awards. Requiring submission of the plan at the JIT phase rather than at the proposal stage minimizes administrative burden for both the applicant and peer reviewers. Shifting the review of plans to NIH staff members rather than volunteer reviewers will also make the process more uniform and streamlined, provided NIH staff are properly trained in reviewing data management plans. One potential drawback of submitting data management plans during JIT, however, is that accurately budgeting for data management

costs at the time of proposal will be challenging as details of the plan remain to be finalized with the NIH Program staff at JIT. UC, therefore, recommends that the NIH allow additional data management costs to be added to the budget at JIT based on the final negotiated data management plan. We also recommend an option that allows grantees to appeal ICO-mandated data sharing requirements to the NIH Policy Office should the requirements be considered unreasonable or inappropriate by the grantee, without fear of reprisal.

Furthermore, UC strongly urges the NIH to provide data management and sharing templates. Rather than having each researcher develop their own plans in two pages or less, having a common template will streamline the process and reduce burden for both submitters and reviewers. It would also be useful if NIH published sample plans. Within these templates, we recommend that NIH provide minimum requirements for researchers to include in a data management and sharing plan, such as data type, standards and metadata, plans for data preservation, and projected data accessibility. In addition, the template could include the usage of persistent identifiers and tools to generate machine-readable and actionable data management plans that support compliance checking and will facilitate the interoperability of data between systems. At a minimum, it would be useful if the NIH policy required the use of ORCID ID's for all individuals mentioned in the data management and sharing plans and for the use of persistent identifiers for all research datasets.

Finally, UC recommends that the NIH provide guidance on appropriate ways to maintain sensitive data. Not only is this administratively burdensome, but it also introduces a dependence on the researcher (and their current contact information) that undermines the goal of long-term data accessibility. NIH should recommend restricted access repositories as well as other kinds of repositories that so that the researchers understand what qualifies as an "established repository." In this context, the NIH has the unique opportunity to lead the community by creating field-specific data repositories that have the added benefit of ensuring relevant security and privacy concerns are addressed.

Section VII: Compliance and Enforcement:

UC appreciates that the NIH will allow researchers to make updates to their data management and sharing plan. In practice, data management plans may evolve over time as repositories, technologies, and data standards change. We are pleased to see some degree of flexibility built into the plans to address unforeseen issues.

The policy is not clear how adherence to a data management and sharing plan is verified at the end of a project. Requiring a final report on a data management and sharing plan would be an excellent way to enforce compliance and ensure that a researcher has followed through on

plans for sharing and depositing data. Adding details how compliance is monitored would give the policy some enforcement capabilities and make sure it is taken seriously.

Additionally, we wondered who would be responsible for reviewing and approving data management and sharing plans during the initial review process. It may be useful to specify the required knowledge and expertise that will be required of reviewers to ensure that they have the necessary backgrounds to effectively analyze and comment upon data management and sharing plans.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

UC appreciates the NIH's recognition of the costs associated with data management and sharing and applauds the inclusion of the supplemental guidance defining possible allowable costs. However, UC is concerned that the guidance only addresses those costs incurred during the term of the award but does not address costs associated with long-term data retention and accessibility, such infrastructure costs related to storing data . We strongly urge the NIH to allow researchers to budget for long-term data curation and preservation costs, or at a minimum clarify that grantee institutions may pre-pay from their awards these long-term costs. We would recommend that if these long-term costs are not permitted on a grant-by-grant basis, that the agency offers additional supplemental funding to institutions to develop infrastructure for data management and storage.

Increasing data management and sharing activities often requires significant support from personnel outside of the traditional laboratory environment, including librarians and data scientists, to provide the necessary expertise and guidance needed to comply with a data sharing policy and build good data management practices into an investigator's research process. NIH should strongly consider including these additional staff as part of the allowable costs. Again, if this is not doable, it will be necessary for the agency to provide supplemental funding to institutions in building up and maintaining services that support scientific data sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

UC suggests that the policy more clearly define acceptable timeframes for data sharing, as "timely manner" could be widely interpreted. These could be conveyed as ranges to preserve flexibility.

Other Considerations Relevant to this DRAFT Policy Proposal:

UC appreciates that the draft acknowledges the importance of human subject protections without including a mandate for IRBs to verify or otherwise be a gatekeeper for data sharing and management. The NIH Genomic Data Sharing Policy model of IRB involvement is not

appropriate for a broader policy such as this. However, UC does suggest that future NIH guidance emphasize three points:

- It is important that PIs prepare protocols and IRB applications that are consistent with the data management and sharing plans.
- It is important that PIs prepare consent forms that are both consistent with the data management and sharing plans and provide adequate notice to subjects about plans to reuse data.
- Institutions and their IRBs should evaluate policies, instructions, application forms and other similar materials to ensure that the ability to comply with this policy is not inadvertently or unduly restricted.

Attachment:

UC Comment Letter-NIH Draft Data Management and Sharing Policy-1-10-2020.pdf

Description:

UC Comment Letter on NIH DRAFT Data Management and Sharing Policy

January 10, 2020

Dr. Andrea Jackson-Dipina
Director of the Division of Scientific Data Sharing Policy
National Institutes of Health
Office of Science Policy
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

RE: UC Comments in Response to NOT-OD-20-013, “Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance”

Dear Dr. Jackson-Dipina:

I write on behalf of the University of California (UC) system with regard to the Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance issued on November 11, 2019.

The UC system comprises ten research-intensive campuses, six medical schools, and three affiliated U.S. Department of Energy national laboratories. As a system, UC receives approximately \$6 billion annually of extramural awards to support research conducted throughout all UC locations. UC generally receives 5 to 6 percent of the NIH’s annual appropriations for research, making UC the largest single recipient of NIH funding for academic research.

The UC believes that the curation and sharing of research data offers benefits to the larger research community and advances public knowledge. UC supports the NIH’s effort to facilitate data sharing and appreciates the NIH’s recognition of the challenges that come with regulating data generated by a broad research community. There are many aspects of the NIH draft policy that we support and were pleased to see included in the policy such the incorporation of compliance checking and routine reporting of data management and sharing plans, the ability to request NIH funding to support data management, and the ability to make data management and sharing plans publicly available. While UC generally agrees with the NIH’s draft data management and sharing policy, however, we believe that in order to make this policy a success, the NIH should make developing templates, further resources and robust tools to better facilitate data sharing, particularly across scientific disciplines a high priority.

Comments on specific aspects of the draft policy are provided below.

I. Purpose Section

UC appreciates the NIH's philosophy that underlies the draft policy as discussed in the Purpose section. It serves as a helpful reminder that investigators are not conducting their work within a vacuum. Proper research data management will inevitably include not just researchers, but also various institutional research administrative units and departments, along with research funders, regulators and journals equally invested in promoting the dissemination of knowledge within the constraints of legal, ethical, technical, security, and privacy considerations. We suggest that the NIH include in this section a recognition of data preservation and sharing as part of the entire research lifecycle.

II. Definition Section

The definition of "scientific data" requires clarification. According to the draft policy, scientific data is "the recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings regardless of whether the data are used to support scholarly publications." The definition is very broad, but at the same time, excludes specific sources of information. This ambiguity could lead to confusion about how best to interpret "scientific data" and implement proper data management and sharing plans. For example, the definition of scientific data incorporated into the draft policy explicitly excludes laboratory notebooks. Information contained in laboratory notebooks, however, might reasonably be seen as "necessary to validate and replicate research findings." Given its breadth, the proposed definition of scientific data seems open to the interpretation that almost all data collected for a study counts, and therefore should be shared. UC asks that NIH provide clarification on this in addition to examples of accepted data standards.

Furthermore, the NIH's expectation that reasonable efforts will be made to digitize all scientific data be digitized raises concerns to UC. Not all scientific records must be digital to be useful. Digitizing also imposes administrative burden upon institutions, and creates risk, both in loss or error in translation, as well as in of the disclosure or use of sensitive material, including medical information. If external parties are necessary to digitize such data, the risk of loss, error or exposure becomes more pronounced. Because of the potential for heightened risk, the decision to digitize data should be left to the principal investigator's discretion. UC recommends that this expectation be removed.

III. Effective Date

UC asks that the NIH choose a policy implementation date that will allow the research community to prepare sufficiently for the policy change and to consider a reasonable embargo period to advance intellectual property in its Supplemental Draft Guidance for data sharing plan elements (*Section 4, Data preservation, Access and Associated Timelines*). We recommend a minimum implementation date of one year after the release of the final policy, with a delay in enforcement actions for at least one year after the implementation deadline. Any determination of non-compliance should follow well-defined and transparent criteria.

IV. Requirements

While UC supports the requirement of a data management and sharing plan for NIH-funded or conducted research, we are concerned about varied supplementary information requirements requested by individual NIH Institutes, Centers, and Offices (ICOs). To minimize confusion and administrative burdens, we strongly urge trans-NIH coordination of these supplemental requests and listing ICO-specific requirements as part of centralized resources associated with the final data management and sharing policy.

V. Data Management and Sharing Plans

UC appreciates the proposal to collect data management and sharing plans as part of Just-in-Time (JIT) documentation for extramural awards. Requiring submission of the plan at the JIT phase rather than at the proposal stage minimizes administrative burden for both the applicant and peer reviewers. Shifting the review of plans to NIH staff members rather than volunteer reviewers will also make the process more uniform and streamlined, provided NIH staff are properly trained in reviewing data management plans. One potential drawback of submitting data management plans during JIT, however, is that accurately budgeting for data management costs at the time of proposal will be challenging as details of the plan remain to be finalized with the NIH Program staff at JIT. UC, therefore, recommends that the NIH allow additional data management costs to be added to the budget at JIT based on the final negotiated data management plan. We also recommend an option that allows grantees to appeal ICO-mandated data sharing requirements to the NIH Policy Office should the requirements be considered unreasonable or inappropriate by the grantee, without fear of reprisal.

Furthermore, UC strongly urges the NIH to provide data management and sharing templates. Rather than having each researcher develop their own plans in two pages or less, having a common template will streamline the process and reduce burden for both submitters and reviewers. It would also be useful if NIH published sample plans. Within these templates, we recommend that NIH provide minimum requirements for researchers to include in a data management and sharing plan, such as data type, standards and metadata, plans for data preservation, and projected data accessibility. In addition, the template could include the usage of persistent identifiers and tools to generate machine-readable and actionable data management plans that support compliance checking and will facilitate the interoperability of data between systems. At a minimum, it would be useful if the NIH policy required the use of ORCID ID's for all individuals mentioned in the data management and sharing plans and for the use of persistent identifiers for all research datasets.

Finally, UC recommends that the NIH provide guidance on appropriate ways to maintain sensitive data. Not only is this administratively burdensome, but it also introduces a dependence on the researcher (and their current contact information) that undermines the goal of long-term data accessibility. NIH should recommend restricted access repositories as well as other kinds of repositories that so that the researchers understand what qualifies as an "established repository." In this context, the NIH has the unique opportunity to lead the community by creating field-specific data repositories that have the added benefit of ensuring relevant security and privacy concerns are addressed.

VI. Compliance and Enforcement

UC appreciates that the NIH will allow researchers to make updates to their data management and sharing plan. In practice, data management plans may evolve over time as repositories, technologies, and data standards change. We are pleased to see some degree of flexibility built into the plans to address unforeseen issues.

The policy is not clear how adherence to a data management and sharing plan is verified at the end of a project. Requiring a final report on a data management and sharing plan would be an excellent way to enforce compliance and ensure that a researcher has followed through on plans for sharing and depositing data. Adding details how compliance is monitored would give the policy some enforcement capabilities and make sure it is taken seriously.

Additionally, we wondered who would be responsible for reviewing and approving data management and sharing plans during the initial review process. It may be useful to specify the required knowledge and expertise that will be required of reviewers to ensure that they have the necessary backgrounds to effectively analyze and comment upon data management and sharing plans.

VII. Supplemental Draft Guidance – Allowable Costs

UC appreciates the NIH's recognition of the costs associated with data management and sharing and applauds the inclusion of the supplemental guidance defining possible allowable costs. However, UC is concerned that the guidance only addresses those costs incurred during the term of the award but does not address costs associated with long-term data retention and accessibility, such infrastructure costs related to storing data. We strongly urge the NIH to allow researchers to budget for long-term data curation and preservation costs, or at a minimum clarify that grantee institutions may pre-pay from their awards these long-term costs. We would recommend that if these long-term costs are not permitted on a grant-by-grant basis, that the agency offers additional supplemental funding to institutions to develop infrastructure for data management and storage.

Increasing data management and sharing activities often requires significant support from personnel outside of the traditional laboratory environment, including librarians and data scientists, to provide the necessary expertise and guidance needed to comply with a data sharing policy and build good data management practices into an investigator's research process. NIH should strongly consider including these additional staff as part of the allowable costs. Again, if this is not doable, it will be necessary for the agency to provide supplemental funding to institutions in building up and maintaining services that support scientific data sharing.

VIII. Supplemental Draft Guidance – Elements of a NIH Data Management and Sharing Plan

UC suggests that the policy more clearly define acceptable timeframes for data sharing, as “timely manner” could be widely interpreted. These could be conveyed as ranges to preserve flexibility.

IX. Other Considerations Relevant to the Draft Policy Proposal

UC appreciates that the draft acknowledges the importance of human subject protections without including a mandate for IRBs to verify or otherwise be a gatekeeper for data sharing and management. The NIH Genomic Data Sharing Policy model of IRB involvement is not appropriate for a broader policy such as this. However, UC does suggest that future NIH guidance emphasize three points:

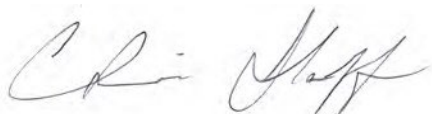
- It is important that PIs prepare protocols and IRB applications that are consistent with the data management and sharing plans.
- It is important that PIs prepare consent forms that are both consistent with the data management and sharing plans and provide adequate notice to subjects about plans to reuse data.
- Institutions and their IRBs should evaluate policies, instructions, application forms and other similar materials to ensure that the ability to comply with this policy is not inadvertently or unduly restricted.

Thank you for the opportunity to comment on this important issue and we look forward to continued engagement on this issue as the data sharing and management draft policy and other guidance is developed.

Sincerely,



Lourdes G. DeMattos
Associate Director
Research Policy Analysis & Coordination
Office of Research & Innovation



Chris Shaffer
University Librarian and Assistant Vice Chancellor
University of California San Francisco

Submission ID: 1432

Date: 1/11/2020

Name: Rajni Samevedam, MPH, Principal/Director

Name of Organization: Booz Allen Hamilton

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Other

Type of Organization - Other: Government consulting firm

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

Multiple domains of life sciences

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Booz Allen supports FAIR data sharing by all publicly funded researchers; in 2019, we joined the Global Alliance for Genomics and Health (GA4GH) as a measure of our commitment to this principle and to ensure we stay current with the standards and APIs being developed by GA4GH.

We note that Section 2014 of the 21st Century Cures Act authorizes the NIH Director to require award recipients to share data in a manner consistent with applicable federal regulations, an endorsement of NIH's existing data sharing policy for grants in excess of \$500,000 and a statutory basis for expanding both the scope and increased enforcement of these policies. The recent NIH Strategic Plan for Data Science also supports storage and sharing of individual datasets under the Modernized Data Ecosystem overarching goal and encourages developing policies for a FAIR data ecosystem under the Stewardship and Sustainability overarching goal. For the former goal, developing NIH data repositories for datasets is the plan, although small data sets can be linked to publications via PubMed Central and NCBI, making data sharing much easier.

We support NIH in providing details as to what needs to go into a data management and sharing plan; perhaps example plans for different common data types, or even sample templates would enhance the guidelines provided. In addition, Booz Allen recommends that Security of the data is as important as Findability, Accessibility, Interoperability, and Reusability, and should not be an afterthought, especially as precision medicine moves into an era where such data will be clinically actionable – and thus cannot be deidentified. Privacy of patient data should be paramount, and not given short shrift in the data sharing plans. However, we also recognize that some data may be difficult or even impossible to share for sensitive privacy, ethical, or previous agreement reasons. For the privacy concern, due to the advent of advanced analytics, it is becoming increasingly possible to assemble disparate data from multiple sources and identify a cohort participant even after de-identification of the data. Therefore, security need to be considered even before data collection and sharing. To mitigate this risk, it may soon be feasible using techniques like homomorphic encryption to analyze data while it remains encrypted.

NIH should emphasize to the research community that the further upfront they embed good data practices, use detailed and carefully chosen metadata with use of common data elements where possible, and make use of standards, the better for the usability of data from the project – doing these items will fundamentally improve the ability to use the data and harmonize it for potential use in combination with other studies. When this is attempted retrospectively, it becomes very problematic, and much more labor-intensive.

However, this begs the question of how NIH can help address this new need for data management for all its awardees? No doubt much of the details in these plans will be new to many researchers. Some form of NIH-offered training on data management, best practices, and even publicly available computational resources might make this policy much more immediately palatable to the community – especially for less well-funded researchers and those at smaller institutions. In addition, common training including notions of what constitutes good metadata and documentation would improve data structures and the use of metadata, so critical for data harmonization.

NIH should be actively promoting the consideration and adoption of full data lifecycles – from data creation through to final archiving. Also, our experience is that scientists strongly prefer unlocked data for sharing and full use by end researchers to allow maximum flexibility in data analysis, and not to tie it up in the hands of middlemen required to analyze and produce interpreted data upon request.

We also note that discovery and cataloging resources will also need to be improved to be able to find all this shared data, perhaps by increased funding and outreach of efforts such as the DataMed prototype developed by the bioCADDIE project originally begun under the BD2K program.

Section II: Definitions:

Whenever possible, research data needs to be digitized in a way that it is consumable by and computable with analytic tools – and not just scans of data into PDFs as unsearchable images

Section III: Scope:

In the definition of what funded research the policy will apply to, perhaps NIH should explicitly mention OTAs and cooperative agreements, and explicitly exclude training grants. Will the policy also be applied to SBIR and STTR grants that generate research data? If so, that should be stated as well.

Section IV: Effective Date(s):

Section V: Requirements:

Booz Allen suggests explicitly adding requirements and considerations for archiving of scientific data, once the frequency of direct access to shared data tails off to a low level. We make a distinction here between placement of data in a repository for active data sharing, and archiving of the data for long term storage, less accessibly. While such archiving is not always necessary, data storage space in most repositories is finite, and less used large datasets may need to be stored differently as part of the data lifecycle. Perhaps in the data sharing plan, a timeline for this inflection point in the data lifecycle can be added, and the plan be required to be revised as warranted.

We also applaud allowing these costs being made allowable under the budget; too often data management has been an unfunded mandate. However, we suggest that NIH also try to give guidance to the investigators as to what kinds of items will be deemed acceptable as allowable costs, and a rough order of magnitude of such costs for common data types to aid in setting realistic budgets.

Section VI: Data Management and Sharing Plans:

Booz Allen urges NIH to look at including data lifecycle considerations in the data sharing plans – especially with guidance on deciding when data should be archived and no longer actively shared, which would most likely be done on declining usage statistics. It could also be based on the return on investment (ROI) calculated after a significantly long period of time, say 10 years after deposition into an active sharing repository.

We are against discarding any older data if the cost burden allows, depending on size. If the data can be archived on local disk storage with sufficient backup provisions or burned to one or a few CD-ROMs or large memory sticks where there are no additional costs for storage, it should be kept indefinitely. For data sizes in the terabyte or petabyte range, this is not feasible, but there are archival storage tiers in the cloud providers' offerings (e.g. S3 Glacier Deep Archive for AWS or Microsoft Azure blob storage) that would allow retention of the data at low cost.

Data security and privacy are important considerations; it may be helpful for NIH to provide some guidance for this for common data types. While human research data is most often de-identified, genomic data is inherently self-identifying, and other security methods (e.g. encryption) may be necessary, especially for translational work where the data is clinically actionable.

Section VII: Compliance and Enforcement:

If this policy is indeed applied to all NIH-funded research, that could be well in excess of 11,000 grants/R&D contracts/OTAs per year. We ask whether NIH has enough staff to adequately review that number of data sharing plans. Are program officers and review staff being trained to judge the adequacy and compliance of plans with the policy, preferably with a set of basic standard criteria? We suggest some standard criteria below. We also suggest that reports on plan compliance with such standard criteria be made periodically across the NIH, so an evaluation can be made of how well this aspect of the data sharing policy is working.

Suggested Standard Criteria:

- Primer for common data types (clinical, genomic, proteomic, transcriptomic, imaging, etc.) and issues to consider with each.
- Guidance as to approximate costs of data storage per terabyte.
- Adequacy of metadata – how comprehensive does it appear? Is it time/date stamped appropriately? Does the data and metadata use any standards? Does it make use of Common Data Elements?
- Are the data and metadata well documented? Is there a data dictionary provided?
- Is security and privacy of the data considered at appropriate levels?
- Are identifiers being used consistent with FAIR?

- Where will the data and metadata be placed for data sharing? Can a public data repository be used? If local to the data generator, is access, disaster recovery, etc. adequate?
- Are needed tools publicly available, preferably in a GitHub or similar software versioning environment?
- How will controlled access data be handled? Are there any agreements for use of the data that must be taken into consideration?
- Timeframes for data sharing, including how long after generation the data will be made available. Are there any limits on the data to be shared, and are those reasonable?
- Who is responsible for the data management? Does that person have adequate knowledge and experience in data management?

It is critical that NIH actively enforce this data sharing policy. Too often, data generated with NIH funds has tended to be viewed as the property of the Principal Investigator on that award. This notion is increasingly deprecated - publicly funded data should be made publicly available, within a reasonable (but fairly short) time limit to give the data generators sufficient time to 1) analyze that data first and 2) get it accepted for publication. There remains some lingering reluctance to share data. In an NIH program with which we were involved, there were line items in FFP contracts for data deposition, yet some awardees would not do it, and were even willing to let that funding lapse. Also, with one of the data sharing platforms we've developed for an NIH IC for genomic and proteomic data, where the PI of the program was from a major research university, the IC had to have NIH lawyers enter into discussions with the university lawyers in order to get the data released. For this same IC, for a viral study with a university and another government agency, a case was made to not put the data into a data sharing platform due to its sensitivity (despite adequate security in the data sharing system).

NIH might want to consider having a more formal Designated Approving Authority (DAA) in the future from the NIH side that would provide approval and guidance on data classification, sharing and retention for each IC, or each research project, as each IC or research project can have varying data sharing/retention requirements. We and other consulting firms can recommend and implement solutions that meet NIH's data sharing and retention requirements.

If the data sharing plans were made publicly available post-award, there might well be a sort of crowdsourced response to data sharing plans with any perceived potential for abuse. We believe that making the plans public will also help improve their quality over a short period of time due to feedback from the readers. No one would want to go through a 5-year grant cycle before ensuring that the metadata practices used were done well for an award.

We applaud NIH's notion that non-compliance after the end of the funding period be considered for future funding decisions for the institution; we suggest that in addition this should also be at the level of individual PI's.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Does NIH have a good understanding of the magnitude of the data sharing costs, now that these policies will apply to all NIH-funded research?

Booz Allen strongly supports data curation, which at a general level is looking for any detectable inconsistencies in a data set by both automated and manual means as required, as well as having detailed supporting documentation for data; these are far too often neglected. We would like to see NIH expand a bit on "accepted community standards", perhaps by giving a few examples for common data types.

For data storage on the cloud, storage costs go on indefinitely, which is an issue once grant funding for data collection, analysis, and sharing ceases. Archival tiers of storage can be used that would substantially lower, but not eliminate, these costs. It is conceivable NIH will need to maintain an archive for data sets whose usefulness extends many years after the original award that generated that data. Determining which data sets warrant such archival treatment will likely need to be done on a case-by-case basis, considering the estimated or demonstrated value to the community, expressed demand for continued access to the data sets, the amount of continuing costs, and projections of how much longer the data is likely to remain useful.

Booz Allen has concerns over "unique and specialized information infrastructure necessary to provide local management, preservation, and access to data..." to ensure that such localized storage does not in any way substitute for reasonably rapid deposition into established, more public repositories. That local infrastructure may not be easy to access from outside that institution, nor be able to be accessed in a standardized way; in our opinion, localized infrastructure should be part of the data gathering process, not the data sharing. This is a fine line and could lead to substantial ongoing costs for such unique, specialized information infrastructure. Any local data sharing may also need disaster recovery plans to prevent research data loss.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Data Sharing plans will require more than 2 pages for adequate detail – the description of the plan elements alone herein take up nearly three pages; setting expectations for 3-4 pages is

more likely, along with a flowchart, checklist, or well-documented workflow with milestones to ensure all requirements of the data sharing process are covered. As noted previously in our comments, a template for the plans might be crafted to reduce free text entry in favor of selecting from a set of approved options for data sharing for common data types where feasible.

We suggest that use of "To Be Determined" in the Plan be both discouraged and minimized. The Data Sharing Plan should be essentially complete to be acceptable to NIH.

1. Data Type

- The second bullet point, "Providing a rationale for decisions about which scientific data are to be preserved and made available for sharing, taking into consideration scientific utility, validation of results, availability of suitable data repositories, privacy and confidentiality, cost, consistency with community practices, and data security." concerns us. Allowing researchers to define the rationale behind what is to be shared can be a slippery slope – that can easily be turned into a loophole to avoid data sharing, or to make only selected data available. Sharing all or as much of the data as possible should be the default case, but it will be difficult in some cases to know where to draw the line on what is to be shared, especially due to costs for very large data sets as well as adequate data security for sensitive data with above average privacy concerns.

As we've asked elsewhere, are program officers and review staff being trained to judge the adequacy and compliance of plans with the data sharing policy, preferably with a set of basic standard criteria? We suggest these criteria would include:

- Primer for common data types (clinical, genomic, proteomic, transcriptomic, imaging, etc.) and issues to consider with each.
- Guidance as to approximate costs of data storage per terabyte.
- Adequacy of metadata – how comprehensive does it appear? Is it time/date stamped appropriately? Does the data and metadata use any standards? Does it make use of Common Data Elements?
- Are the data and metadata well documented? Is there a data dictionary provided?
- Is security and privacy of the data considered at appropriate levels?
- Are identifiers being used consistent with FAIR?

- Where will the data and metadata be placed for data sharing? Can a public data repository be used? If local to the data generator, is access, disaster recovery, etc. adequate?
- Are needed tools publicly available, preferably in a GitHub or similar software versioning environment?
- How will controlled access data be handled? Are there any agreements for use of the data that must be taken into consideration?
- Timeframes for data sharing, including how long after generation the data will be made available. Are there any limits on the data to be shared, and are those reasonable?
- Who is responsible for the data management? Does that person have adequate knowledge and experience in data management?

2. Related Tools, Software, and or Code

- We suggest that open source code or tools be required to be placed into a GitHub repository that is freely accessible to the research community, along with documentation as to any parameters used in the analysis of the data and versioning of the code/tools. We suggest minimizing having tools and code "available only from the research team..."
- Automation of data extraction, transformation, and loading should be encouraged to minimize errors in loading data into a database manually [1].
- NIH should also encourage wide reuse of code when possible. This lowers costs and often improves interoperability.

[1] Our chief medical officer and other Booz Allen leadership recently published an article in the New England Journal of Medicine (Big Data and the Intelligence Community — Lessons for Health Care, Kevin Vigilante, M.D., M.P.H., Steve Escaravage, M.S., and Mike McConnell, M.P.A. (2019) NEJM 380:20, 1888-1890) that looked at tools the intelligence community uses that could be applicable to health data. They suggest adopting less structured data-storage technologies, such as data lakes, to integrate disparate data with minimal data modeling coupled with incorporating automated metadata labeling to characterize properties of data. Also, natural-language processing solutions could be used to translate and structure text for analysis.

3. Standards

- NIH should make an even stronger push for the use of standards, especially for controlled vocabularies and ontologies. Perhaps NLM could be engaged to help provide national resources for standards-based controlled data elements to be used.

- We also suggest inclusion in the given examples of the Global Alliance for Genomics and Health (GA4GH) evolving standards, as well as the FHIR standard being supported by NIH (cf. the recent NIH FHIR comments request).

- NIH should ask for well-documented APIs whenever possible to access the data.

4. Data Preservation, Access, and Associated Timelines

- We suggest striking the word "consider" from the first bullet point. The Data Sharing Plan should clearly state where the data is being placed for sharing and for eventual archiving, with a requirement to update the plan if the location or method of archival storage changes. This is critical for federated data sharing/access and for interoperability.

- We would strongly encourage the use of GUIDs/UUIDs for data as the primary identifier, with secondary identifiers being used to make the data more human-readable.

- The need for restricted access data should be minimized wherever possible. Use of the current authentication via dbGaP should be encouraged for now, until a better authentication and authorization mechanism such as NIH RAS is more widely available.

- Data should be made available to other users after a reasonable time for analysis and arranging for publication of the initial paper by the data generators. Having the data made available upon journal submission risk undermine the data generator's rights to publish first on such data.

5. Data Sharing Agreements, Licenses, and Other Use Limitations –

- NIH should ensure as much as is practical that such agreements or limitations were not made with the primary intention of blocking data sharing, but that there are legitimate scientific or privacy reasons behind such agreements.

6. Oversight of Data Management –

- We suggest that each project designate a data manager who will be responsible for the execution of the data sharing plan, and that the data manager be a different individual than the PI in most cases, with a preference for long term laboratory members with sufficient technical proficiency for data management. We are concerned that many awardees may not have staff qualified to be doing the data engineering that may be required to share data well.

Other Considerations Relevant to this DRAFT Policy Proposal:

If a set of standard criteria were to be used for the data sharing plans, against which investigators are scored, NIH could calculate a kind of 'data sharing index' for PIs, that might function as a pseudo- reward mechanism and provide feedback to the investigators as to the quality of their data sharing plans.

Attachment:

Description:

Submission ID: 1434

Date: 1/11/2020

Name: Amanda Gentzel

Name of Organization: University of Massachusetts Amherst

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Causality

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I am encouraged to see the NIH's commitment to data sharing. As a researcher primarily interested in methods to estimate causal effect, access to data to evaluate these methods is vitally important, and improving the accessibility of data from NIH-funded and conducted research will be a great boon to the field.

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

The draft plan states that "NIH encourages the use of established repositories for preserving and sharing scientific data." The primary difficulty I have faced with accessing data for use with causal methods is non-standardized ways to access it. Many papers that use collected data state that their data is available upon request. However, the need to individually contact authors to find out if and how I can get access to the data significantly limits the speed and ease with which data can be accessed in practice. The use of existing repositories, or the creation of new specialized repositories, would be a great help in standardizing data access. As someone

who is a part of the causal modeling community, and not part of a medical or biological research community, I am highly interested in using data from clinical trials and other experimental studies, but I don't know what repositories are favored by those communities. I would hope, then, that the NIH could provide information to the greater scientific and public communities on what repositories are being used to store data from NIH-funded or conducted projects.

The current proposal states that the NIH may make plans publicly available. I would encourage them to do so if possible, and to take any other reasonable steps to create a centralized space where researchers can see which data sets have been released from NIH-funded projects, and how to access them.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1435

Date: 1/11/2020

Name: Katherine Boronow and Julia Brody

Name of Organization: Silent Spring Institute

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Environmental Health (includes genomic and clinical data as well as exposure data such as occupation, housing information, and chemical measurements)

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Environmental health

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Summary: Environmental health study data are vulnerable to re-identification, the process of attaching a name or address to a person's data after overt identifiers are redacted. Data types including housing characteristics, demographics, location, genetics, medical records, and employment are vulnerable because they can be matched to publicly-available data that contain identifiers. For example, genetics data have been re-identified using ancestry websites. In NIH-supported research, we demonstrated that chemical exposure measurements alone could be used with 80-98% accuracy to cluster study participants by state or city, showing how a characteristic useful for re-identification could be inferred even if it is excluded when study data are shared. The risks of re-identification by linking study data to outside data are poorly

understood but may be significant. Until these risks are clarified, we urge caution about sharing research data without IRB oversight or protection in a registered-access data enclave.

Comment

As scientists at Silent Spring Institute, an environmental health research organization dedicated to breast cancer prevention, we recognize the value of sharing data to maximize its contribution to science and health. However, we caution NIH to ensure that its data sharing and management Policy actively protects the privacy of research participants who have donated their personal data to scientific research.

The draft Policy requires researchers to outline how they will protect human participants' privacy by using methods like de-identification. However, it does not provide practical guidance on how to fulfill their responsibilities. We believe this is a serious gap.

A primary concern about "de-identified" data is that study participants can be re-identified—when identifiable information sufficient to contact a person (such as name or address) is matched to "de-identified" data. If study participants have been promised in the informed consent that their confidentiality will be protected—as in nearly all human subjects studies—re-identification is a breach of that promise. It can result in material harm, such as employment or insurance discrimination, property value loss, or legal repercussions, as well as stigma or loss of dignity.

One strategy for re-identification is data linkage, whereby data containing identifiers are matched to "de-identified" data using other fields that are common to both data sets. Our NIH-supported research and other studies have demonstrated that environmental health research data are vulnerable to this type of attack.

To empirically evaluate the risk of re-identification in our own data, we conducted an experiment in which we shared anonymized data from Silent Spring's Northern California Household Exposure Study with Harvard University researchers skilled in re-identification techniques. By linking housing and demographic information from the study to publicly-available tax assessor data, and using information from the published study, the re-

identification team correctly re-identified 25 percent of participants by name from one housing development [1].

To better understand the scope of re-identification risks in environmental health data more generally, we convened an Advisory Council and assessed 12 major environmental health studies for features that may contribute to re-identification [2]. We found that all of the studies included at least two of five data types that overlap with outside datasets: geographic location (9 studies), medical data (9 studies), occupation (10 studies), housing characteristics (10 studies), and genetic data (7 studies). Seemingly non-sensitive or non-protected data (such as data related to housing, occupation, daily activities, or product use) can lead to breaches of sensitive information (like personal health records or personal chemical measurements) through linkage re-identification. Examples of databases that could be used in data linkage include voter lists, tax and real estate data, professional licensing lists, and ancestry websites. New risks will arise as databases move online and personal sensors become more affordable.

To investigate how environmental measurement data (for which there are currently few public repositories of matching data) might contribute to re-identification risk, we also conducted a cluster analysis to learn whether participants' region of residence could be inferred from measurements of chemicals in household air and dust [2]. Using data from Silent Spring's Household Exposure Study in California and Massachusetts, and the CDC's Green Housing Study in Boston and Cincinnati, we used k-means cluster analysis of the raw chemical measurements to partition the data from each study. The groups identified by the algorithm corresponded to geographic location with 80-98% accuracy, demonstrating the possibility of inferring a feature of the data even if it is not shared in a public dataset. Subgroup information can increase the likelihood of successful linkage by narrowing the pool of potential matches.

To our knowledge, the only Federal regulation that provides explicit guidance on methods for de-identification is the Health Information Portability and Accountability Act (HIPAA) Privacy Rule, which governs sharing of health information. The Safe Harbor provision is a prescriptive method of de-identification that requires the removal of certain overt identifiers (e.g., name, street address, date of birth) and makes no numerical assessment of the privacy risk. However, as demonstrated by our research [1] and others', Safe Harbor "de-identification" may not adequately protect against re-identification. Other technical approaches, such as k-anonymity, are not practical and may result in data that are no longer useful for their intended purpose.

It is clear to us that nearly all NIH researchers will lack the technical expertise to produce a truly "de-identified" dataset, and that asking researchers to do so poses an undue burden.

Therefore, we ask NIH to shoulder the responsibility of defining what constitutes a "de-identified" human subjects data set in their Policy. The definition should improve upon HIPAA, such that the risk of re-identification is quantifiably low.

Because many types of data will not meet a restrictive definition of "de-identified" and also retain its utility, we recommend caution in making human subjects data available for sharing outside of IRB oversight—even after overt personal identifiers are removed. IRB oversight requires that the recipients of the data have completed human subjects ethics training, creates a record of persons having had access to the data, and provides a vehicle for consequences if data are misused.

Another model is the NIH All of Us Study, which has two tiers of data access to an online research hub. The public tier contains only anonymized, aggregate data. To access the registered tier, which includes more robust data, researchers will have to register, complete research ethics training, and sign a data use agreement. Following a similar model, NIH could establish a centralized registered-access data repository for use by awardees across all institutes and centers.

Finally, we urge NIH to support federal policies that provide legal recognition of the sensitivity of research data and protect study participants from harm due to privacy loss. Policies such as the HIPAA Privacy Rule and Genetic Information Nondiscrimination Act of 2008, which protects people from insurance and employment discrimination based on genetic data, however imperfect, can be models for parallel policies for environmental health and other research data.

1. Sweeney L, et al. 2017. Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. *Tech Sci* 2017082801.
2. Boronow K, et al. 2020. Privacy Risks of Sharing Data from Environmental Health Studies. *Environ Health Perspect* 128:1.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Please refer to our comment in Section VI: Data Management and Sharing Plans. In brief, the draft Guidance requires researchers to outline how they will protect human participants'

privacy by using methods like de-identification. However, it does not provide practical guidance on how to fulfill their responsibilities. We believe this is a serious gap.

To our knowledge, the only Federal regulation that provides explicit guidance on methods for de-identification is the Health Information Portability and Accountability Act (HIPAA) Privacy Rule, which governs sharing of health information. The Safe Harbor provision is a prescriptive method of de-identification that requires the removal of certain overt identifiers (e.g., name, street address, date of birth) and makes no numerical assessment of the privacy risk. However, as demonstrated by our research [1] and others', Safe Harbor "de-identification" may not adequately protect against re-identification. Other technical approaches, such as k-anonymity, are not practical and may result in data that are no longer useful for their intended purpose.

It is clear to us that nearly all NIH researchers will lack the technical expertise to produce a truly "de-identified" dataset, and that asking researchers to do so poses an undue burden. Therefore, we ask NIH to shoulder the responsibility of defining what constitutes a "de-identified" human subjects data set in their Guidance. The definition should improve upon HIPAA, such that the risk of re-identification is quantifiably low.

Because many types of data will not meet a restrictive definition of "de-identified" and also retain its utility, we recommend caution in making human subjects data available for sharing outside of IRB oversight—even after overt personal identifiers are removed. IRB oversight requires that the recipients of the data have completed human subjects ethics training, creates a record of persons having had access to the data, and provides a vehicle for consequences if data are misused.

1. Sweeney L, et al. 2017. Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. Tech Sci 2017082801.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Boronow_KE2020.pdf

Description:

Research article describing privacy risks of environmental health data

Privacy Risks of Sharing Data from Environmental Health Studies

Katherine E. Boronow,¹ Laura J. Perovich,^{1,2} Latanya Sweeney,³ Ji Su Yoo,³ Ruthann A. Rudel,¹ Phil Brown,⁴ and Julia Green Brody¹

¹Silent Spring Institute, Newton, Massachusetts, USA

²MIT Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

³Department of Government, Harvard University, Cambridge, Massachusetts, USA

⁴Department of Sociology and Anthropology and Department of Health Sciences, Northeastern University, Boston, Massachusetts, USA

BACKGROUND: Sharing research data uses resources effectively; enables large, diverse data sets; and supports rigor and reproducibility. However, sharing such data increases privacy risks for participants who may be re-identified by linking study data to outside data sets. These risks have been investigated for genetic and medical records but rarely for environmental data.

OBJECTIVES: We evaluated how data in environmental health (EH) studies may be vulnerable to linkage and we investigated, in a case study, whether environmental measurements could contribute to inferring latent categories (e.g., geographic location), which increases privacy risks.

METHODS: We identified 12 prominent EH studies, reviewed the data types collected, and evaluated the availability of outside data sets that overlap with study data. With data from the Household Exposure Study in California and Massachusetts and the Green Housing Study in Boston, Massachusetts, and Cincinnati, Ohio, we used *k*-means clustering and principal component analysis to investigate whether participants' region of residence could be inferred from measurements of chemicals in household air and dust.

RESULTS: All 12 studies included at least two of five data types that overlap with outside data sets: geographic location (9 studies), medical data (9 studies), occupation (10 studies), housing characteristics (10 studies), and genetic data (7 studies). In our cluster analysis, participants' region of residence could be inferred with 80%–98% accuracy using environmental measurements with original laboratory reporting limits.

DISCUSSION: EH studies frequently include data that are vulnerable to linkage with voter lists, tax and real estate data, professional licensing lists, and ancestry websites, and exposure measurements may be used to identify subgroup membership, increasing likelihood of linkage. Thus, unsupervised sharing of EH research data potentially raises substantial privacy risks. Empirical research can help characterize risks and evaluate technical solutions. Our findings reinforce the need for legal and policy protections to shield participants from potential harms of re-identification from data sharing. <https://doi.org/10.1289/EHP4817>

Introduction

The trade-off between sharing personal data and the risks to privacy has become an everyday concern for consumers as social networks, search engines, ride-sharing apps, credit cards, and smart home devices, for example, ask consumers to “allow” access to location, purchases, internet searches, and more. Privacy researchers have demonstrated that diverse types of data—for example, movie rentals (Narayanan and Shmatikov 2008), bicycle shares (Pennarola et al. 2017), electricity meter readings (Buchmann et al. 2013), and hospital visits (Sweeney 2013)—can be linked back to individuals, even when they are shared without names or other overt identifiers such as address or exact birth date. In 2013, researchers demonstrated that surnames can be identified from genetic sequencing data (Gymrek et al. 2013), and, recently, genetics data posted by consumers on genealogy websites were used to identify crime suspects (Justin 2018)—such databases may soon have sufficient coverage to facilitate re-identification (re-ID) of nearly any American of European descent (Erlich et al. 2018). Most recently, Rocher et al. (2019) estimated that nearly all Americans can be identified in any data set by using 15 demographic attributes. This process of linking

“de-identified” data that lack obvious personal identifiers, such as name, birth date, or address, back to one individual or a few likely matches is referred to as re-ID. Reports of successful re-ID increased rapidly in the last decade, although a recent review found that only 8 of 55 reports were published in academic journals, limiting dissemination of these risks to the broader research community (Henriksen-Bulmer and Jeary 2016).

Re-ID is increasingly relevant to environmental health (EH) research, because of growing pressures to share data, more personalized exposure measurements, and rapidly expanding repositories of public and commercial data. Environmental exposure measurements are often individual- or home-specific, such as chemical biomonitoring data or measurements in personal spaces. The advent of wearable sensors (e.g., smartphones and devices like Fitbit[®]) that continuously collect data such as location, exposure, and biometrics creates added vulnerability. In addition, EH studies can include genetic, medical, or household data that are themselves vulnerable to re-ID, creating disclosure risks for the entire data set. Loss of privacy from re-ID could result in stigma for individuals and communities; affect property values, insurance, employability, and legal obligations; or reveal embarrassing or illegal activity (Goho 2016; Zarate et al. 2016). It could damage trust in research, harming the study and research more generally. Because EH studies often focus on groups with the highest exposures, privacy risks potentially compound harms faced by the most vulnerable communities. Entities that might be motivated to re-identify EH data include, for example, employers or insurance companies (who may wish to discriminate against individuals or properties on the basis of environmental exposures) and corporations affected by environmental regulations (who may wish to discredit litigants or studies demonstrating EH harms, or to discourage participation in EH research). Other parties might leverage the environmental variables to gain access to other parts of the data set, such as sensitive health information.

At the same time, researchers, funders, the public, and study participants may want to share data to maximize its value to science and health. Biological and environmental samples are

Address correspondence to Julia Green Brody, Silent Spring Institute, 320 Nevada St., Ste. 302, Newton, MA 02460, USA. Email: Brody@silentspring.org
Supplemental Material is available online (<https://doi.org/10.1289/EHP4817>).

L. Sweeney owns Privacert, a for-profit company that issues determinations about whether data are sufficiently deidentified in accordance with specific legal and regulatory requirements. The other authors declare they have no actual or potential competing financial interests.

Received 30 November 2018; Revised 4 December 2019; Accepted 5 December 2019; Published 10 January 2020.

Note to readers with disabilities: *EHP* strives to ensure that all journal content is accessible to all readers. However, some figures and Supplemental Material published in *EHP* articles may not conform to 508 standards due to the complexity of the information being presented. If you need assistance accessing journal content, please contact ehponline@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.

expensive to collect and analyze. Sometimes they are nonrepeatable, for example, in the case of samples collected after environmental disasters. Sharing data creates new opportunities to gain knowledge from an initial and often public investment. Data sharing can also facilitate the creation of larger and geographically and demographically more diverse data sets. Research consortia, such as the Environmental Health Influences on Child Health Outcomes Program (ECHO), and, ultimately, large-scale research, such as the All of Us Study, are specifically designed around the concept of multiple researchers pooling data collected using common protocols (NIH 2017, n.d.-c).

Paradigms for data sharing span a continuum from agreements among researchers under Institutional Review Board (IRB) oversight—consistent with strong pledges of confidentiality to study participants—to data without overt identifiers shared without restriction and without IRB oversight, and, in its most open form, publicly. Public research funding is increasingly tied to open access requirements for digital data. In 2013, the White House Office of Science and Technology Policy mandated that U.S. federal agencies develop plans to make scholarly publications created with federal funds—and their underlying digital data—publicly available (Holdren 2013). Since then, 22 agencies have issued public access plans, including the National Institutes of Health (NIH) and the U.S. Environmental Protection Agency (U.S. EPA) (<https://www.science.gov/publicAccess.html>). The European Commission has issued similar open access recommendations for publicly funded research (European Commission 2018), leading to implementation efforts such as the Horizon 2020 Open Research Data pilot and European Open Science Cloud (European Commission 2017, n.d.). Scholarly journals also favor policies requiring access to data to ensure the rigor and reproducibility of statistical analyses [e.g., the American Association for the Advancement of Science (AAAS n.d.)].

In the United States, scientists have faced particular pressure from government officials and private interests to make data publicly available from studies used to support regulatory decisions. In the late 1990s, the tobacco industry sought to discredit scientific findings through a multipronged “sound science” campaign that included efforts to legislate data access to federally funded research (Baba et al. 2005). At the same time, industries affected by air pollution standards sought raw, individual data from the NIH-funded Six Cities Study, which was cited in regulations under the Clean Air Act (Fischer 2013). In response to these coordinated efforts, the U.S. Congress passed a law in 1999 that included a provision often referred to as the Shelby Amendment, which extended the reach of Freedom of Information Act (FOIA) requests (Fischer 2013). Previously, the U.S. Supreme Court had found that data from federal grantees were not subject to FOIA requests (in contrast to data from intramural agency research) (Fischer 2013). With the Shelby amendment, federal grantees at outside nonprofit institutions are required to provide data in response to FOIA requests when the research was used to support regulation (OMB 1999). More recently, in 2018, the U.S. EPA issued a proposed rule titled, “Strengthening Transparency in Regulatory Science,” which would require data to be publicly available as a precondition for using it to support regulatory decisions (EPA 2018). The proposed rule raised concerns among researchers about threats to participants’ privacy, violations of assurances in informed consent, and barriers to recruiting participants for future studies (Schwartz 2018). The U.S. EPA is expected to issue a supplemental proposed rule on this topic in 2020 (Science and Technology at the Environmental Protection Agency 2019).

Although all the policies discussed include stated exemptions to protect personal privacy, they do not provide guidance

regarding which data constitute a risk of privacy violations. As expectations for publicly available data increase, EH researchers and decision-makers need to better understand privacy risks in order to make responsible choices that optimize data sharing while protecting privacy. These issues came to our attention through our Household Exposure Study (HES). The HES measured 87 endocrine-disrupting compounds (EDCs) in 120 people and their homes on Cape Cod in Massachusetts in the first comprehensive report on indoor exposure to these chemicals (Rudel et al. 2003), and the study later expanded to California (Brody et al. 2009; Rudel et al. 2010). The study also collected information about demographics, housing characteristics, and behaviors that might be related to the chemical measurements. These data provided a unique resource for understanding exposure and health risks from consumer product chemicals, and the U.S. EPA staff approached us about sharing the data online in ExpoCast™ (Cohen-Hubal 2009). We were uncertain whether the data might be vulnerable to re-ID, so we began an empirical investigation of that risk.

A common re-ID strategy uses linkage of two or more data sets with overlapping fields (Sweeney et al. 2017). When study data overlap with externally available data sets, the combined data can be used to re-ID participants. Using this approach, we conducted a re-ID experiment using data from the Northern California HES in Bolinas, California, and in selected neighborhoods of Richmond, California (Sweeney et al. 2017). HES researchers first redacted the data set to exclude information that cannot be shared under the Safe Harbor provision of the U.S. Health Information Portability and Accountability Act (HIPAA) and removed or aggregated other variables that we considered to be vulnerable—based on our knowledge of which data were likely to be publicly or commercially available—while maintaining the scientific utility of the data. For example, we removed data on pet ownership and aggregated dates, such as the year the home was built, into categories. The remaining data set included chemical measurements from the homes and variables such as race, gender, birth year, home ownership, square footage of living area, number and types of rooms, and decade group for when the house was built and when the participant moved in. Using information in the peer-reviewed articles from the study (Brody et al. 2009; Rudel et al. 2003, 2010), property data from local tax assessor records, and person-level information acquired from public data brokers, the re-ID team, led by L. Sweeney, was able to correctly separate records for residents of Bolinas from records for Richmond residents, and they correctly and uniquely identified 8 of 50 participants (16%) by name in a data set that met HIPAA requirements. One participant (2%) was correctly named even when the data set was further redacted to birth year by decade (Sweeney et al. 2017). Matches associate a name to a study participant’s record, which includes environmental exposure measurements in this study and could include protected health and genetic information or other personal data in other studies.

To further understand how data linkage (also called record linkage) can contribute to privacy risks in EH data, we conducted two additional investigations reported here. We sought to evaluate how data in well-known EH studies may be vulnerable to linkage: a) We evaluated whether data types in 12 major environmental studies are currently available as part of public or commercial registries and therefore potentially pose privacy risks from linkage, and b) We used clustering methods to investigate how environmental measurements, which themselves are not currently vulnerable to data linkage, might contribute to re-ID by identifying subgroups (e.g., location, race, disease status) within a data set so that linkage is limited to smaller numbers of possible matches. For example, Gymrek et al. (2013) used U.S. state of residency as part of the strategy to match individuals to genetic

data. In the Northern California HES, partitioning the data set by city (Richmond vs. Bolinas) and housing development (Liberty Village vs. Atchison Village) was crucial to the success of re-ID. Could similar partitions be achieved with the chemical measurement data alone? Answering this question is helpful for evaluating whether investigators must consider the possibility that variables such as location can be reconstructed from environmental measurements after they have been redacted.

Through these investigations, we aim to determine the availability of vulnerable data types in EH studies and assess additional risk of re-ID introduced by sharing detailed environmental sampling results. Understanding re-ID risks can contribute to realistic informed consent statements and the development of privacy-preserving research policies and practices.

Methods

EH Data Types That Are Vulnerable to Linkage

We selected 12 EH studies to evaluate for the presence of data vulnerable to linkage to existing public or commercial registries. We first selected candidate studies, based on our knowledge and experience as NIEHS-supported EH researchers, as examples of significant EH studies that have made or are expected to make important contributions to the field. Eleven studies were chosen to include a range of scenarios in EH. Because we were motivated by the HES, we included other studies of household exposures in addition to biomonitoring studies. We visited the websites of the selected studies and recorded information about the data that are publicly available or are offered for restricted sharing with other researchers. We sent our descriptions of the studies to the investigators to verify their accuracy. We categorized the data in the studies into broad categories of data types that are important for EH studies, because they represent exposures or outcomes of interest. To focus our investigation, we did not include demographic variables or survey data, which are less distinctive to EH, although these are commonly collected and could be used in re-ID. We also searched for public information about each study's data sharing policy, which governs who can access the data outside the original research team. Because the first step in re-ID is gaining access to the data, data sharing practices are an important contributor to risk of re-ID. We convened an Advisory Council of EH researchers, privacy experts, ethicists, study-participant representatives, and public officials to consider the candidate studies and discuss data sharing and privacy issues in EH. The Advisory Council concurred with our choices of EH studies and data types. Advisory Council members are shown in Table S1. The selected studies are all U.S.-based, reflecting the expertise of the authors and Advisory Council. We later added the Green Housing Study because of its inclusion in the clustering analysis, and we revisited the websites for the studies to update information about available data.

Finally, we evaluated the availability of external public or commercial data that could be used in linkage strategies to match to the EH study data types. We searched for data sets online and drew on the experience of author L. Sweeney with obtaining commercial data sets and performing re-IDs (Sweeney 2002, 2013; Sweeney et al. 2017, 2018). We also searched for published examples of re-IDs where these data types were used as part of the linkage strategy.

Unsupervised Clustering of Environmental Chemical Measurements

The biological and individual-level environmental chemical measurements that are fundamental to EH studies are not usually

expected to contribute directly to re-ID via data linkage, because currently there are very few public repositories of matching data (lead results, which may become public in some jurisdictions, are a possible example of matching data). However, as noted earlier, chemical measurements can potentially contribute to re-ID by identifying subgroups that narrow the matching task for other variables. We tested the ability to infer subgroup membership in a data set by using a data clustering approach in two studies of household exposure to environmental chemicals, each conducted in multiple geographic locations. We selected these data sets because they contained relevant exposure measurements, and we had access to the "true" group membership (location in this case) to score the success of clustering. We hypothesized that some exposures vary systematically between locations, creating an opportunity to infer likely geographic location from the chemical measurements.

Description of data. Massachusetts and Northern California HES. The Massachusetts Household Exposure Study was conducted in 1999–2001 (Rudel et al. 2003). Dust and indoor air samples were collected from 120 households in Cape Cod, Massachusetts, but the current analysis is restricted to 72 participants who were identified as deceased as of May 2016, because of restrictions by the Massachusetts Cancer Registry on use of the data for this analysis. The Northern California Household Exposure Study was conducted in 2006 (Rudel et al. 2010). Dust and indoor and outdoor air samples were collected from 50 households in Richmond and Bolinas. Samples in both studies were analyzed for a broad suite of chemicals, including pesticides, phthalates, polycyclic aromatic hydrocarbons (PAHs), polybrominated diphenyl ethers (PBDEs), polychlorinated biphenyls (PCBs), alkylphenols, parabens, and other phenols and biphenyls identified as EDCs.

We created a combined data set restricted to analytes measured in the same medium (dust or indoor air) in both locations in a majority of homes. The data set included 24 chemicals measured in indoor air in 122 homes and 44 chemicals measured in dust in 120 homes.

Green Housing Study. The Green Housing Study (GHS)—a project of the Centers for Disease Control and Prevention and Department of Housing and Urban Development—is evaluating the effects of "green" housing on indoor environmental quality and children's respiratory health. The study sites are in Boston, Massachusetts, Cincinnati, Ohio, and New Orleans, Louisiana (Coombs et al. 2016; Dodson et al. 2017). Indoor air samples were collected from households with children ages 7–12 with physician-diagnosed asthma living in subsidized housing in Boston ($n = 44$) and Cincinnati ($n = 33$) in 2012–2013. Air samples were analyzed for 35 semivolatile organic compounds, including flame retardants, phthalates, environmental phenols, fragrance chemicals, and PCBs. New Orleans data were not available at the time of this analysis. Some homes ($n = 28$) had two air samples collected approximately 6 months apart; for these homes, we calculated the average exposure for each chemical. Homes that have two visits are not expected to differ systematically with respect to exposure from homes that have one visit.

Cluster analysis. Chemicals with detection frequencies above 10% were used for cluster analysis. Detection frequencies were calculated as the percent of samples with measured masses above the method reporting limit (MRL). Samples below the MRL were reported as either not detected or as estimated values. Estimated values were used in the cluster analysis. For samples reported as not detected, concentrations were substituted with the sample-specific reporting limit (SSRL), which was calculated as the MRL divided by the sample-specific volume of air or sample-specific mass of dust. Concentration data were natural log transformed. All

analyses were conducted using R (version 3.5.0; R Development Core Team).

We applied k -means clustering to the chemical exposure data. k -Means is a common method for unsupervised clustering that can be used to blindly classify groups of similar observations when no information about group membership is available. It uses an iterative process that partitions observations into k clusters based on distance to the cluster centroid. The number of clusters, k , is specified by the investigator *a priori*. The initial cluster centroids are assigned randomly to points in the n -dimensional space that spans all n chemicals in the analysis, and observations are assigned to the nearest centroid. Then, centroids are recalculated as the arithmetic means of the chemical measurements assigned to the cluster. Next, observations are reassigned to the new nearest centroid, and the process is repeated until no observations are reassigned to a different centroid. We used the Hartigan-Wong algorithm, which assigns observations to centroids by minimizing the within-cluster sum of squared errors (Hartigan and Wong 1979), as implemented in the “kmeans” function from the R base stats package (version 3.5.0; R Development Core Team), and we specified 100 random starts and a maximum of 10,000 iterations to converge on a stable solution.

We hypothesized that cluster analysis would partition the data by geographic region, so we specified two clusters ($k = 2$) for each analysis (i.e., Massachusetts and California for HES, Boston and Cincinnati for GHS). This correct answer for the number of geographic centers would be known to anyone from publications about the studies. We were not aware of subgroupings in these studies other than location (such as gender or race/ethnicity) that were likely to be relevant to household exposures and did not also covary with location, so we did not explore additional numbers of clusters.

To score the results of the k -means analysis, we checked the true location associated with each data point after clustering was complete. We assigned each cluster a geographic identity (e.g., Massachusetts or California) based on the site to which the majority of records belonged. We calculated the accuracy of the clustering by counting the correctly grouped records (e.g., Massachusetts records in the Massachusetts cluster and California records in the California cluster) and dividing by the total number of records analyzed. We also calculated the adjusted Rand index (Hubert and Arabie 1985), a measure of similarity between two partitions—in this case, the k -means clustering result and the true distribution. The adjusted Rand index has an expected value of zero for random clusters and a maximum value of 1 in the case of perfect agreement.

To examine which chemicals were influential in the clustering analysis, we ran principal component analysis (PCA) on the same data sets using eigendecomposition of the covariance matrix [function “princomp” from R base stats (version 3.5.0; R Development Core Team)]. We visually examined the clustering results on a plot of PC2 vs. PC1. We assessed whether variation along either axis contributed to separation between the clusters and examined the PC loadings to determine which chemicals contributed most strongly to the separation.

Differences in data-collection protocols or analytic methods between study sites may make it easier to partition participants by study site, thus increasing the risk of re-ID. Therefore, when study data are shared, researchers may want to consider data-masking methods to obscure site differences that represent methodological artifacts. To illustrate the potential influence of data-masking approaches on our ability to partition study participants by location using chemical exposure data, we created censored data sets that eliminated systematic site differences in reporting limits (in the HES) and sample volumes (in the GHS). In the

HES, MRLs systematically differed between Massachusetts and California for some chemicals. For each chemical, we calculated the most frequent MRL reported in each site (in cases of ties, we used the lower value) and defined the MRL for the censored analysis (MRL_{censored}) as the higher of the two modal MRLs. We calculated censored sample-specific reporting limits ($SSRL_{\text{censored}}$) as MRL_{censored} divided by the sample-specific volume of air or sample-specific mass of dust. For all records where the original SSRL or detected or estimated concentration was lower than $SSRL_{\text{censored}}$, the concentration was substituted with $SSRL_{\text{censored}}$. This substitution effectively masked differences in the exposure distribution resulting from differences in reporting limits. We repeated the cluster analysis using the same procedures described above but using the censored concentration. We did not mask differences by site in sample volume or sample mass, which were small relative to differences in MRL in the HES. In GHS, MRLs did not differ by site, but the sample-specific volume of air was systematically higher in Cincinnati (mean \pm standard deviation = $18.6 \pm 1.7 \text{ m}^3$) than Boston ($15.3 \pm 2.7 \text{ m}^3$) (Welch’s two sample t -test: $t = -7.9$, $df = 102.5$, $p < 0.001$). To assess whether differences in sample volume were driving the cluster results, we repeated the cluster analysis with nondetects substituted with the MRL divided by the median volume of air across Boston, Massachusetts, and Cincinnati, Ohio (i.e., a constant value).

Results

EH Data Types That Are Vulnerable to Linkage

The 12 studies chosen for investigation included the National Health and Nutrition Examination Survey (NHANES), a cross-sectional sample designed to be representative of the U.S. population (Zipf et al. 2013); cohort studies, including occupational groups [Agricultural Health Study (Alavanja et al. 1996) and California Teachers Study (Bernstein et al. 2002)], a disease-risk group (Sister Study; Sandler et al. 2017), and children [Breast Cancer and the Environment Research Program (BCERP) (Biro et al. 2010) and Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) Study (Eskenazi et al. 2003)]; environmental disaster response [Gulf Long Term Follow-Up (GuLF) Study (Kwok et al. 2017)]; and household chemical exposures [American Healthy Homes Survey (Stout et al. 2009), Relationships of Indoor Outdoor and Personal Air (RIOPA) (Weisel et al. 2005), Pesticide and Chemical Exposure (PACE) Study (Adamkiewicz et al. 2011), HES, and GHS]. We selected eight features of EH studies for investigation: a focus on specific locations, inclusion of multiple family members in the study, medical data, genetic data, occupation data, housing data, exposure data from biological samples, and exposure data from home or personal environment samples. We also investigated each study’s data-sharing policy. The studies and study features are identified in Table 1.

EH data types and potential for linkage. Each study included between three and eight of the EH features we investigated. Here we review publicly available data sets that could match to each of these features in linkage-based re-ID. The data sets that follow illustrate the most apparent vulnerabilities but are not an exhaustive list of currently available data relevant to re-ID.

Focus on specific locations. Nine studies are limited to specific geographic locations, so linkage efforts can focus on matching to data from that location. Locations may be statewide, such as the California Teachers Study and Agricultural Health Study (two states), or an environmentally defined region, such as the CHAMACOS Study in the intensive-agriculture region of the Salinas Valley and the GuLF Study of the area affected by the Deepwater Horizon oil spill. Other studies are limited to

Table 1. Study characteristics and data types that may contribute to re-identification risk in selected environmental health studies.

Study	Study characteristics				Data types				
	Focus on specific locations ^a	Family members in study ^b	Medical data	Genetic data	Occupation data	Housing data ^c	Exposure data from biological samples	Exposure data from home/personal environment samples	
<i>Agricultural Health Study</i> . Private pesticide applicators and their spouses in Iowa and North Carolina; licensed pesticide applicators in Iowa.	x	x	x	x	x	x	x	x	
<i>American Healthy Homes Survey</i> . Representative sample of U.S. homes 2005–2006.	—	—	—	—	x	x	—	x	
<i>Breast Cancer and the Environment Research Program (BCERP) Puberty Study</i> . Girls recruited at 6–8 years of age in New York City, California Bay Area, and Greater Cincinnati.	x	—	x	x	—	—	x	—	
<i>California Teachers Study</i> . Current and former female public school teachers or administrators.	x	—	x	x	x	—	x	—	
<i>CHAMACOS Study</i> . Mothers and children in Salinas Valley, CA.	x	x	x	x	x	x	x	x	
<i>Green Housing Study</i> . Children with physician-diagnosed asthma living in public housing in greater Boston, Cincinnati, and New Orleans.	x	—	x	—	—	x	x	x	
<i>Gulf Long Term Follow-up (GuLF) Study</i> . Participants in the Deepwater Horizon oil spill cleanup or training.	x	—	x	x	x	x	x	x	
<i>National Health and Nutrition Examination Survey (NHANES)</i> . Nationally representative sample of the U.S. population, collected from specific locations in each two-year cycle.	—	—	x	x	x	x	x	—	
<i>Pesticide and Chemical Exposure (PACE) Study</i> . Residents of urban MA and rural FL neighborhoods.	x	—	—	—	x	x	—	x	
<i>Relationships of Indoor, Outdoor, and Personal Air (RIOPA)</i> . Adults and children in Elizabeth, NH; Houston, TX; and Los Angeles, CA.	x	x	x	—	x	x	—	x	
<i>Silent Spring Institute Household Exposure Study (HES)</i> . Residents in Cape Cod, MA, and Bolinas, CA, and Richmond, CA.	x	—	—	—	x	x	x	x	
<i>Sister Study</i> . Women (cancer-free at enrollment) with a sister diagnosed with breast cancer.	—	—	x	x	x	x	x	x	

Note: —, not a study characteristic or a data type collected in the study.

^aOne enrollment criterion was living or working in a publicly defined geographic area. In addition, NHANES samples from 15 locations per year, although these locations are not intended to be a focus of study.

^bThe study enrolled family members as part of its study design. Additional studies, for example NHANES and the Sister Study, allow enrollment of multiple members of the same family.

^cCharacteristics of participants' homes, such as number or type of rooms; square footage; year built; information about heating, ventilation, and air conditioning; presence of certain furnishings or appliances, etc.

metropolitan areas (BCERP Puberty Study cohorts), cities (GHS, RIOPA, Northern California HES), or counties (PACE Study, Massachusetts HES). In the Northern California HES, specific neighborhoods were named in one city. NHANES is a nationally representative sample that does not overtly include location, but the locations and dates associated with data collection cycles can sometimes be discovered in local news stories and photographs online. However, because multiple locations are pooled in each cycle, identifying subsets of data associated with specific locations would require further work, such as data matching or clustering. Otherwise, linkage analysis must be performed between the entire NHANES data set and external data from all locations included in each cycle. The ability to identify location improves the ease and likelihood of matching demographic information, such as gender, race, and age, to voter lists or commercial lists of residents (e.g., Sweeney et al. 2017). Lists of residents by name, address, and demographic characteristics—along with countless other personal data elements—are readily available by geographic area from data brokers (Ramirez et al. 2014). Examples

of major data brokers include Acxiom, Experian, Equifax, CoreLogic, and TowerData. Since January 2019, the State of Vermont has required data brokers who trade data of Vermont residents to register in a public database; as of this writing, there are 154 active registrations (Vermont Secretary of State 2019). In addition, knowing location narrows the search and improves the likelihood of matching data in other domains, such as housing and occupation, described below.

Occupation data. Ten studies have information about occupation of the participant or the parent of the participant, providing data that potentially matches to licensing lists for pesticide applicators, teachers, nurses, and other professions, or to publicly accessible LinkedIn profiles, professional society membership lists, and institutional employer websites (e.g., Sweeney et al. 2018). In the same way that location can narrow sources of matching data, lists of licensed professionals can serve as the population registry or be cross-referenced with other population registries (e.g., voter lists) to restrict the pool of possible matches. Lists of licensed professionals may also contain linkable information,

such as the year in which a professional obtained a license or allowed it to expire.

Genetic data. Seven studies have genetic data that could potentially be matched to ancestry sites (e.g., [Erlich et al. 2018](#)); indeed, these sites are designed to facilitate record matching. Conversely, it is possible to begin from a publicly available genome on an ancestry site that also includes the individual's name or family surname and then infer whether the individual participated in a research study reporting some genomic information. The shared research data could be as limited as statistical measures of linkage disequilibrium from a genome-wide association study ([Wang et al. 2009](#)), or a genomic data-sharing beacon that returns a yes or no answer for the presence of a single nucleotide polymorphism (SNP) in the data set ([von Thenen et al. 2019](#)). Identifying someone as a participant can associate the individual with any sensitive characteristics of the study population, such as residence in a community with environmental contamination or diagnosis of disease.

Medical data. Nine studies collected medical data that may be linked to data disclosed in hospital discharge records, pharmacy sales, local news stories and obituaries, social media, and disease-centered online communities ([Culnane et al. 2017](#); [El Emam et al. 2011](#); [Sweeney 2013](#)). A 2013 survey found that 33 states release some form of publicly available—but not necessarily free—patient-level hospital discharge data ([Hooley and Sweeney 2013](#)), and we found current examples of these practices (e.g., [New York State Department of Health 2019](#); [Vermont Department of Health 2019](#); [Virginia Health Information 2018](#)). In addition to some demographic information, hospital discharge records may include information like admission and discharge dates, diagnoses, and cost of stays. Anonymized patient records are also being created by health information–data-mining companies that aggregate data purchased from pharmacies, insurance companies, and health care providers ([Tanner 2017](#)). Some of these data sources, such as news stories and obituaries, contain direct identifiers that could be used to re-ID participants; others, like hospital discharge records, may contain additional quasi-identifiers that could be used as part of a multistage re-ID strategy.

Housing data. Ten studies collected housing data to characterize potential exposures, such as construction years and materials, residence years, and floor plan characteristics. These data may be linked to local records such as tax assessor data, real estate transactions, and building permits, as well as records available on real estate websites, such as Zillow (e.g., [Sweeney et al. 2017](#)).

Biological or personal environment exposure data. All 12 EH studies contained at least one type of exposure data. Because few public repositories contain any matching exposure data, these data are less vulnerable to straightforward linkage approaches. However, biological and household samples tested for consumer product chemicals could potentially be linked to commercial data on credit card purchases. In addition, we evaluate in this article their potential use for identifying subgroup membership using cluster analysis.

Multiple family members. Three studies included multiple family members in the same study by design (e.g., spouses, parents and children). If one member is re-identified, then it becomes trivial to identify all family members in the study.

Data sharing practices of the studies. Among the selected studies, RIOPA and NHANES post selected data publicly online ([CDC/NCHS 2018](#); [Health Effects Institute n.d.](#)), and exposure data from the American Healthy Homes Survey are intended for inclusion in the public U.S. EPA ExpoCast™ database but are not yet available online ([NCCT 2017](#)). For NHANES, re-ID is

prohibited by law and online instructions state that use of the data signifies agreement to use it “only for the purpose of health statistical reporting and analysis” ([CDC 2015a](#)). Additional NHANES data beyond the public-use files are available by application at a restricted data center; researchers must receive approval in advance for their analysis plans and code, and they enter the controlled area without laptops, smartphones, or other electronic communications devices ([CDC 2015b](#)). Four studies—the Agricultural Health Study, California Teacher's Study, GuLF Study, and Sister Study—have established data access procedures that consider requests for specific variables for specific research purposes under agreements to protect the confidentiality of individual participants ([California Teachers Study n.d.](#); [Freeman et al. 2017](#); [NIH n.d.d](#); [Sister Study n.d.](#)). The Agricultural Health Study has shared selected data in response to FOIA requests (e.g., see Supplemental Material of [Goodman et al. 2017](#)), and other federally funded studies must also follow the requirements set forth in that Act and in the Shelby amendment. For the other studies we considered, data-sharing policies were not publicly specified; however, they are required to follow policies set by IRBS and funding agencies, such as NIH.

Unsupervised Clustering of Environmental Chemical Measurements

Massachusetts and Northern California HES. Air. A total of 13 chemicals were detected in residential indoor air samples in at least 10% of the 122 homes in the data set using the original reporting limits. *k*-Means clustering of these uncensored chemical measurements grouped 70 of 72 Massachusetts homes into one cluster with positive scores for the first principal component (PC1 in [Figure 1A](#)) and grouped all 50 California homes (and 2 misclassified Massachusetts homes) into a second cluster with negative scores for PC1, resulting in a partition accuracy of 98.4% ([Table 2](#)). When differences in MRLs between the Massachusetts and California sites were masked by substituting censored sample-specific reporting limits for some observations, all 13 chemicals were retained for analysis based on detection frequency. *k*-Means analysis of the data set with censored detection limits grouped 63 of 72 Massachusetts homes into one cluster with positive scores on PC1, and grouped all 50 California homes (and 9 misclassified Massachusetts homes) into a second cluster with negative scores for PC1, for a partition accuracy of 92.6% ([Table 2](#); [Figure 1B](#)). The chemicals with the highest loadings on PC1 in both analyses include banned organochlorine pesticides (e.g., heptachlor and chlordane) and the disinfectant *o*-phenylphenol ([Table S2](#)). Massachusetts homes had positive scores for PC1, reflecting higher levels of these chemicals—which is consistent with published findings from the HES study ([Rudel et al. 2010](#)).

Dust. A total of 25 chemicals were detected in residential indoor dust samples in at least 10% of the 120 homes in the data set using original reporting limits. *k*-Means clustering of these uncensored chemical measurements grouped 68 of 71 Massachusetts homes (and 1 misclassified California home) into one cluster with positive scores for PC1, and grouped 48 of 49 California homes (and 3 misclassified Massachusetts homes) into a second cluster with negative scores for PC1, resulting in a partition accuracy of 96.7% ([Table 2](#), [Figure 1C](#)). When differences in MRLs between the Massachusetts and California sites were masked by substituting censored sample-specific reporting limits for some observations, only 18 chemicals were detected in at least 10% of homes. *k*-Means analysis of the data set with censored detection limits grouped 24 of 71 Massachusetts homes (and 6 misclassified California homes) into one cluster with higher scores on PC1, and grouped 43 of 49 California homes (and 47 misclassified Massachusetts homes) into

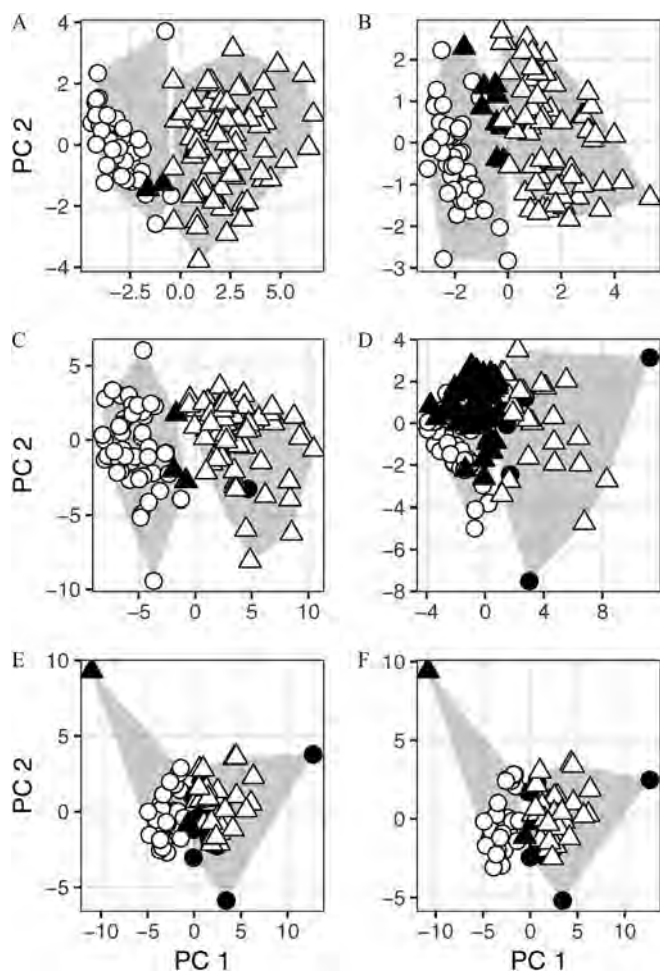


Figure 1. Individual homes plotted by principal component scores (PC1 and PC2) of residential chemical concentration data and overlaid on gray convex hulls indicating the bounds of two clusters generated using unsupervised *k*-means cluster analysis of the same data. All panels show homes from two regions. Homes were classified as correctly clustered (white symbols) if they were grouped in the cluster containing the majority of homes from their region; otherwise, they were classified as incorrectly clustered (black symbols). The shape of the symbol indicates the home's true location. (A) and (B): *k*-means classification of 122 homes in the Household Exposure Study (72 from Massachusetts, triangles; 50 from California, circles) based on chemical concentrations in indoor air using original (A) or censored (B) reporting limits. (C) and (D): *k*-means classification of 120 homes in the Household Exposure Study (71 from Massachusetts, triangles; 49 from California, circles) based on chemical concentrations in indoor dust using original (C) or censored (D) reporting limits. (E) and (F): *k*-means classification of 77 homes in the Green Housing Study (33 from Cincinnati, Ohio, triangles, and 44 from Boston, Massachusetts, circles) based on chemical concentrations in indoor air using original (E) or constant (F) reporting limits.

a second cluster with lower scores for PC1, for a partition accuracy of 55.8% (Table 2, Figure 1D). Four of the top six chemicals that contributed to separation along PC1 (diazinon, PCB 105, PCB 153, and PCB 52) were among those excluded from analysis when reporting limits were censored (Table S2), which likely contributed to the reduction in accuracy of the unsupervised clustering.

Green Housing Study. Air. A total of 28 chemicals were detected in residential indoor air samples in at least 10% of the 105 samples collected, before duplicate samples were averaged for 28 homes. *k*-Means clustering of the data set with original detection limits grouped 31 of 33 Cincinnati homes (and 13 misclassified Boston homes) into one cluster with positive scores for

PC1, and grouped 31 of 44 Boston homes (and 2 misclassified Cincinnati homes) into a second cluster with negative scores for PC1, resulting in a partition accuracy of 80.5% (Table 2; Figure 1E). When differences in sample volumes between Boston and Cincinnati were masked by using the median volume of air across sites to calculate reporting limits, rather than a sample-specific value, *k*-means analysis produced an identical partition to the analysis using original reporting limits (Table 2; Figure 1F). The chemicals with the highest loadings on PC1 in both analyses include fragrance chemicals (musk ketone, musk xylene, tonalide, galaxolide) and flame retardants (TDCIPP, BDE 47, BDE 100, BDE99, BDE 28) (Table S3). Cincinnati homes had positive scores for PC1, reflecting higher levels of these chemicals.

These results show that region of residence of a particular study participant can be inferred with substantial confidence from their cluster membership. Thus, the ability to partition data sets using differences in chemical levels can narrow the pool of potential matches for re-ID for each participant.

Discussion

This investigation considered two types of vulnerabilities that increase risk of re-ID in EH data: *a*) study variables that overlap with public or commercial data sets and *b*) environmental measurements that vary by subgroup, such as location. Our results show that EH studies collect several types of data that could facilitate re-ID, suggesting that public sharing of study data will often create privacy threats to study participants. Researchers must consider vulnerabilities in their data when deciding which data to share, with whom, and with which restrictions. Additional legal protections are also needed, parallel to protections for genetic data.

Linking EH Data in Practice

We found that information collected in EH studies (Table 1) extensively overlaps with publicly available data sets that could result in re-ID using linkage strategies. For example, occupational data may be linked to lists of licensed professionals, housing characteristics may be linked to tax and real estate data, genetics data are stored on ancestry websites, and medical information may be linked to hospital release data or news stories about accidents, illnesses, or deaths. Studies limited to well-defined geographic areas constrain the population of prospective matches, especially in combination with other data fields (e.g., registered pesticide applicators in Iowa). The data types we identified in environmental studies also are commonly collected in other types of cohort studies, posing similar risks.

As noted earlier, we do not explicitly discuss demographic information (e.g., age, gender, race/ethnicity); however, these data are nearly always collected in EH studies and are known to be vulnerable in linkage-based re-ID strategies. We also did not consider mobility traces (i.e., detailed location information) in our analysis. At the time the 12 EH studies were identified, personal sensors and mobile health technologies were not as widely and cheaply available. Recent EH studies, however, are precisely tracking individual location, for example, to infer collocated environmental data such as weather, traffic, and air quality (Habre et al. 2018). Mobility traces are extremely vulnerable to re-ID (de Montjoye et al. 2013; Douriez et al. 2016; Siddle 2014). Finally, we did not consider linkage to data sets that are not intended to be public but enter the public domain (such as through unintentional or intentional leaks); data sets obtained through malicious or illegal activities (such as hacking); or privately held data sets accessible only to the holder (such as data held by a health insurer or bank) (Culnane et al. 2017). Linkage with privately held data

Table 2. Accuracy of *k*-means cluster analysis for subgrouping homes by region in the household exposure study (HES; Massachusetts and California) and Green Housing Study (GHS; Boston, Massachusetts, and Cincinnati, Ohio) using concentrations of chemicals detected in at least 10 percent of residential indoor air or dust samples.

Study	Homes (<i>n</i>)	Sample matrix	Chemicals in study ^a (<i>n</i>)	Chemicals in cluster analysis ^b (<i>n</i>)	Reporting limits	Accuracy ^c (%)	Adjusted Rand index ^d
HES	122	Air	24	13	Original ^e	98.4	0.93
HES	122	Air	24	13	Censored ^f	92.6	0.72
HES	120	Dust	44	25	Original	96.7	0.87
HES	120	Dust	44	18	Censored	55.8	0
GHS	77 ^g	Air	35	28	Original	80.5	0.36
GHS	77 ^g	Air	35	28	Constant ^h	80.5	0.36

^aNumber of chemicals measured in the same medium in all homes in each cluster analysis.

^bNumber of chemicals detected in at least 10% of homes given the reporting limits used in each analysis.

^cNumber of homes correctly grouped by region using *k*-means clustering divided by the total number of homes analyzed.

^dThe adjusted Rand index measures similarity between the two clusters identified by *k*-means analysis and the two true regional subgroups in the data. It has an expected value of zero for random clusters and a maximum value of 1 in the case of perfect agreement.

^eIn analyses using the original reporting limits, concentrations that were not detected were substituted with the sample-specific reporting limit (SSRL). We calculated the SSRL as the method reporting limit (MRL) divided by the sample-specific volume of air or sample-specific mass of dust.

^fIn analyses with censored reporting limits, we calculated the most frequent MRL reported in each site (in cases of ties we used the lower value). We defined MRL_{censored} as the higher of the two modal MRLs and calculated censored sample-specific reporting limits (SSRL_{censored}) as MRL_{censored} divided by the sample-specific volume of air or sample-specific mass of dust. For all records where the original SSRL or detected or estimated concentration was lower than SSRL_{censored}, the concentration was substituted with SSRL_{censored}.

^gCluster analysis was performed on 77 homes comprising 105 samples. A total of 49 homes were sampled once, and 28 homes were sampled twice approximately six months apart. For homes sampled twice, we used the average exposure for each chemical.

^hNon-detects were substituted with the MRL divided by the median volume of air across Boston, Massachusetts, and Cincinnati, Ohio.

could be used by the holder to inform decisions without necessitating disclosure.

The fact that environmental data sets include vulnerable fields does not necessarily mean that re-ID will occur, however. One factor in real-world vulnerability is ease of access to data with the potential for linkage. Data available online have the greatest ease of access, and many public government records have migrated online in recent years. However, for older studies, or in less well-resourced municipalities, records of interest for linkage may not be available electronically, and attackers may be less likely to seek out original print records housed in government offices. Another consideration is for what years the data of interest are available. Some data may exist only historically, whereas other data may have been available at the time the study was conducted but were not archived. For example, in our previous re-ID experiment using HES data, historical tax data were available for purchase in Bolinas, California, but these data did not include detailed housing characteristics. Rather, current housing characteristics were mined from the Marin County (California) Tax Assessor's office website up to 10 y after the study data were collected. This factor likely contributed to not obtaining any correct re-IDs in Bolinas (Sweeney et al. 2017). Because the availability of property records on Cape Cod created a similar situation, we predicted that a re-ID with the Massachusetts HES data would not provide meaningful results. Other types of data may have started to be collected only in recent years. For instance, as services have migrated to the internet, data sets containing personal data have proliferated. In addition, even when data are available, completeness and accuracy of the data will affect their ultimate utility. However, even incomplete data can create risk, and completeness and quality of data are likely to increase over time. Cost of the data is a final potential barrier. Although cost—especially of data obtained through commercial brokers—could deter an individual attacker, better-resourced entities are unlikely to be deterred, and costs may be offset by financial or other gains, such as identifying individuals who would be costly to insure or learning information relevant to a legal case. All three of these factors (cost, availability, and ease of access) can change rapidly as old files are digitized, electronic data storage becomes cheaper, and regulations about personal data evolve. Therefore, linkage strategies that are not currently possible may become possible in the future, for example, as new data become available, and current strategies could expire as data are lost to the public domain. A final consideration is the availability of EH research data. As

investigators face mounting pressure from funding agencies and regulators to release data (EPA 2018; Holdren 2013), EH data could become increasingly accessible.

Inferring Group Membership Using Environmental Measurements

Even when EH data sets do not contain certain fields explicitly, they can often be inferred and therefore help triangulate re-ID approaches. In NHANES, geographic location is not specified in the data set, but the counties sampled in 2-y time periods can be identified using multiple strategies—for example, by locating news articles that announce NHANES's presence in a community (e.g., Bawab 2018; Kowalick 2018) or by examining search engine data for locations where searches for “NHANES” or related terms have experienced spikes. We illustrated in the experiment described here that another approach could use environmental chemical measurements to facilitate re-ID by uncovering latent variables (perhaps intentionally removed to protect privacy). We found that unsupervised clustering of chemical exposure measurements alone successfully discriminated geographic location in five out of six test cases (with accuracy rates ranging from 80% to 98%). Although this study was limited to inferring geographic region, the same approach could be used with other variables. For example, behavioral correlates of gender and race/ethnicity (e.g., frequency of use of makeup or fragrance) can influence exposure levels, as can biological sex (e.g., sex differences in elimination rates contributes to sex differences in levels of per- and polyfluoroalkyl substances).

In this experiment we were able to use the true identity of the homes to assign cluster identity, but an attacker would not have access to the true identities. However, an attacker could use the cluster means, or loadings from a principal component analysis, to infer the identity of the cluster. For example, after clustering the HES, we observed one cluster with higher organochlorine pesticide levels than the other. An attacker could easily look to the published findings from the study to find this result. In this case, Rudel et al. (2010) note that “Indoor air concentrations in Cape Cod were higher than those in this California study for banned organochlorine pesticides (but not contemporary pesticides) and PCBs, and for the commercial chemicals nonylphenol and o-phenylphenol.” In addition, both papers describing the HES include a supplement with summary statistics of all chemicals measured in the study, so an attacker can learn that the 95th

percentile of heptachlor in indoor air in the Northern California HES was 0:37 ng=m³ (Rudel et al. 2010), whereas the 90th percentile in the Cape Cod HES was 19 ng=m³ (Rudel et al. 2003). If these published data were not available, an attacker might logically infer that Cape Cod's older housing and humid climate would be associated with higher residues of chlordane used for termites and that additional pesticides would be found from the historical proximity to agriculture, protecting trees from gypsy moths, and manicured greenery for the tourist economy. In contrast, a primary difference observed in the GHS data—higher fragrance levels in Cincinnati, Ohio, than in Boston, Massachusetts—would be difficult to infer without the published study findings. It is possible that an attacker would be more likely to attribute the differences to another subgroup categorization, such as ethnicity, rather than geography. The accuracy of the partition is also lower, likely because fragrance levels reflect current rather than historical use, and fragrance use has a strong personal component in addition to reflecting larger cultural trends.

A limitation of this experiment is that we used a relatively simple clustering algorithm, *k*-means, that is easy to implement.

Other more sophisticated unsupervised methods, such as using the Expectation-Maximization algorithm to estimate Gaussian mixture models, could produce even more accurate clusters than those observed here or be used to assess the optimal number of clusters without having an *a priori* prediction. The failed test case (HES dust data with censored reporting limits) demonstrates that artifacts in the data set—in this case, laboratory reporting limits—can contribute to identifiability, but also that researchers can use data-masking approaches to reduce identifiability while retaining sufficient data for some analyses. In the HES, the differing reporting limits between sites resulted from advances in laboratory analytical capabilities in the years between when the data were collected. In other data sets, differences could result from samples being analyzed by different laboratories—for example, as might occur in a large consortium. By censoring the data to the same reporting limits at both sites, we masked this simple discriminator. However, we also lost some information about low levels of these chemicals at one site, which could slightly reduce the utility to other researchers using the shared data set. A final limitation is that this technique is likely only sensitive enough to reveal high-level group membership (e.g., sex, location, disease status), variables that may most often be critical to the utility of the data set and therefore unlikely to be masked in the first place.

To establish the generalizability—and limits—of using *k*-means or other clustering approaches for inferring subgroups from environmental data, future research should test techniques on data sets for which the true group memberships are not initially known to the study team, and on larger data sets that have more participants, contributing sites, or both. Results will help to evaluate the likely magnitude of risk associated with chemical measurement data alone in different types of studies. The *k*-means technique relies on every site having a unique multidimensional “exposure signature” with minimal overlap with other sites. Therefore, the combination of intrasite and intersite variation will determine cluster accuracy (i.e., subgroup identifiability). We expect that larger numbers of participants per site would decrease identifiability only if the additional participants increase intrasite variation. Similarly, we expect that additional sites would decrease identifiability only if the new sites decrease intersite variation. Greater numbers of measured analytes may increase identifiability insofar as the analyte set is more likely to contain one or more distinguishing chemicals that increase intersite variation. In this study, for example, the HES Dust censored analysis had reduced identifiability because seven chemicals were excluded that no longer met minimum

detection requirements. In addition, clustering techniques should be evaluated for efficacy at uncovering latent variables other than geographic location in isolation (i.e., distinguishing sex in a single geographic location) and in combination with location. Empirical research in a variety of studies will build knowledge about factors associated with higher or lower subgroup identifiability. This knowledge will inform decisions about whether to share data, and it can also spur subsequent research to evaluate and optimize data-masking strategies that reduce identifiability while retaining the utility of the data for other analyses. Results will help researchers understand risks and consider solutions when sharing exposure measurements.

Technical and Policy Solutions

To evaluate data sharing plans, researchers and IRBs would benefit from empirically based methods to quantitatively evaluate re-ID potential and inform solutions. For example, the Privacert computational model (developed by Privacert, Inc., owned by coauthor L. Sweeney) predicts re-ID risk in relation to HIPAA standards, but like HIPAA itself, the model does not provide a quantitative estimate of risk (Privacert n.d.; Sweeney 2011). New models could be developed for other types of data and measures of privacy. Other benchmarks for privacy are *k*-anonymity (Sweeney 2002) and unicity (de Montjoye et al. 2013, 2015). *K*-anonymity requires that every combination of fields that could be used for linkage occurs at least *k* times, such that linking on these fields never reveals a set of individuals smaller than *k*. However, it is not clear what value of *k* is acceptable, because small groups can be equally harmed by re-ID. In a similar vein, unicity measures the proportion of a data set that can be uniquely re-identified given some number of outside pieces of information. Higher values of *k* and lower values of unicity (i.e., less uniquely identifiable data) can be achieved by redacting fields or coarsening the data into larger categories. For example, we used this method in the previous HES re-ID study by aggregating year a house was built into decade or longer groups, rather than exact year (Sweeney et al. 2017). However, decisions of how to redact and coarsen data often are not guided by a formal definition of privacy, in part because producing an optimal *k*-anonymous data set (one that minimizes data loss) is unlikely to be achievable with the technologies and computational methods available in the present and near future (Meyerson and Williams 2004). In addition, coarsening or suppressing data may do little to lessen identifiability (de Montjoye et al. 2013, 2015) and can completely negate the utility of the data (Aggarwal 2005; Brickell and Shmatikov 2008). More quantitative research and translational guidance is needed before researchers rely on redaction or coarsening to protect their data sets. However, as the data landscape constantly expands, new “big data” analytical techniques are developed, and computing power increases, researchers who wish to protect their data using technical solutions face an unending arms race against those parties seeking to re-identify data.

In comparison with these technical approaches, the more protective strategy is to restrict access to potentially vulnerable data sets and enact public policies or laws that protect study participants from harm due to privacy loss. Restricted access refers to any access limitation beyond open, publicly available data. Some forms of restricted access include sharing identifiable data under IRB oversight, typically with other researchers; restricted-access data centers (e.g., similar to that used by NHANES to give researchers limited access to more sensitive data fields); and legally binding data-use agreements that establish allowed uses of the data and penalties for misuse. Although NHANES data are subject to the general data-use agreement for public-use files

from the National Center for Health Statistics (NCHS), which has the force of law and expressly prohibits re-ID, the agreement does not state penalties for violations, nor is it readily accessible from the NHANES website at the time of this writing (CDC 2015a). Rather, the agreement is housed in the data access section of the NCHS website. A more proactive strategy would be to require users to actively accept the data-use agreement (perhaps revised to include penalties) at the time when NHANES or other NCHS data are downloaded. Data access can also be mediated by a fixed query interface (i.e., raw data are not available). Simple fixed query systems are vulnerable to attack when results of repeated queries are combined, but a query system that uses differentially private algorithms would be less vulnerable (NASSEM 2017). Differentially private algorithms, which rely on noise, provide a formal guarantee that no individual private data can be inferred from the output of a statistical query. Differentially private algorithms are promising but not ready for practical implementation, because more work is needed to understand the relationships between amount of noise introduced, accuracy of the algorithm, complexity of the statistical task, and size of the data set. The All of Us Study is confronting these issues with a two-tiered system of data access to an online research hub. The public tier contains only anonymized, aggregate data. To access the registered tier, which includes more robust data, researchers will have to register, complete research ethics training, and sign a data-use agreement. All activity on the research hub will be tracked (NIH n.d.-a, n.d.-b).

Federal policies that govern medical records research, however imperfect, can stimulate the development of parallel policies for EH. The HIPAA Privacy Rule sets limits on the use and disclosure of personal medical information (DHHS 2002), and the Genetic Information Nondiscrimination Act of 2008 protects people from insurance and employment discrimination based on genetic data (United States Congress 2008). Although a prescriptive approach to privacy protection like HIPAA Safe Harbor would be ineffective for EH data (as evidenced by Sweeney et al. 2017), legal recognition of the sensitivity of EH information is needed.

Privacy-protective policies for data sharing are necessary to fulfill researchers' preexisting pledges of confidentiality to study participants in the informed consent. Researchers who violate these pledges, or who do not offer privacy protections in future studies, will risk suppressing research participation among people who fear loss of privacy. Some people are comfortable with open data sharing (Zarate et al. 2016), and others may be willing to accept low to moderate privacy risks for the benefit of public health. However, requiring consent for permissive data sharing could negatively affect participation of racial and ethnic minorities, populations that are already underrepresented in health research (Konkel 2015), and overburdened by diseases with environmental triggers, such as asthma (Forno and Celedón 2012). In multiple studies, African Americans have shown significantly less acceptance of broad consent than white participants (Ewing et al. 2015; Platt et al. 2014; Sanderson et al. 2017). In addition, environmental justice communities are also less able to cope with the economic harms of privacy loss. The new and growing "data justice" movement has pointed to many abuses of data for surveillance and discrimination and has called for combatting such abuses and using data for rights, justice, and fairness (Taylor 2017). Our research reflects such concerns.

Our survey of prominent EH studies and case study of clustering of environmental measurements illuminate features of EH studies that increase vulnerability to re-ID. Researchers and institutions face large knowledge gaps about the nature and magnitude of these risks. In addition, they lack technical and policy guidance, and legal protections have not been developed for potential harms

from release of EH data. Our work represents a beginning effort to stimulate further consideration of a fast-moving landscape of increasing vulnerability to re-ID as more data becomes accessible online. Empirical assessments of privacy risks in EH data can contribute to decisions about when and how to share data, accurate descriptions of risk in informed consent documents, and discussions about whether new legal protections are needed to shield study participants from harm. During this time when privacy risks and solutions remain substantially underinvestigated, researchers and agencies should be cautious about sharing EH study data outside IRB protections or other explicit privacy agreements.

Acknowledgments

The authors thank O. Zarate for contributions to the analysis of vulnerable data in EH studies, including the first draft of Table 1, and R. Dodson for access to selected environmental chemicals data from the GHS. This work was supported by the National Institute of Environmental Health Sciences (R01ES021726).

References

- AAAS (American Association for the Advancement of Science). N.d. Science Journals: editorial policies. <https://www.sciencemag.org/authors/science-journals-editorial-policies> [accessed 3 June 2019].
- Adamkiewicz G, Dodson R, Zota A, Perovich L, Brody J, Rudel R, et al. 2011. Semi-volatile organic compounds distributions in residential dust samples from 5 US communities: key lessons for improving residential exposure assessment. *Epidemiology* 22(1):S160–S161, <https://doi.org/10.1097/01.ede.0000392166.33641.22>.
- Aggarwal CC. 2005. On k-anonymity and the curse of dimensionality. In: *Proceedings of the 31st International Conference on Very Large Data Bases*. 30 August–2 September 2005, Trondheim, Norway. Trondheim: VLDB Endowment, 901–909.
- Alavanja MC, Sandler DP, McMaster SB, Zahm SH, McDonnell CJ, Lynch CF, et al. 1996. The Agricultural Health Study. *Environ Health Perspect* 104(4):362–369, PMID: 8732939, <https://doi.org/10.1289/ehp.96104362>.
- Baba A, Cook DM, McGarity TO, Bero LA. 2005. Legislating "Sound Science": the role of the tobacco industry. *Am J Public Health* 95(5):S20–S27, PMID: 16030333, <https://doi.org/10.2105/AJPH.2004.050963>.
- Bawab N. 2018. CDC is conducting a national health survey in Dallas. *Dallas Observer*, 2 November 2018. <https://www.dallasobserver.com/news/the-cdc-is-conducting-a-national-health-survey-in-dallas-11320061> [accessed 21 November 2018].
- Bernstein L, Allen M, Anton-Culver H, Deapen D, Horn-Ross PL, Peel D, et al. 2002. High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). *Cancer Causes Control* 13(7):625–635, PMID: 12296510, <https://doi.org/10.1023/a:1019552126105>.
- Biro FM, Galvez MP, Greenspan LC, Succop PA, Vangeepuram N, Pinney SM, et al. 2010. Pubertal assessment method and baseline characteristics in a mixed longitudinal study of girls. *Pediatrics* 126(3):e583, PMID: 20696727, <https://doi.org/10.1542/peds.2009-3079>.
- Brickell J, Shmatikov V. 2008. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2008, Las Vegas, NV. New York, NY: ACM, 70–78.
- Brody JG, Morello-Frosch R, Zota A, Brown P, Pérez C, Rudel RA. 2009. Linking exposure assessment science with policy objectives for environmental justice and breast cancer advocacy: the Northern California Household Exposure Study. *Am J Public Health* 99 (suppl 3):S600–609, PMID: 19890164, <https://doi.org/10.2105/AJPH.2008.149088>.
- Buchmann E, Böhm K, Burghardt T, Kessler S. 2013. Re-identification of smart meter data. *Pers Ubiquit Comput* 17(4):653–662, <https://doi.org/10.1007/s00779-012-0513-6>.
- California Teachers Study. n.d. California Teachers Study: for researchers. <https://www.calteachersstudy.org/for-researchers> [accessed 18 November 2019].
- CDC/NCHS (Centers for Disease Control and Prevention/National Center for Health Statistics). 2015a. Data User Agreement. https://www.cdc.gov/nchs/data_access/restrictions.htm [accessed 2 July 2019].
- CDC/NCHS. 2015b. On Site at an NCHS RDC. <https://www.cdc.gov/rdc/b2accessmod/acs210.htm> [accessed 30 July 2018].
- CDC/NCHS. 2018. Public-Use Data Files and Documentation. https://www.cdc.gov/nchs/data_access/ftp_data.htm [accessed 18 November 2019].
- Cohen-Hubal EA. 2009. ExpoCast: exposure science for prioritization and toxicity testing. In: *International Society of Exposure Science, Annual Meeting*.

- Computational Toxicology Board of Scientific Counselors Review, Research Triangle Park, NC, 29–30 September 2009. Minneapolis, MN.
- Coombs KC, Chew GL, Schaffer C, Ryan PH, Brokamp C, Grinshpun SA, et al. 2016. Indoor air quality in green-renovated vs. non-green low-income homes of children living in a temperate region of US (Ohio). *Sci Total Environ* 554-555:178–185, PMID: 26950631, <https://doi.org/10.1016/j.scitotenv.2016.02.136>.
- Culnane C, Rubinstein BI, Teague V. 2017. Health data in an open world. *arXiv*: 1712.05627.
- de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD. 2013. Unique in the crowd: the privacy bounds of human mobility. *Sci Rep* 3:1376, PMID: 23524645, <https://doi.org/10.1038/srep01376>.
- de Montjoye YA, Radaelli L, Singh VK, Pentland AS. 2015. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science* 347(6221):536–539, PMID: 25635097, <https://doi.org/10.1126/science.1256297>.
- DHHS (Department of Health and Human Services). 2002. Standards for Privacy of Individually Identifiable Health Information; Final Rule. *Fed Reg* 67(157):53181–53273, PMID: 12180470.
- Dodson RE, Udesky JO, Colton MD, McCauley M, Camann DE, Yau AY, et al. 2017. Chemical exposures in recently renovated low-income housing: Influence of building materials and occupant activities. *Environ Int* 109:114–127, PMID: 28916131, <https://doi.org/10.1016/j.envint.2017.07.007>.
- Douriez M, Doraiswamy H, Freire J, Silva CT. 2016. Anonymizing NYC taxi data: does it matter? In: *Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. October 2016. Montreal, Canada: IEEE, 140–148.
- El Emam K, Jonker E, Arbuckle L, Malin B. 2011. A systematic review of re-identification attacks on health data. *PLoS One* 6(12):e28071, PMID: 22164229, <https://doi.org/10.1371/journal.pone.0028071>.
- EPA (U.S. Environmental Protection Agency). 2018. Strengthening transparency in regulatory science. *Fed Reg* 83:18768–18774.
- Erlich Y, Shor T, Pe'er I, Carmi S. 2018. Identity inference of genomic data using long-range familial searches. *Science* 362(6415):690–694, PMID: 30309907, <https://doi.org/10.1126/science.aau4832>.
- Eskenazi B, Bradman A, Gladstone EA, Jaramillo S, Birch K, Holland N. 2003. CHAMACOS, a longitudinal birth cohort study: lessons from the fields. *J Children's Health* 1(1):3–27, <https://doi.org/10.3109/713610244>.
- European Commission. 2017. *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Version 3.2*. Brussels, Belgium: European Commission Directorate-General for Research & Innovation.
- European Commission. 2018. Commission Recommendation of 25.4.2018 on access to and preservation of scientific information. 2375 final. Brussels, Belgium: European Commission.
- European Commission. n.d. European Open Science Cloud (EOSC). <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud> [accessed 2 July 2019].
- Ewing AT, Erby LA, Bollinger J, Tetteyio E, Ricks-Santi LJ, Kaufman D. 2015. Demographic differences in willingness to provide broad and narrow consent for biobank research. *Biopreserv Biobank* 13(2):98–106, PMID: 25825819, <https://doi.org/10.1089/bio.2014.0032>.
- Fischer EA. 2013. *Public Access to Data from Federally Funded Research: Provisions in OMB Circular A-110*. Washington, DC: Congressional Research Service, Library of Congress.
- Forno E, Celedón JC. 2012. Health disparities in asthma. *Am J Respir Crit Care Med* 185(10):1033–1035, PMID: 22589306, <https://doi.org/10.1164/rccm.201202-0350ED>.
- Freeman LB, Blair A, Hofmann J, Sandler DP, Parks CG, Thomas K. 2017. Agricultural Health Study Policy 2-4: Guidelines for Collaboration. https://aghealth.nih.gov/collaboration/AHS%20Policy%202-4%20Guidelines%20for%20Collaboration_2017.1.pdf [accessed 18 November 2019].
- Goho SA. 2016. The legal implications of report back in household exposure studies. *Environ Health Perspect* 124(11):1662–1670, PMID: 27153111, <https://doi.org/10.1289/EHP187>.
- Goodman JE, Loftus CT, Zu K. 2017. 2,4-Dichlorophenoxyacetic acid and non-Hodgkin's lymphoma: results from the Agricultural Health Study and an updated meta-analysis. *Ann Epidemiol* 27(4):290–292, PMID: 28292638, <https://doi.org/10.1016/j.annepidem.2017.01.008>.
- Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. *Science* 339(6117):321–324, PMID: 23329047, <https://doi.org/10.1126/science.1229566>.
- Habre R, Rocchio R, Hosseini A, van Vliet E, Eckel S, Valencia L. 2018. An mHealth platform for predicting risk of pediatric asthma exacerbation using personal sensor monitoring systems: The Los Angeles Prisms Center. In: *Proceedings of the ISEE Conference Abstracts 2018*.
- Hartigan JA, Wong MA. 1979. Algorithm AS 136: a k-means clustering algorithm. *J Royal Stat Soc Series C (Appl Stat)* 28(1):100–108, <https://doi.org/10.2307/2346830>.
- Health Effects Institute. n.d. Databases. <https://www.healtheffects.org/research/databases> [accessed 18 November 2019].
- Henriksen-Bulmer J, Jeary S. 2016. Re-identification attacks—A systematic literature review. *Int J Inform Manage* 36(6, part B):1184–1192, <https://doi.org/10.1016/j.ijinfomgt.2016.08.002>.
- Holdren JP. 2013. *Increasing Access to the Results of Federally Funded Scientific Research*. Washington, D.C: Office of Science and Technology Policy. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [accessed 13 May 2019].
- Hooley S, Sweeney L. 2013. Survey of publicly available state health databases. *arXiv* 1306:2564.
- Hubert L, Arabie P. 1985. Comparing partitions. *J Classification* 2(1):193–218, <https://doi.org/10.1007/BF01908075>.
- Justin J. 2018. To find alleged Golden State Killer, investigators first found his great-great-great-grandparents. *Washington Post*. Washington, D.C. 30 April 2018. https://www.washingtonpost.com/local/public-safety/to-find-alleged-golden-state-killer-investigators-first-found-his-great-great-great-grandparents/2018/04/30/3c865fe7-dfcc-4a0e-b6b2-0bec548d501f_story.html [accessed 31 October 2018].
- Konkel L. 2015. Racial and ethnic disparities in research studies: the challenge of creating more diverse cohorts. *Environ Health Perspect* 123(12):A297–302, PMID: 26625444, <https://doi.org/10.1289/ehp.123-A297>.
- Kowalick C. 2018. Wichita County selected to participate in national health survey. *Times Record News Wichita Falls, KS*. 8 September 2018. <https://www.timesrecordnews.com/story/news/local/2018/09/08/wichita-county-selected-participate-national-health-survey/1212460002/> [accessed 11 September 2018].
- Kwok RK, Engel LS, Miller AK, Blair A, Curry MD, Jackson WB, et al. 2017. The Gulf STUDY: a prospective study of persons involved in the *Deepwater Horizon* oil spill response and clean-up. *Environ Health Perspect* 125(4):570–578, PMID: 28362265, <https://doi.org/10.1289/EHP715>.
- Meyerson A, Williams R. 2004. On the complexity of optimal K-anonymity. In: *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. June 2004, Paris, France. 23:223–228, <https://doi.org/10.1145/1055558.1055591>.
- Narayanan A, Shmatikov V. 2008. Robust de-anonymization of large sparse datasets. In: *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 2008, Oakland, CA. 111–125, <https://doi.org/10.1109/SP.2008.33>.
- NASEM (National Academies of Sciences Engineering and Medicine). 2017. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press.
- NCCT (United States Environmental Protection Agency National Center for Computational Toxicology). 2017. Exposure Forecaster. <https://catalog.data.gov/dataset/exposure-forecaster-a16a1> [accessed 10 July 2019].
- New York State Department of Health. 2019. Hospital Inpatient Discharges (SPARCS De-Identified): 2017. <https://health.data.ny.gov/dataset/Hospital-Inpatient-Discharges-SPARCS-De-Identified/22g3-7e7> [accessed 21 November 2019].
- NIH (National Institutes of Health). 2017. About ECHO. <https://www.nih.gov/echo/about-echo> [accessed 26 July 2018].
- NIH. n.d.-a. Privacy Safeguards. <https://www.joinallofus.org/en/privacy-safeguards> [accessed 21 November 2018].
- NIH. n.d.-b. Workbench. <https://www.researchallofus.org/data/workbench/> [accessed 5 June 2019].
- NIH. n.d.-c. Program Overview—All of Us. <https://allofus.nih.gov/about/about-all-us-research-program> [accessed 1 November].
- NIH. n.d.-d. Gulf STUDY: For Researchers. <https://gulfstudy.nih.gov/en/forresearchers.html> [accessed 18 November 2019].
- OMB (Office of Management and Budget). 1999. OMB Circular A-110, “Uniform Administrative Requirements for Grants and Agreements with Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations.” *Fed Reg* 64(195):54926–54930.
- Pennarola F, Pistilli L, Chau M. 2017. Angels and demons: is more knowledge better than less privacy? an empirical study on a k-anonymized openly available dataset. *38th International Conference on Information Systems ICIS 2017 Proceedings. December 2017*. Seoul, South Korea. 1318.
- Platt J, Bollinger J, Dvoskin R, Kardia SL, Kaufman D. 2014. Public preferences regarding informed consent models for participation in population-based genomic research. *Genet Med* 16(1):11–18, PMID: 23660530, <https://doi.org/10.1038/gim.2013.59>.
- Privacert. n.d. HIPAA Solutions. <http://privacert.com/hipaa/index.html> [accessed 1 November 2018].
- Ramirez E, Brill J, Ohlhausen MK, Wright JD, McSweeney T. 2014. Data brokers: a call for transparency and accountability. Washington, DC: Federal Trade Commission. <https://www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014> [accessed 18 November 2019].

- Rocher L, Hendrickx JM, de Montjoye YA. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 10(1):3069, PMID: 31337762, <https://doi.org/10.1038/s41467-019-10933-3>.
- Rudel RA, Camann DE, Spengler JD, Korn LR, Brody JG. 2003. Phthalates, alkylphenols, pesticides, polybrominated diphenyl ethers, and other endocrine-disrupting compounds in indoor air and dust. *Environ Sci Technol* 37(20):4543–4553, PMID: 14594359, <https://doi.org/10.1021/es0264596>.
- Rudel RA, Dodson RE, Perovich LJ, Morello-Frosch R, Camann DE, Zuniga MM, et al. 2010. Semivolatile endocrine-disrupting compounds in paired indoor and outdoor air in two northern California communities. *Environ Sci Technol* 44(17):6583–6590, PMID: 20681565, <https://doi.org/10.1021/es100159c>.
- Sanderson SC, Brothers KB, Mercaldo ND, Clayton EW, Antommaria AHM, Aufox SA, et al. 2017. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the US. *Am J Hum Genet* 100(3):414–427, PMID: 28190457, <https://doi.org/10.1016/j.ajhg.2017.01.021>.
- Sandler DP, Hodgson ME, Deming-Halverson SL, Juras PS, D'Aloisio AA, Suarez LM, et al. 2017. The Sister Study Cohort: baseline methods and participant characteristics. *Environ Health Perspect* 125(12):127003, PMID: 29373861, <https://doi.org/10.1289/EHP1923>.
- Schwartz J. 2018. Transparency” as mask? The EPA’s proposed rule on scientific data. *N Engl J Med* 379(16):1496–1497, PMID: 30156992, <https://doi.org/10.1056/NEJMp1807751>.
- Science and Technology at the Environmental Protection Agency. 2019. Hearing Before the Committee on Science, Space, and Technology, U.S. House of Representatives, 116th Congress (September 19, 2019) (testimony of Andrew R. Wheeler, Administrator U.S. Environmental Protection Agency). <https://science.house.gov/imo/media/doc/9.19.19%20Wheeler%20Testimony.pdf> [accessed 20 November 2019].
- Siddle J. 2014. I know where you were last summer: London’s public bike data is telling everyone where you’ve been. <https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html> [accessed 22 August 2019].
- Sister Study. n.d. The Sister Study data sharing policy. <https://www.sisterstudystars.org/Public/Sister/Documents/Data%20Access%20Policies%20and%20Procedures.pdf> [accessed 18 November 2019].
- Stout DMI 2nd, Bradham KD, Egeghy PP, Jones PA, Croghan CW, Ashley PA, et al. 2009. American Healthy Homes Survey: a national study of residential pesticides measured from floor wipes. *Environ Sci Technol* 43(12):4294–4300, PMID: 19603637, <https://doi.org/10.1021/es8030243>.
- Sweeney L. 2002. k-Anonymity: a model for protecting privacy. *Int J Unc Fuzz Knowl Based Syst* 10(5):557–570, <https://doi.org/10.1142/S0218488502001648>.
- Sweeney L. 2011. Patient Identifiability in Pharmaceutical Marketing Data. Data Privacy Lab Working Paper 1015. <https://dataprivacylab.org/projects/identifiability/pharma1.pdf> [accessed 1 November 2018].
- Sweeney L. 2013. *Matching Known Patients to Health Records in Washington State Data*. ID 2289850. Rochester, NY: Social Science Research Network.
- Sweeney L, Von Loewenfeldt M, Perry M. 2018. Saying it’s anonymous doesn’t make it so: re-identifications of “anonymized” law school data. *Technol Sci*. 2018111301, PMID: 25078429, <https://techscience.org/a/2017082801> [accessed 17 January 2018].
- Sweeney L, Yoo J, Perovich L, Boronow K, Brown P, Brody J. 2017. Re-identification Risks in HIPAA Safe Harbor Data: a study of data from one environmental health study. *Technol Sci*. 2017082801, PMID: 30687852.
- Tanner A. 2017. *Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records*. Boston: Beacon Press.
- Taylor L. 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society* 4(2):2053951717736335, <https://doi.org/10.1177/2053951717736335>.
- United States Congress. 2008. The Genetic Information Nondiscrimination Act of 2008 (GINA). Pub L No. 110–233. H.R. 493 (110th Congress 28 May 2008), <https://www.gpo.gov/fdsys/pkg/PLAW-110publ233/html/PLAW-110publ233.htm> [accessed 1 November 2018].
- Vermont Department of Health. 2019. Vermont Uniform Hospital Discharge Data System. <https://www.healthvermont.gov/health-statistics-vital-records/health-care-systems-reporting/hospital-discharge-data> [accessed 21 November 2019].
- Vermont Secretary of State. 2019. Data Broker Search. <https://www.vtsonline.com/online/DataBrokerInquire/> [accessed 18 November 2019].
- Virginia Health Information. 2018. Data Directory: Health Care Decision Support Data (updated 2018-01-23). https://vhi.org/files/PDFs_to_download_from_web/Data_Directory_2017.pdf [accessed 21 November 2019].
- von Thenen N, Ayday E, Cicek AE. 2019. Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics* 35(3):365–371, PMID: 30052749, <https://doi.org/10.1093/bioinformatics/bty643>.
- Wang R, Li YF, Wang X, Tang H, Zhou X. 2009. Learning your identity and disease from research papers: information leaks in genome wide association study. In: *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS '09)*. CCS 2009. 9–13 November 2009. Chicago, IL. New York, NY: ACM, 534–544, <https://doi.org/10.1145/1653662.1653726>.
- Weisel CP, Zhang J, Turpin BJ, Morandi MT, Colome S, Stock TH, et al. 2005. Relationship of Indoor, Outdoor and Personal Air (RIOPA) study: study design, methods and quality assurance/control results. *J Expo Anal Environ Epidemiol* 15(2):123–137, PMID: 15213705, <https://doi.org/10.1038/sj.jea.7500379>.
- Zarate OA, Brody JG, Brown P, Ramirez-Andreotta MD, Perovich L, Matz J. 2016. Balancing benefits and risks of immortal data: participants’ views of open consent in the Personal Genome Project. *Hastings Cent Rep* 46(1):36–45, PMID: 26678513, <https://doi.org/10.1002/hast.523>.
- Zipf G, Chiappa M, Porter K, Ostchega Y, Lewis B, Dostal J. 2013. National Health and Nutrition Examination Survey: plan and operations, 1999–2010. *Vital and Health Statistics* 1(56): PMID: 25078429.

Submission ID: 1436

Date: 1/11/2020

Name: Kenneth J. Ottenbacher

Name of Organization: University of Texas Medical Branch, Galveston, TX

Type of Data of Primary Interest: Clinical

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Aging, Rehabilitation and Disability

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Consider including data sharing plans in the scored criteria for NIH study section reviewers. The current Draft Plan indicates that internal ICO staff will review the adequacy of the plans and provide feedback to investigators. We contend that peer review would be more efficient and more likely to lead to best practices. The Draft Plan indicates that investigators can request funding for creating public use data and making them available. All funded grant activities should be peer reviewed and scored.

Section VI: Data Management and Sharing Plans:

A "persistent identifier or other standard indexing tools" is mentioned as optional in the Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan (element #4. Data Preservation, Access, and Associated Timelines). A unique digital object identifier (DOI) is critical for efficiently crediting the original grant as well as documenting the longer-term return on investment of public funds by NIH. This should not simply be a preferred step, but

rather a required one with a standardized approach and supported infrastructure. Further, properly citing the data source and DOI should also be required of data users. This may best be enforced by scientific journals and editors.

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Description:

Submission ID: 1437

Date: 1/11/2020

Name: Maureen McArthur Hart

Name of Organization: Global Genes

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Patient Advocacy Organization

Type of Organization - Other:

Role: Patient Advocate

Role - Other:

Domain of Research Most Important to You or Your Organization:

Rare disease

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Data sharing is crucial to advancing rare disease research efforts. However, if data are not shared rapidly, resources in terms of funding, time, and effort on the part of both researchers and patients/advocates are often duplicated or, in the worse case scenarios, are wasted. The draft NIH Policy for Data Management and Sharing specifies that data must be shared in a "timely manner" but provides no specific definition. A specific definition, and one that requires data sharing as soon as possible, should be included. As a suggestion, for the purpose of speeding research and minimizing duplication of resources, data sharing upon the first publication using data generated should be considered.

In addition, while data sharing is critical, it may not have much value without the tools necessary to replicate the data. The Data Sharing and Management Plan should be required to include a discussion of the management, preservation, and sharing of software, animal models, cell lines, etc.

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:**Section VI: Data Management and Sharing Plans:**

Again, because of the significant need for data sharing to maximize the impact of research findings and minimize duplicative or wasteful efforts, the Data Management and Sharing Plan should be considered as a significant part of any application, worthy of review as a part of consideration of the Overall Impact Score, assessing the likelihood for both an extended and significant impact. As such, the Data Management and Sharing Plan should be a part of the review process, considered when reviewing and scoring the grant award or contract.

Rather than stating that "NIH encourages shared scientific data to be made available...", the policy should state that "NIH requires shared scientific data to be made available..."

The policy states that "...shared scientific data to be made available as long as it is deemed useful to the research community or the public" but does not include any criteria by which this judgment will be made or by whom. Potential further specifications could include a time period be specified following the last publication generated using the scientific data or consideration of whether the data are no longer useful be made by an advisory committee.

Section VII: Compliance and Enforcement:

The inclusion of compliance and enforcement is recognized as needed to ensure active and broad data sharing. However, the requirement should be strengthened for post award period. The policy states "After the end of the funding period, non-compliance with the NIH ICO-approved Plan may be taken into account by the funding NIH ICO for future funding decisions for the recipient institution." This should be strengthened to state "...non-compliance with the NIH approved Plan will be taken into account...." In addition, non-compliance with a Data Sharing and Management Plan must be taken into account for future applications from the researcher as a Principal Investigator (PI)/Project Lead, co-PI, consultant, collaborator, or subaward.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:**Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:****Other Considerations Relevant to this DRAFT Policy Proposal:**

Investigators should be rewarded for the impact of their work when they actively and broadly share scientific data. A section of grant, contract, and scholarship applications could be devoted to examples of data sharing and the resulting impact of data sharing. This section should be considered in the review and scoring, with investigators demonstrating past significant sharing

receiving credit in investigator scoring. This could be particularly important for young investigators where they have not yet built a significant publication record, but could demonstrate significant impact through data sharing.

Attachment:

Description:

Submission ID: 1438

Date: 1/11/2020

Name: Kimberly Sabelko

Name of Organization: Susan G Komen

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Patient Advocacy Organization

Type of Organization - Other:

Role: Patient Advocate

Role - Other:

Domain of Research Most Important to You or Your Organization:

Breast cancer

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Susan G Komen applauds the NIH for revising its policy for data management and sharing as part of NIH's ongoing efforts to lead the biomedical community toward a culture of broad and routine data sharing that will improve the scientific process, enabling research that will impact health outcomes and patients' lives. Data sharing is critical to the scientific process as it allows researchers to build upon current knowledge, reduce duplicative research, independently validate findings, promote reproducibility, and generate new knowledge faster, including exploring new frontiers using big data approaches. As a representative and advocate for people impacted by a breast cancer diagnosis--i.e., beneficiaries of and participants in research--Komen supports policies such as this that enable researchers to maximize the use of data to solve problems in biomedical/breast cancer research.

Section II: Definitions:

Scientific data is defined as encompassing the data and metadata needed to replicate and validate research findings. The definition excludes "laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens." We found the definition to be clear and succinct and the exclusions appropriate.

We recommend that the NIH emphasize that this definition represents the minimal data to be shared and encourage (not require) researchers to share both data and methods applied to data to the greatest extent possible (beyond the definition of scientific data). Some of the excluded items listed in the definition could be of importance for other researchers to build on prior findings and/or improve reproducibility. Additionally, we recommend that supporting documentation – when and how the data was generated, what methods were used to collect data, address missing and erroneous data, etc.-- be shared alongside the data themselves. It would prevent the many datasets currently shared that cannot be used due to a lack of supporting documentation associated with them. It is important to state that data sharing must also include negative data to better reduce overlap in efforts by multiple groups. By having this type data widely available, researchers may be more efficient with funding and reduce duplication of effort on similar questions that have already been asked and disproven.

Section III: Scope:

We were pleased to see that the revised policy will apply to all NIH-funded or conducted research generating scientific data, regardless of data type, size, or the requested amount of funding.

Section IV: Effective Date(s):

Section V: Requirements:

To be valuable, the shared data needs to be collected consistently and well characterized with all data elements and methods documented and recorded. We agree that having a well-documented data sharing plan for all research projects is vital. It is also important that these plans be measurable, compliance be monitored, and investigators be held accountable for managing and sharing their data.

The draft policy states "the funding NIH ICO may request additional or specific information to be included within the Plan in order to meet expectations for data management and data sharing in support of programmatic priorities or to expand the utility of the scientific data generated from the research." We recognize that each of the NIH Institutes, Centers, and Offices (ICOs) need flexibility in adapting the NIH's Data Management and Sharing policy and that the DRAFT Policy outlines the minimum expectations for Data Management and Sharing Plans. We also recognize the need to ease the burden on investigators and are concerned that having data sharing policies that differ based on the ICO will lead to confusion and increased administrative work for investigators. Consistent recommendations and requirements for data sharing plans would simplify requirements for investigators as well as the NIH and better ensure the ability to share and integrate data across agencies meeting the goal of the policy. We agree that the flexibility of the DRAFT Policy will require additional implementation guidance for the ICOs.

Section VI: Data Management and Sharing Plans:

Having a well-described data sharing plan and protocols governing data collection and management prior to the start of the research project is a vital component of data sharing, affecting data quality, consistency, and usability. We agree that the plans should "prospectively outline where, when, and how scientific data will be managed and shared" as well as "which of these scientific data will be shared." We also agree with the NIH's emphasis that investigators ensure patient privacy is preserved and confidential data is protected at all times. This is especially important when the data relate to small and/or underserved populations where there may be distrust with the current system and where data sharing may have the greatest impact.

The DRAFT policy proposes that for extramural awards, the data management and sharing plans could be submitted at "Just-In-Time" such that "only those applicants likely to be funded would submit Plans." We recommend that the data management and sharing plans be evaluated during peer review, by NIH staff and/or peer reviewers. We agree it's important to minimize the burden on applicants, and in some instances, depending on the review process, those applicants not likely to be funded could be identified earlier in the process (for example, due to lack of alignment with request for letters of intent) and not required to submit detailed, full applications with data sharing plans. We believe including an assessment of the plan during peer review emphasizes the importance and value of data sharing and encourages investigators to consider their data sharing plan as an inherent part of designing the research project. Criteria will need to be established to guide staff/reviewers in determining whether Plans are feasible and adequate. Applications that include inadequate data management and sharing plans should not score as highly in peer review.

The DRAFT policy states that "NIH may make Plans publicly available." We recommend that the NIH state that plans will be made publicly available. In addition to allowing the public to see how researchers are planning to share their data, it will help other researchers drafting their own plans, especially those who are currently not well versed in sharing their data.

Section VII: Compliance and Enforcement:

It is important that data management and sharing plans be measurable, compliance be monitored, and investigators be held accountable for managing and sharing their data and evaluated throughout the grant term. Data sharing updates should be included as part of any submitted progress reports. The DRAFT policy states "Failure to comply with the Terms and Conditions (...) may affect future funding decisions" and "after the end of the funding period, non-compliance with the NIH ICO-approved Plan may be taken into account by the funding NIH ICO for future funding decisions for the recipient institution." We recommend that the policy state that non-compliance will be taken into account for future funding decisions. We believe penalties and incentives are needed to enforce and encourage appropriate and meaningful

data sharing, and we encourage NIH to continue to consider appropriate carrots and sticks that will facilitate compliance with data sharing plans.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

We agree on the importance of allowing costs that will facilitate data management and sharing. The DRAFT Guidance states "local data management considerations, such as unique and specialized information infrastructure necessary to provide local management, preservation, and access to data" could be allowable costs. It is unclear how the assessment for "unique and specialized information" will be made and whether this will allow institutions to continue to create silos of information counter to the philosophy of data sharing.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

We commend NIH for developing this supplemental DRAFT Guidance, as it provides clarity regarding the key elements of a good data sharing plan. As stated earlier, we agree that the plans should specify which scientific data will be shared, as well as when, how and where the data will be managed and shared. We wish to emphasize the importance of sharing any additional metadata, information, or documentation about the scientific data that will be made publicly available to ensure datasets are all useable and enhance their value to the research community.

Shared data will only enable research that will impact health outcomes and patients' lives. If it is of high quality and used by the research community. The selection of data repositories is important and we recommend that justifications for the selected repository(ies) be a required element, especially in cases where there are multiple options. The ease by which shared data can be accessed varies among repositories and should be considered in the data sharing plan. NIH supported repositories should be strongly considered as a first option for bringing scientific data resulting from NIH funded or conducted research and the plan should clarify why a different repository would be a better choice.

The DRAFT Guidance states "if certain elements of a Plan have not been determined at the time of submission, an entry of 'to be determined' may be acceptable if a justification is provided along with a timeline or appropriate milestone at which a determination will be made." Information about data sharing is of importance in reviewing applications. We recommend that the guidance require applicants to include their overall expectations/approach to data management and sharing for any elements of a Plan for which details are not available at the time of submission.

Other Considerations Relevant to this DRAFT Policy Proposal:

We appreciate the opportunity to provide comments re: the DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance.

Submission ID: 1439

Date: 1/11/2020

Name: Sarah Nusser

Name of Organization: Iowa State University

Type of Data of Primary Interest:

Type of Data of Primary Interest - Other:

Type of Organization: University

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

Iowa State University has a growing NIH funded research portfolio across a broad range of species and with focal areas of interest in anti-microbial resistance, molecular mechanisms, gene editing, addiction and social behavior, neuroscience, mechanisms of infectious disease, vaccines and virology.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

We support the focus on prospectively planning for which scientific data will be shared as this is an important aspect to ensuring quality and value in the data to be shared.

Section II: Definitions:

Computer code is often used to manipulate data, produce indices or other measures, or implement other types of computations. Would code be considered part of scientific data?

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

We support the use of a Just-in-Time approach to providing the data management and sharing plan (DMSP) only when an award is to be made.

We hope that the RFP will ask that key elements be articulated in the grant proposal, such as what data would be collected and how, how data would be manipulated, and what data would be of value to share.

The two-page restriction will be viewed favorably by researchers. However, a large number of data elements are specified in the supporting documentation, some of which may be complex, e.g., assuring appropriate restrictions are put in place for sensitive data. This raises questions about how much information can reasonably be provided in two pages. Further, if the DMSP is meant to be a compliance document, articulating these more complex problems might be a concern unless a link to other documentation is expected.

Thinking ahead, presumably the DMSP would ultimately be in software system to enable visibility to NIH program staff, institutions and researchers, so is the 2-page restriction sensical?

We appreciate the explicit articulation of allowable costs for data management and sharing. Researchers are not always aware that these costs are allowed and this is a step forward.

Section VII: Compliance and Enforcement:

We appreciate the weight that NIH is placing on DMSPs as contractual terms and support this as a way to ensure that researchers are adhering to the intent to publicly share research data when appropriate and worthwhile.

That said, this is a new kind of compliance process and universities are not well positioned to expand their staff to support a new form of compliance. Implementation details need to be worked out. It would be useful for NIH to work with research institutions on implementation details.

For examples, how might contract terms might actually be expressed and evaluated? Presumably, the institution would need to have an approval process prior to submitting the plan to ensure that the proposed plan can actually be executed by the institution. In addition, institutions would need access to any version of the plan through a portal.

We strongly encourage NIH to work with research institutions to ensure that this new form of compliance is well understand and can be implemented effectively and at minimal cost.

Submission ID: 1440

Date: 1/11/2020

Name: Peter Sorger

Name of Organization: Harvard Program in Therapeutic Science

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: gene expression, proteomics, imaging, basic biomedical, etc.

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

systems pharmacology

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

LSP_Response_to_DRAFT NIH Policy for Data Management and Sharing_20200110.pdf

Description:

Response to “DRAFT NIH Policy for Data Management and Sharing”

Peter Sorger (Otto Kraye Professor of Systems Biology at Harvard Medical School, Head of the Harvard Program in Therapeutic Science, Founding Director of the HMS Laboratory of Systems Pharmacology)

Laura Maliszewski (Executive Director of the Harvard Program in Therapeutic Science and the HMS Laboratory of Systems Pharmacology)

Catherine Luria* (Scientific Program Manager, Harvard Program in Therapeutic Science and the HMS Laboratory of Systems Pharmacology) *Please contact catherine_luria@hms.harvard.edu for further information.

Over the past five years we have been involved in multiple large NIH/NCI grants that involve data sharing and management activities (including the NIH LINCS and IDG programs and the NCI HTAN program). We therefore have considerable experience in implementing such activities from the perspective of practicing scientists and NIH grantees. We have commented on this topic in a perspective in *Science Translational Medicine*¹ and written multiple papers attempting to improve the reproducibility of preclinical assays of drug response²⁻⁴.

Overall, we are highly supportive of the development of a more robust set of policies and associated infrastructure for data management and sharing that conform with FAIR principles. NIH’s effort in seeking and incorporating feedback from the scientific community on the October 2018 *Proposed Provisions for a Draft NIH Data Management and Sharing Policy* has resulted in what we view as a number of positive changes. In particular, the statement that “costs associated with data management and data sharing may be allowable under the budget for the proposed project” and the associated *Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing* acknowledges the resources required to comply with FAIR principles, a critical point that was not addressed in the previous draft.

However, some of the concerns that we expressed in response to the *Proposed Provisions* remain. We believe that the implementation of policies under the current draft as well as successful, consistent adherence to data management plans proposed by individual investigators can only be accomplished as a part of a multi-part multi-year strategy that also includes: (i) education – including education of graduate students and fellows (ii) development of infrastructure for validating, storing and disseminating diverse types of data (iii) substantial investment in innovation and computational tools and approaches (iv) incentives for timely and useful data deposition as opposed to simple mandates and penalties.

First, we feel strongly that NIH should provide education and training materials to principal investigators before the requirements described in the November 2019 draft document are implemented. NIH has developed a very flexible draft document to accommodate a wide range of research areas and approaches, but more guidance is needed to ensure that PIs provide reasonable data plans that can be implemented successfully once a grant is funded.

Controlled vocabulary for data types and how data will be shared (e.g. individual vs. aggregated vs. summarized) should be provided along with guidance on minimum standards for the types of materials that should accompany scientific data, including standardized formats for study protocols and information about data collection instrumentation, software and code.

It is also essential to recognize that for most of the data we generate, there are no generally accepted standards or established repositories. The annotation and re-use of heterogeneous data arising from perturbational studies (the vast majority of the mechanism-oriented research in the NIH portfolio) is fundamentally different from storing and disseminating a single type of data on a steady-state sample (e.g. a genome sequence). Formats and reporting standards have not yet been developed for most types of microscopy data, the many variants of mass spectrometry, multiplex immuno-assays on cell and tissues lysates or on components of the microenvironment and emerging data types such as spatial transcriptomics or multiplex imaging among others. New standards will be required to adequately annotate experiments in which these types of data are collected over time following genetic or drug-mediated perturbation of a system.

It is not sufficient to establish ontologies for these data: tools must be developed to annotate data according to relevant standards, to impose uniform vocabularies and to validate annotations. While the current draft policy addresses the possibility of budgeting for data management, substantial investment in software development and hardening is required to implement FAIR data standards across all data types. The apps we all enjoy using (e.g. Google Maps) have involved a much higher level of refinement than any of the code we use for storage and annotation of scientific data. cBioPortal and Cytoscape are two examples of well-developed code – both required large teams and many years of investment. In addition to the extensive work required for data types (and even file formats) for which no tools currently exist, existing infrastructure must be more actively supported. For example, the OMERO image management standard that we helped to develop over a decade at MIT (now in wide-spread use) has never received any NIH support despite multiple attempts. The entire development team was moved from the US to the UK, where it is now headed by Jason Swedlow with EU/UK funding.

Other considerations:

- Assessment of data management plans: Plans will be assessed as part of Just-in-Time for extramural awards. While this saves effort on the part of reviewers and of proposal developers who will ultimately not be funded, it prevents data management plans from being assessed by a team of scientific reviewers, which is concerning.
- Data archiving: Draft guidance on allowable cost states that recurring fees for sharing and preserving data in existing repositories may be included in proposal budgets. NIH does not explicitly address whether data storage will continue to be funded after the grants which supported data production have ended, how long data must be preserved, and how these decisions might be revised if a particular type of data is no longer of interest to the community due to technological advancements. The possibility of retiring some types of data (e.g. RAW files or image-based screening data) after a predetermined period must also be considered.

- Standards development: We hope that NIH will catalyze community groups to develop community-based standards for data types for which no standards exist already. The task of developing these standards is too large for individual grantees and standards developed by small groups are unlikely to result in FAIR data. Standardization efforts should be international; particularly in the area of pre-clinical, basic research data, EMBL/EBI is well ahead of anything in the US.
- Standards dissemination: Before the proposed data management and data sharing requirements are implemented, existing data standards should be more actively supported and disseminated. Many existing standards are currently difficult to locate and sometimes poorly documented, meaning that research groups struggle to find and correctly implement existing standards. The Common Data Element Resource Portal suggested by NIH in *Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan* is largely limited to disease-specific clinical studies and is not necessarily relevant to many basic research projects. A more broadly applicable and more easily searchable resource than the Common Data Element Resource Portal should be in place to provide information about existing data standards for NIH investigators. For data types for which no NIH data repository exists, a list of accepted non-NIH repositories will be required; these will need persistent unique identifiers for deposited data. For example, the PRIDE database maintained by EMBL/EBI is becoming the standard for deposition of proteomics data. NIH should support and mirror this and similar types of repositories, by analogy with mirrored genomics databases. If existing standards are not easy to find, some groups may “reinvent the wheel” and develop new, redundant standards, which ultimately reduce data FAIRness.
- Software tools: As noted above, a standard is useless without the software infrastructure needed to implement and validate it. In our experience, this often requires some ability in scripting – we teach all of our trainees basic Python coding skills. It is for this reason that data annotation and education and closely interrelated.

References

1. AlQuraishi M, Sorger PK. Reproducibility will only come with data liberation. *Sci Transl Med*. 2016 May 18;8(339):339ed7. PMID: PMC5084089
2. Hafner M, Niepel M, Chung M, Sorger PK. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat Methods*. 2016 Jun;13(6):521–7. PMID: PMC4887336.
3. Niepel M, Hafner M, Mills CE, Subramanian K, Williams EH, Chung M, Gaudio B, Barrette AM, Stern AD, Hu B, Korkola JE, LINCS Consortium, Gray JW, Birtwistle MR, Heiser LM, Sorger PK. A Multi-center Study on the Reproducibility of Drug-Response Assays in Mammalian Cell Lines. *Cell Syst*. 2019 Jul 24;9(1):35-48.e5. PMID: PMC6700527
4. Hafner M, Niepel M, Sorger PK. Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics. *Nat Biotechnol*. 2017 Jun 7;35(6):500–502. PMID: PMC5668135

Submission ID: 1441

Date: 1/11/2020

Name: Moffitt Cancer Center

Name of Organization: H. Lee Moffitt Cancer Center & Research Institute, Inc.

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Nonprofit Research Organization

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

epidemiology, infectious diseases, genomic, cancer, etc.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Moffitt Cancer Center (MCC) greatly appreciates the opportunity to submit comments regarding the National Institutes of Health's (NIH) request for public input on trans-NIH data management and sharing policies. MCC is an NCI Designated Comprehensive Cancer Center that is home to more than 5,000 medical and scientific professionals and the proud generator of guideline, policy, and paradigm shifting advances in basic, population, translational, and clinical science within cancer research, treatment, and beyond.

As an institution, MCC is committed to increasing scientific communication and transparency by making data as widely and freely available as possible, in compliance with appropriate federal policies, while protecting participant & investigator personal rights.

On Behalf of Moffitt Cancer Center:

Dr. Shelley Tworoger Dr. Heather Jim Dr. James Mule

Dr. Dana Rollison Dr. Jhanelle Gray Dr. Travis Gerke

Dr. John Cleveland Dr. Jose Conejo-Garcia Dr. Kristen Scott

Dr. Peter Kanetsky Dr. Frederick Locke Brian Springer

Dr. Elsa Flores Dr. Derek Duckett Lowell Smith

Dr. Thomas Brandon Dr. Brooke Fridley Susan Sharpe

Dr. Erin Siegel Dr. Susan Vadapampil

Section I: Purpose

MCC concurs with the recommendations and opinions provided by NIH regarding the importance of data sharing and the role of good data management practices within publicly funded research and beyond. In order to best bring about change and compliance within the broader research community, MCC recommends that non-specific language be replaced with clear and decisive verbiage similar to what is included within the NIH Public Access Policy of Division F. Section 217 of PL 111-8 (Omnibus Appropriations Act, 2009); which states "The Director of the National Institutes of Health ("NIH") shall require in the current fiscal year and thereafter that all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine's PubMed Central an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication: Provided, that the NIH shall implement the public access policy in a manner consistent with copyright law."

In doing so, the NIH will remove the temptation for researchers and research entities to withhold data for unspecified lengths of time as can be the current case. Furthermore, by applying specific deadlines and terminology to the request, NIH will be removing the temptation for administration to regard NIH's recommendations as mere "suggestions" which are easily dismissed in favor of simplified processes, or legal obfuscation.

MCC recommends that NIH call for the "Effective Date(s)" to be negotiated specifically in accordance with the final publication dates, i.e. within 12 months of final commercialized publication and/or upon the release of the embargo period— and maintain the same timeline for retrieval and dissemination. Should no publication be forthcoming from the funded project, then MCC recommends that a curated selection of data and/or results of the funded project be released to a federal repository within 12 months of the funding period closure date. Doing so provides multiple benefits, the chief of which is the elimination of the publication bias of only reporting positive results.

Likewise, MCC believes strongly that any requirements passed by NIH should be applied to newly collected data or for new studies only upon the emergence of the final DMP resolution; however, special consideration should be granted to studies leveraging previously collected data as these studies should be grandfathered into an exception because new studies may not be able to comply with new rules based on historical or legal (e.g., human subjects) reasons.

Section II: Definitions:

Per the NIH's recommendation's in Section I: Purpose, it is MCC's recommendation that this section addresses the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles and that these principles should be included within the formal definitions of required components for newly collected data and/or for new studies commissioned after the formal adoption of NIH DMP policy. In the same vein, MCC recommends that a FAIR policy checklist be included in an actionable, measurable component of scoring rather than limiting to their inclusion at the JIT submission.

MCC recommends providing a list of existing acceptable categories of Scientific Data types, to be provided to investigators in accompaniment of the statement "The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings, regardless of whether the data are used to support scholarly publications."

The remainder of the definitions proffered are acceptable and concise.

Section III: Scope:

Section IV: Effective Date(s):

MCC recommends the adoption of the effective dates of 12 months after the official date of final-publication, and/or conclusion of the funding period. MCC encourages selecting a specifically defined comprehensive and definitive timeline with clear beginning and ending periods which will apply to newly collected data and/or studies opened after the formal adoption of the new DMP requirements.

Section V: Requirements:

MCC requests that NIH unite the NIH Institute and Center Operations (ICO) Plans to reduce the efforts necessary to maintain disparate recommendations and requirements, and enable scientists to focus on their research, instead of comparing ICO DMP policies. If it is not the intention of NIH to unite the ICO specific policies and plans for data management at this time, then MCC recommends an improved plan to assemble investigator feedback and commentary on ICO requirements at such time when these plans are being assembled.

Section VI: Data Management and Sharing Plans:

The disambiguation of the Data Management and Sharing Plan from the application, and instead its incorporation into the Just-in-Time information submitted to the funding ICO is unsatisfactory, and presents risks to peer review and accountability, by inviting investigators at the time of submission to circumnavigate requirements and submit their DMP as "to be determined." MCC recommends the removal of non-specific language such as "encourages" and "deemed useful" and the incorporation of specific and precise terminology clearly identifying expected requirements. In the case of "deemed useful" there exists a disparity of understanding between what the originator PI may consider as "useful" data, vs. what a

scientist within a different field might find useful, failure to clarify this point will result in ambiguity and continue to expand the gap between what data is reported and what data is not.

While, MCC investigators appreciate the ability to update DMP plans during regular reporting intervals, the mechanism for seeking approval and submitting such revisions requires precise definition & guidance per ICO.

Regarding the listed Plan Assessment, MCC suggest that any approved DMP be able to satisfy a checklist of FAIR principles at all stages of submission for all awards and contracts. Again, MCC recommends that this should be applied to newly collected data or for new funded studies; keeping in mind that, studies leveraging previously collected data will need to be grandfathered in under any new rulings.

We acknowledge that NIH has been at the forefront of creating and maintaining reliably operated public data repositories and recommend that NIH to create a curated list of trustworthy & Peer-Reviewed repositories consisting of both NIH and third-party independent options to which funded NIH investigators must submit. MCC would recommend that such a list become a requirement for newly conducted studies and for extensions of existing studies when ethically and legally possible, and it is our thought, that through requiring a specific list and a set of specific locations for data, NIH will ensure the continued federally funded data can be reliably found across the spectrum of approved NIH resources. If posting to such data repositories is not possible, the reasons for that should be clearly articulated and alternate plan for ensuring data accessibility should be described in detail.

Section VII: Compliance and Enforcement:

It is the opinion of MCC that non-compliance with NIH and the ICO's requirements for a data sharing plan be considered in the same light as a publication non-compliance, in that the funding of the Principle Investigator who is non-compliant with their DMP has their funding withheld and/or risks additional punishments consistent with research misconduct. MCC formally recommends against punishing the entire institution, and instead enabling the institution to continue supporting compliant researchers, while both NIH, the ICO, and institution work to resolve the non-compliant researchers.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

The guidance related to allowable costs should be more explicit with the inclusion of acceptable cost types, such as long-term maintenance fees.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

MCC wishes NIH to reconsider the following sentences "If certain elements of a Plan have not been determined at the time of submission, an entry of "to be determined" may be acceptable if a justification is provided along with a timeline or appropriate milestone at which a determination will be made. Note, NIH does not expect researchers to share all scientific data

generated in a study." Specifically, we recommend that datasets used to generate the publications coming out of the study should be made available for other researchers based on appropriate legal and/or human subjects' considerations.

Considering this segment, researchers can effectively determine at some future date which data are useful to share. As a result, the community could see Plans with "We will determine which data are technically aligned with public consumption on [future date]" and then decide no data are fit for purpose after funding has been delivered.

MCC recommends that NIH consider the potential repercussions of the numerous exit strategies that exist within the stated DMP plan for those not wishing to share data:

- (scientific utility) = all findings are null so there is no utility in sharing unpublished data,
- (validation) = we provide independent validation in a separate cohort in this publication negating the need for others to validate on the present data,
- (privacy) = PHI is contained within these data so they cannot be shared (with no justification for why deidentification is not possible),
- (cost) = the costs to share these data exceed the scope of the project, (consistency with community practices) = community practices around data sharing are virtually absent, thus reinforcing the status quo, and
- (data security) = to minimize risk to patients with regard to release of PHI we opted not to publicly share data (again with no justification for why deidentification is not possible).

Regarding the restrictions on sharing, the NIH plan does not address the complexities faced by large-scale consortia studies (e.g. ORIEN), and MCC requests that language addressing these public-private partnerships be added to the Draft DMP, and in the context of ensuring the minimum essential data given varying institutional standards and requirements. Similarly, consortia that bring together existing data from many on-going or completed studies often include research studies that have ICFs that do not allow public data sharing, are not NIH funded, and/or are based in other countries with different laws regarding data privacy. Given the emphasis that NIH has placed on leveraging existing data in consortia to advance science – more specific guidance should be provided to researchers who conduct this type of research in which it can be difficult to comply with the data sharing policies as currently articulated due to needing to address a myriad of different compliance and legal issues.

Other Considerations Relevant to this DRAFT Policy Proposal:

MCC requests 3 pages minimum be assigned to the DMP.

Attachment:

NIH Data Comment v3.docx

Description:

V3 of comment

Comment: DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Moffitt Cancer Center (MCC) greatly appreciates the opportunity to submit comments regarding the National Institutes of Health's (NIH) request for public input on trans-NIH data management and sharing policies. MCC is an NCI Designated Comprehensive Cancer Center that is home to more than 5,000 medical and scientific professionals and the proud generator of guideline, policy, and paradigm shifting advances in basic, population, translational, and clinical science within cancer research, treatment, and beyond.

As an institution, MCC is committed to increasing scientific communication and transparency by making data as widely and freely available as possible, in compliance with appropriate federal policies, while protecting participant & investigator personal rights.

On Behalf of Moffitt Cancer Center:

Dr. Shelley Tworoger
 Dr. Dana Rollison
 Dr. John Cleveland
 Dr. Peter Kanetsky
 Dr. Elsa Flores
 Dr. Thomas Brandon
 Dr. Erin Siegel

Dr. Heather Jim
 Dr. Jhanelle Gray
 Dr. Jose Conejo-Garcia
 Dr. Frederick Locke
 Dr. Derek Duckett
 Dr. Brooke Fridley
 Dr. Susan Vadaparampil

Dr. James Mule
 Dr. Travis Gerke
 Dr. Kristen Scott
 Brian Springer
 Lowell Smith
 Susan Sharpe

Section I: Purpose

MCC concurs with the recommendations and opinions provided by NIH regarding the importance of data sharing and the role of good data management practices within publicly funded research and beyond. In order to best bring about change and compliance within the broader research community, MCC recommends that non-specific language be replaced with clear and decisive verbiage similar to what is included within the NIH Public Access Policy of Division F. Section 217 of PL 111-8 (Omnibus Appropriations Act, 2009); which states "The Director of the National Institutes of Health ("NIH") shall require in the current fiscal year and thereafter that all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine's PubMed Central an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication: Provided, that the NIH shall implement the public access policy in a manner consistent with copyright law."

In doing so, the NIH will remove the temptation for researchers and research entities to withhold data for unspecified lengths of time as can be the current case. Furthermore, by applying specific deadlines and terminology to the request, NIH will be removing the temptation for administration to regard NIH's recommendations as mere "suggestions" which are easily dismissed in favor of simplified processes, or legal obfuscation.

MCC recommends that NIH call for the "Effective Date(s)" to be negotiated specifically in accordance with the final publication dates, i.e. within 12 months of final commercialized publication and/or upon the release of the embargo period— and maintain the same timeline for retrieval and dissemination. Should no publication be forthcoming from the funded project, then MCC recommends that a curated selection of data and/or results of the funded project be released to a federal repository within 12 months of the funding period closure date. Doing so provides multiple benefits, the chief of which is the elimination of the publication bias of only reporting positive results.

Likewise, MCC believes strongly that any requirements passed by NIH should be applied to newly collected data or for new studies only upon the emergence of the final DMP resolution; however, special consideration should be granted to studies leveraging previously collected data as these studies should be grandfathered into an exception because new studies may not be able to comply with new rules based on historical or legal (e.g., human subjects) reasons.

Section II: Definitions

Per the NIH's recommendation's in Section I: Purpose, it is MCC's recommendation that this section addresses the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles and that these principles should be included within the formal definitions of required components for newly collected data and/or for new studies commissioned after the formal adoption of NIH DMP policy. In the same vein, MCC recommends that a FAIR policy checklist be included in an actionable, measurable component of scoring rather than limiting to their inclusion at the JIT submission.

MCC recommends providing a list of existing acceptable categories of Scientific Data types, to be provided to investigators in accompaniment of the statement "The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings, regardless of whether the data are used to support scholarly publications."

The remainder of the definitions proffered are acceptable and concise.

Section III: Scope

[no comment]

Section IV: Effective Date(s)

MCC recommends the adoption of the effective dates of 12 months after the official date of final-publication, and/or conclusion of the funding period. MCC encourages selecting a specifically defined comprehensive and definitive timeline with clear beginning and ending periods which will apply to newly collected data and/or studies opened after the formal adoption of the new DMP requirements.

Section V: Requirements

MCC requests that NIH unite the NIH Institute and Center Operations (ICO) Plans to reduce the efforts necessary to maintain disparate recommendations and requirements, and enable scientists to focus on their research, instead of comparing ICO DMP policies. If it is not the intention of NIH to unite the ICO specific policies and plans for data management at this time, then MCC recommends an improved plan to assemble investigator feedback and commentary on ICO requirements at such time when these plans are being assembled.

Section VI: Data Management & Sharing Plans

The disambiguation of the Data Management and Sharing Plan from the application, and instead its incorporation into the Just-in-Time information submitted to the funding ICO is unsatisfactory, and presents risks to peer review and accountability, by inviting investigators at the time of submission to circumnavigate requirements and submit their DMP as "to be determined." MCC recommends the removal of non-specific language such as "encourages" and "deemed useful" and the incorporation of specific and precise terminology clearly identifying expected requirements. In the case of "deemed useful" there exists a disparity of understanding between what the originator PI may consider as "useful" data, vs. what a scientist within a different field might find useful, failure to clarify this point will result in ambiguity and continue to expand the gap between what data is reported and what data is not.

While, MCC investigators appreciate the ability to update DMP plans during regular reporting intervals, the mechanism for seeking approval and submitting such revisions requires precise definition & guidance per ICO.

Regarding the listed *Plan Assessment*, MCC suggest that any approved DMP be able to satisfy a checklist of FAIR principles at all stages of submission for all awards and contracts. Again, MCC recommends that this should be applied to newly collected data or for new funded studies; keeping in mind that, studies leveraging previously collected data will need to be grandfathered in under any new rulings.

We acknowledge that NIH has been at the forefront of creating and maintaining reliably operated public data repositories and recommend that NIH to create a curated list of trustworthy & Peer-Reviewed repositories consisting of both NIH and third-party independent options to which funded NIH investigators must submit. MCC would recommend that such a list become a requirement for newly conducted studies and for extensions of existing studies when ethically and legally possible, and it is our thought, that through requiring a specific list and a set of specific locations for data, NIH will ensure the continued federally funded data can be reliably found across the spectrum of approved NIH resources. If posting to such data repositories is not possible, the reasons for that should be clearly articulated and alternate plan for ensuring data accessibility should be described in detail.

Section VII: Compliance & Enforcement

It is the opinion of MCC that non-compliance with NIH and the ICO's requirements for a data sharing plan be considered in the same light as a publication non-compliance, in that the funding of the Principle Investigator who is non-compliant with their DMP has their funding withheld and/or risks additional punishments consistent with research misconduct. MCC formally recommends against punishing the entire institution, and instead enabling the institution to continue supporting compliant researchers, while both NIH, the ICO, and institution work to resolve the non-compliant researchers.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

The guidance related to allowable costs should be more explicit with the inclusion of acceptable cost types, such as long-term maintenance fees.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan

MCC wishes NIH to reconsider the following sentences "If certain elements of a Plan have not been determined at the time of submission, an entry of "to be determined" may be acceptable if a justification is provided along with a timeline or appropriate milestone at which a determination will be made. Note, NIH does not expect researchers to share all scientific data generated in a study." Specifically, we recommend that datasets used to generate the publications coming out of the study should be made available for other researchers based on appropriate legal and/or human subjects' considerations.

Considering this segment, researchers can effectively determine at some future date which data are useful to share. As a result, the community could see Plans with "We will determine which data are technically aligned with public consumption on [future date]" and then decide no data are fit for purpose after funding has been delivered.

MCC recommends that NIH consider the potential repercussions of the numerous exit strategies that exist within the stated DMP plan for those not wishing to share data:

- (scientific utility) = all findings are null so there is no utility in sharing unpublished data,
- (validation) = we provide independent validation in a separate cohort in this publication negating the need for others to validate on the present data,
- (privacy) = PHI is contained within these data so they cannot be shared (with no justification for why deidentification is not possible),
- (cost) = the costs to share these data exceed the scope of the project, (consistency with community practices) = community practices around data sharing are virtually absent, thus reinforcing the status quo, and
- (data security) = to minimize risk to patients with regard to release of PHI we opted not to publicly share data (again with no justification for why deidentification is not possible).

Regarding the restrictions on sharing, the NIH plan does not address the complexities faced by large-scale consortia studies (e.g. ORIEN), and MCC requests that language addressing these public-private partnerships be added to the Draft DMP, and in the context of ensuring the minimum essential data given varying institutional standards and

requirements. Similarly, consortia that bring together existing data from many on-going or completed studies often include research studies that have ICFs that do not allow public data sharing, are not NIH funded, and/or are based in other countries with different laws regarding data privacy. Given the emphasis that NIH has placed on leveraging existing data in consortia to advance science – more specific guidance should be provided to researchers who conduct this type of research in which it can be difficult to comply with the data sharing policies as currently articulated due to needing to address a myriad of different compliance and legal issues.

Other Considerations Relevant to this DRAFT Policy Proposal

MCC requests 3 pages minimum be assigned to the DMP.

Submission ID: 1442

Date: 1/11/2020

Name: Merce Crosas

Name of Organization: Harvard University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Data Management; Data Repositories

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

NIH Data Sharing Comments -Crosas.pdf

Description:

Response to DRAFT NIH Policy for Data Management and Sharing

January 10, 2020

Francis S. Collins, MD, PhD
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Submitted electronically: <https://osp.od.nih.gov/draft-data-sharing-and-management/>

RE: DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Dear Dr. Collins:

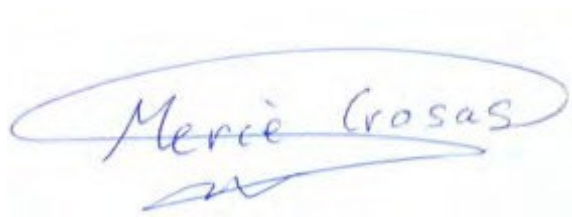
At the urging of several colleagues and collaborators, I'm responding to the new draft of the NIH Policy for Data Management and Sharing with a brief note to point out three main comments. My feedback is based on 15 years of experience working actively on data sharing and data management, as lead of the Dataverse project for sharing research data (<http://dataverse.org>), co-author of the Joint Declaration of Data Citation Principles and the FAIR Data principles, co-PI for the NIH Data Commons Consortium, and more recently, as the University-wide Research Data Officer at Harvard University. I'm supportive of the intent of the new policy, but I would like to strongly suggest making the changes below:

- The policy focuses on sharing and managing scientific data. However, it is increasingly the case that the **code associated with the study** is necessary to enable reproducibility of published results. The policy should specifically require sharing all relevant code related to the scientific results, in addition to the data, in an archival repository.
- The statement "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public" is too soft. It makes it easy for researchers to choose not to share their data because, for example, they simply consider them not useful to the public. The language here should be stronger and **explicitly require** sharing of the data, or at least the metadata, except in some well-defined cases.
- Most of the legitimate difficulties in data-sharing relate to privacy concerns. For those cases, a **tiered access** to the data, together with privacy-preserving tools to query the

data, could resolve the majority of the concerns. For example, it could be required to **share a minimum set of descriptive metadata fields** to ensure that the dataset is findable and citable, and the research community knows that the dataset exist. The data could then be in a secure data enclave where only those with granted permissions could access. With the availability of new privacy-preserving tools (e.g., differential privacy), data aggregates or summaries with privacy guarantees could be released.

Thank you for considering these changes to the current draft.

Sincerely,

A handwritten signature in blue ink that reads "Mercè Crosas". The signature is written in a cursive style and is enclosed within a light blue oval shape.

Mercè Crosas, Ph.D.
University Research Data Officer
Chief Data Science and Technology Officer, Institute for Quantitative Social Science
Harvard University
<https://scholar.harvard.edu/mercecrosas>



Submission ID: 1443

Date: 1/11/2020

Name: Jennifer K. Wagner and Michelle N. Meyer

Name of Organization: Geisinger

Type of Data of Primary Interest: Genomic

Type of Data of Primary Interest - Other:

Type of Organization: Health Care Delivery Organization

Type of Organization - Other:

Role: Biothecist/Social Science Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Diverse data

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

1. We support the stated benefits of data sharing; however, it is disappointing that "maximizing research participants' contributions" is omitted from this section. This deserves highlighting here, just as it deserved mentioning in the background of the request for comments.

2. While we appreciate the need for this policy to have flexibility by design, the statement that "[s]hared data should be made accessible in a timely manner" requires further specification to be meaningful. A reasonable approach would be to require access (1) to all scientific data underlying any publications from the funded research at the time of publication and (2) to all remaining data by a specified period of time from the date the data management and sharing plan is approved.

3. Data management deserves additional attention in this section so that researchers appreciate the ongoing effort involved and diligent care required to do this effectively and efficiently. Similarly, some explanation of the benefits of prospectively planning for data preservation and sharing would be helpful. For instance, failure to anticipate data sharing at the time of participant consent and IRB review frequently leads to situations in which sharing is needlessly put into conflict with promises made to participants that data will not be shared

outside the study team or simply silence about sharing (see M.N. Meyer (2018), Practical Tips for Ethical Data Sharing, *Advances in Methods and Practices in Psychological Science*, 1(1), 131-144, available at <https://journals.sagepub.com/doi/pdf/10.1177/2515245917747656>).

4. In general, the Purpose section contains weak language that could be used to justify nonsharing, e.g., "NIH encourages data management and data sharing practices," "NIH expects researchers to . . . plan for which scientific data will be preserved and shared." The opening section of the Policy should establish a strong default that all scientific data will be preserved and shared.

Section II: Definitions:

1. FAIR data principles should be defined or definitions from elsewhere incorporated by reference so that all researchers have adequate notice of them.

2. The definition of "data management" should be revised to clarify the "integral role" of data management within the scientific process so that researchers are deterred from relegating data management issues and resources to the periphery. This revised definition should explicitly include "upstream management" issues that affect the quality of scientific data ultimately shared.

3. "Scientific software artifacts" should be defined.

4. As acknowledged by the NIH in the Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing, data sharing policies impose future burdens on researchers. This necessitates a definition for "covered period" to establish reasonable time limits on researcher or institution responsibilities.

Section III: Scope:

1. The policy should expressly indicate that data management and sharing obligations continue beyond the NIH funding period. A defined "covered period" would enable this.

2. Scientific software artifacts should be included as a scientific data asset.

Section IV: Effective Date(s):

1. The NIH should commit to a timeline for implementation following the issuance of the final policy.

Section V: Requirements:

1. The policy could be strengthened and anticipated compliance with the policy improved if data management and sharing plan templates would be provided for researchers. A reasonable approach would be to have a standardized template for particular funding mechanisms (e.g., R01, R21, R03) to make it easier for researchers when applying for funding and peer reviewers when evaluating funding applications.

2. A tiered approach to the data sharing requirement, such as that proposed by AMIA in 2018 (available at <https://www.amia.org/sites/default/files/AMIA-Response-to-Draft-DMSP-RFI.pdf>), would be useful.

3. Failing to share data responsibly is unethical, as it frustrates, undermines, and wastes the scientific contributions of voluntary research participants. It is important that the policy preclude researchers and institutions from using data privacy as subterfuge to evade scientific data sharing obligations.

Section VI: Data Management and Sharing Plans:

1. The data management and sharing plans should be required during the regular submission date as part of the application for research funding so that peer reviewers can evaluate the plans appropriately. This should not be merely a Just-In-Time requirement. For instance, data sharing plans can affect human subjects protections, which reviewers consider well in advance of the Just-In-Time period.

2. The data management and sharing plan should be a scorable part of the funding application (in recognition of its integral part to the scientific process) and made publicly available in NIH RePORTER.

3. The current draft policy states in relevant part, "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." It is unclear why NIH would fund the collection of scientific data that scientists know at the outset (i.e., at the time of submitting their Plan) will not be useful to other researchers or to the public. It is in fact exceedingly unlikely that data generated from NIH funds will have no utility beyond the researcher, but there is reason to believe that many researchers will latch onto this

language to avoid responsible data sharing. For instance, science has a history of deeming research that results in null effects as not "useful," leading to the well-known "file drawer problem," publication bias, and untold wasted funds and person-hours testing hypotheses that an unbiased scientific record would have suggested were poor bets. As recently as 2012, 45-50% of academic psychologists reported selectively reporting in their publications only studies that "worked" (L.K. John, G. Loewenstein & D. Prelec (2012), Measuring the Prevalence of Questionable Research Practices With Incentives for Truth-Telling, *Psychological Science* 23(5), 524–532, available at <https://www.cmu.edu/dietrich/sds/docs/loewenstein/MeasPrevalQuestTruthTelling.pdf>). If many scientists (and journals) continue to assume that the results of null effect studies are not worth publishing, it stands to reason that many will similarly assume, when invited by this language in the Policy, that the data underlying such studies is not worth sharing. Even assuming that NIH would fund the collection of unuseful scientific data, who determines data usefulness, how, and when exactly is this determination to be made? Research communities vary in norms and practices, so how does this diversity affect the NIH's expectation that data be shared? This sentence creates an unworkable and unnecessary standard. Instead, scientific data sharing should be a default requirement for receipt of NIH funding unless there are compelling justifications to the contrary, which will usually concern well-founded—not speculative or pretextual—legal or ethical concerns.

4. "NIH recognizes that certain factors (e.g., legal, ethical, technical) may limit the ability to preserve and share data. Plans should include consideration of these factors, when applicable, in describing the approach to data management and data sharing." The Policy should make clear that when researchers cite these factors to justify non-preservation and non-sharing, they must specify the precise concerns (not simply reference "participant privacy") and show that concerns are well-founded rather than speculative or pretextual and ICOs and other NIH staff reviewing Plans should be trained to hold Plans to this standard.

Section VII: Compliance and Enforcement:

1. The draft policy lacks sufficient detail for how and to whom one might report suspected noncompliance and what the process for determining noncompliance would entail.

2. Policy compliance should be integrated with the current annual progress reports, and it would be useful for the annual review forms. As with publications listed in the annual progress report form, the NIH could request a digital object identifier (DOI) or URL to a FAIR data file.

3. While negative incentives such as becoming ineligible for future funding are anticipated in this policy, particularly egregious violations might warrant the NIH to recover funding already received by a researcher or institution.

4. Positive incentives (such as requiring the data management and sharing plan as part of the scorable funding application) should be considered.

5. A process for NIH certification of data commons or repositories that are compliant with the policy would be helpful in promoting compliance.

6. The appropriate time for much data sharing will occur after the funding or support period. It is critical that NIH find ways after that period to maximize compliance with the Policy. Data from the field of psychological science about (non)compliance with journal and professional associated data sharing requirements are sobering. In an effort to obtain data for reanalysis, researchers e-mailed the corresponding authors of 141 articles published in American Psychology Association (APA) journals (J.M. Wicherts, D. Borsboom, J. Kats & D. Molenaar (2006), The Poor Availability of Psychological Research Data for Reanalysis, *American Psychologist* 61, 726–728). All authors who publish in these journals must sign the APA Certification of Compliance With APA Ethical Principles, Principle 8.14 of which requires that psychologists share data with other "competent professionals who seek to verify the substantive claims through reanalysis" (American Psychological Association (2003), Certification of Compliance with APA Ethical Principles, available at <https://www.apa.org/pubs/authors/ethics02.pdf>). Wicherts et al. sent more than 400 e-mails, often including detailed descriptions of their study's aims, IRB approvals, signed assurances not to share the data further, and their curricula vitae. Yet after 6 months, 73% of the authors had still failed to share their data. Most of those authors explicitly refused or said they were unable to share, whereas others promised to share but did not or simply never responded to the requests. Only 11% of the authors shared their data after the first request. To better enable NIH to take noncompliance into account in future funding decisions, and to incentivize compliance, NIH should consider requiring grant applicants with completed NIH awards governed by this Policy to submit a document in which they provide links to repositories where all scientific data for each is shared and provide any justifications for noncompliance.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

1. This guidance should be revised to clarify that data management and sharing costs (including both data preservation costs as well as personnel costs) that continue after the funding period are allowable and specify the period of time covered.

2. This guidance currently lacks details regarding how data management and sharing costs will affect funding decisions. There should be additional details about what levels of cost are acceptable and how that is determined, recognizing that this could significantly affect researcher behaviors not only in the design of the data management and sharing plans but also in their budget justifications for those plans.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

1. The current draft guidance specifies coverage of six elements (data type; related tools, software, and/or code; standards; data preservation, access, and associated timelines; data sharing agreements, licenses and other use limitations; and oversight of data management)"in two pages or less." This page limitation is arbitrary and lacks justification. The constraint would arguably punish those researchers who are exhibiting the characteristics of responsible, proactive data stewardship that this NIH data sharing policy is intended to promote. The two-page limitation is particularly problematic when one considers the importance of understanding the rationales underlying various decisions reflected in data management and sharing plans.

2. It would be useful to have further development of this guidance so researchers are equipped with information necessary to make responsible decisions involving data sharing trade-offs. This could include examples of trade-off decisions that researchers might commonly face and preferences or balancing criteria that the NIH would use in evaluating not only what decisions have been made but how those decisions have been made.

3. The draft guidance permits"an entry of 'to be determined'" if there is"justification provided along with a timeline or appropriate milestone at which a determination will be made." The policy needs additional details articulating what types of justifications would be sufficiently compelling, and this should be rare.

Other Considerations Relevant to this DRAFT Policy Proposal:

N/A.

We appreciate the opportunity to provide this input and look forward to further development of this policy and supplemental guidances.

Attachment:

JKW MNM Comments 2020.01.10 NIH Draft Data Mgmt Sharing Policy Final.pdf

Jennifer K. Wagner, J.D., Ph.D.

Associate Director, Bioethics Research
Assistant Professor, Center for Translational Bioethics & Health Care Policy
Geisinger Health System, 100 North Academy Avenue, MC 30-42, Danville, PA 17822
570.214.3774 | jwagner1@geisinger.edu

Michelle N. Meyer, Ph.D., J.D.

Associate Director, Research Ethics
Assistant Professor, Center for Translational Bioethics & Health Care Policy
Faculty Co-Director, Behavioral Insights Team, Steele Institute for Health Innovation
Geisinger Health System, 100 North Academy Avenue, MC 30-42, Danville, PA 17822
570.214.3380 | mmeyer@geisinger.edu

January 10, 2020

Dr. Andrea Jackson-Dipina
Director of the Division of Scientific Data Sharing Policy
National Institutes of Health
Office of Science Policy
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892
SciencePolicy@mail.nih.gov

Re: Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance, 84 FR 60398-60402 (Nov. 8, 2019)

Dear Director:

The following comments are provided in response to the DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance and informed by those submitted by Michael Hoffman and John Wilbanks. These comments are our own and are not attributable to Geisinger. Although Dr. Meyer is a member of the National Library of Medicine-sponsored National Academies Study Committee on Forecasting Costs for Preserving and Promoting Access to Biomedical Data (which NIH has referenced in webinars and materials associated with this draft policy), she is writing in her individual capacity and these comments are not attributable to, and should not be assumed to reflect the views of, the Committee or other of its members. As per the instructions, in addition to submitting this letter, we have uploaded specific comments into the appropriate fields using the online form at <https://osp.od.nih.gov/draft-data-sharing-and-management/>.

DRAFT NIH Policy for Management and Sharing***Section I: Purpose***

1. We support the stated benefits of data sharing; however, it is disappointing that “maximizing research participants’ contributions” is omitted from this section. This deserves highlighting here, just as it deserved mentioning in the background of the request for comments.

2. While we appreciate the need for this policy to have flexibility by design, the statement that “[s]hared data should be made accessible in a timely manner” requires further specification to be meaningful. A

Jennifer K. Wagner, J.D., Ph.D. and Michelle N. Meyer, Ph.D., J.D.

reasonable approach would be to require access (1) to all scientific data underlying any publications from the funded research at the time of publication and (2) to all remaining data by a specified period of time from the date the data management and sharing plan is approved.

3. Data management deserves additional attention in this section so that researchers appreciate the ongoing effort involved and diligent care required to do this effectively and efficiently. Similarly, some explanation of the benefits of prospectively planning for data preservation and sharing would be helpful. For instance, failure to anticipate data sharing at the time of participant consent and IRB review frequently leads to situations in which sharing is needlessly put into conflict with promises made to participants that data will not be shared outside the study team or simply silence about sharing (see M.N. Meyer (2018), Practical Tips for Ethical Data Sharing, *Advances in Methods and Practices in Psychological Science*, 1(1), 131-144, available at <https://journals.sagepub.com/doi/pdf/10.1177/2515245917747656>).

4. In general, the Purpose section contains weak language that could be used to justify nonsharing, e.g., “NIH encourages data management and data sharing practices,” “NIH expects researchers to . . . plan for which scientific data will be preserved and shared.” The opening section of the Policy should establish a strong default that all scientific data will be preserved and shared.

Section II: Definitions

1. FAIR data principles should be defined or definitions from elsewhere incorporated by reference so that all researchers have adequate notice of them.

2. The definition of “data management” should be revised to clarify the “integral role” of data management within the scientific process so that researchers are deterred from relegating data management issues and resources to the periphery. This revised definition should explicitly include “upstream management” issues that affect the quality of scientific data ultimately shared.

3. “Scientific software artifacts” should be defined.

4. As acknowledged by the NIH in the Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing, data sharing policies impose future burdens on researchers. This necessitates a definition for “covered period” to establish reasonable time limits on researcher or institution responsibilities.

Section III: Scope

1. The policy should expressly indicate that data management and sharing obligations continue beyond the NIH funding period. A defined “covered period” would enable this.

2. Scientific software artifacts should be included as a scientific data asset.

Section IV: Effective Date(s)

1. The NIH should commit to a timeline for implementation following the issuance of the final policy.

Section V: Requirements

1. The policy could be strengthened and anticipated compliance with the policy improved if data management and sharing plan templates would be provided for researchers. A reasonable approach would be to have a standardized template for particular funding mechanisms (e.g., R01, R21, R03) to make it

Jennifer K. Wagner, J.D., Ph.D. and Michelle N. Meyer, Ph.D., J.D.

easier for researchers when applying for funding and peer reviewers when evaluating funding applications.

2. A tiered approach to the data sharing requirement, such as that proposed by AMIA in 2018 (available at <https://www.amia.org/sites/default/files/AMIA-Response-to-Draft-DMSP-RFI.pdf>), would be useful.

3. Failing to share data responsibly is unethical, as it frustrates, undermines, and wastes the scientific contributions of voluntary research participants. It is important that the policy preclude researchers and institutions from using data privacy as subterfuge to evade scientific data sharing obligations.

Section VI: Data Management and Sharing Plans

1. The data management and sharing plans should be required during the regular submission date as part of the application for research funding so that peer reviewers can evaluate the plans appropriately. This should not be merely a Just-In-Time requirement. For instance, data sharing plans can affect human subjects protections, which reviewers consider well in advance of the Just-In-Time period.

2. The data management and sharing plan should be a scorable part of the funding application (in recognition of its integral part to the scientific process) and made publicly available in NIH RePORTER.

3. The current draft policy states in relevant part, “NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public.” It is unclear why NIH would fund the collection of scientific data that scientists know at the outset (i.e., at the time of submitting their Plan) will not be useful to other researchers or to the public. It is in fact exceedingly unlikely that data generated from NIH funds will have no utility beyond the researcher, but there is reason to believe that many researchers will latch onto this language to avoid responsible data sharing. For instance, science has a history of deeming research that results in null effects as not “useful,” leading to the well-known “file drawer problem,” publication bias, and untold wasted funds and person-hours testing hypotheses that an unbiased scientific record would have suggested were poor bets. As recently as 2012, 45-50% of academic psychologists reported selectively reporting in their publications only studies that “worked” (L.K. John, G. Loewenstein & D. Prelec (2012), Measuring the Prevalence of Questionable Research Practices With Incentives for Truth-Telling, *Psychological Science* 23(5), 524–532, available at <https://www.cmu.edu/dietrich/sds/docs/loewenstein/MeasPrevalQuestTruthTelling.pdf>). If many scientists (and journals) continue to assume that the results of null effect studies are not worth publishing, it stands to reason that many will similarly assume, when invited by this language in the Policy, that the data underlying such studies is not worth sharing. Even assuming that NIH would fund the collection of unuseful scientific data, who determines data usefulness, how, and when exactly is this determination to be made? Research communities vary in norms and practices, so how does this diversity affect the NIH’s expectation that data be shared? This sentence creates an unworkable and unnecessary standard. Instead, scientific data sharing should be a default requirement for receipt of NIH funding unless there are compelling justifications to the contrary, which will usually concern well-founded—not speculative or pretextual—legal or ethical concerns.

4. “NIH recognizes that certain factors (e.g., legal, ethical, technical) may limit the ability to preserve and share data. Plans should include consideration of these factors, when applicable, in describing the approach to data management and data sharing.” The Policy should make clear that when researchers cite these factors to justify non-preservation and non-sharing, they must specify the precise concerns (not simply reference “participant privacy”) and show that concerns are well-founded rather than speculative or pretextual and ICOs and other NIH staff reviewing Plans should be trained to hold Plans to this standard.

Jennifer K. Wagner, J.D., Ph.D. and Michelle N. Meyer, Ph.D., J.D.

Section VII: Compliance and Enforcement

1. The draft policy lacks sufficient detail for how and to whom one might report suspected noncompliance and what the process for determining noncompliance would entail.
2. Policy compliance should be integrated with the current annual progress reports, and it would be useful for the annual review forms. As with publications listed in the annual progress report form, the NIH could request a digital object identifier (DOI) or URL to a FAIR data file.
3. While negative incentives such as becoming ineligible for future funding are anticipated in this policy, particularly egregious violations might warrant the NIH to recover funding already received by a researcher or institution.
4. Positive incentives (such as requiring the data management and sharing plan as part of the scorable funding application) should be considered.
5. A process for NIH certification of data commons or repositories that are compliant with the policy would be helpful in promoting compliance.
6. The appropriate time for much data sharing will occur after the funding or support period. It is critical that NIH find ways after that period to maximize compliance with the Policy. Data from the field of psychological science about (non)compliance with journal and professional associated data sharing requirements are sobering. In an effort to obtain data for reanalysis, researchers e-mailed the corresponding authors of 141 articles published in American Psychological Association (APA) journals (J.M. Wicherts, D. Borsboom, J. Kats & D. Molenaar (2006), The Poor Availability of Psychological Research Data for Reanalysis, *American Psychologist* 61, 726–728). All authors who publish in these journals must sign the APA Certification of Compliance With APA Ethical Principles, Principle 8.14 of which requires that psychologists share data with other “competent professionals who seek to verify the substantive claims through reanalysis” (American Psychological Association (2003), Certification of Compliance with APA Ethical Principles, available at <https://www.apa.org/pubs/authors/ethics02.pdf>). Wicherts et al. sent more than 400 e-mails, often including detailed descriptions of their study’s aims, IRB approvals, signed assurances not to share the data further, and their curricula vitae. Yet after 6 months, 73% of the authors had still failed to share their data. Most of those authors explicitly refused or said they were unable to share, whereas others promised to share but did not or simply never responded to the requests. Only 11% of the authors shared their data after the first request. To better enable NIH to take noncompliance into account in future funding decisions, and to incentivize compliance, NIH should consider requiring grant applicants with completed NIH awards governed by this Policy to submit a document in which they provide links to repositories where all scientific data for each is shared and provide any justifications for noncompliance.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

1. This guidance should be revised to clarify that data management and sharing costs (including both data preservation costs as well as personnel costs) that continue after the funding period are allowable and specify the period of time covered.
2. This guidance currently lacks details regarding how data management and sharing costs will affect funding decisions. There should be additional details about what levels of cost are acceptable and how

Jennifer K. Wagner, J.D., Ph.D. and Michelle N. Meyer, Ph.D., J.D.

that is determined, recognizing that this could significantly affect researcher behaviors not only in the design of the data management and sharing plans but also in their budget justifications for those plans.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan

1. The current draft guidance specifies coverage of six elements (data type; related tools, software, and/or code; standards; data preservation, access, and associated timelines; data sharing agreements, licenses and other use limitations; and oversight of data management) “in two pages or less.” This page limitation is arbitrary and lacks justification. The constraint would arguably punish those researchers who are exhibiting the characteristics of responsible, proactive data stewardship that this NIH data sharing policy is intended to promote. The two-page limitation is particularly problematic when one considers the importance of understanding the rationales underlying various decisions reflected in data management and sharing plans.

2. It would be useful to have further development of this guidance so researchers are equipped with information necessary to make responsible decisions involving data sharing trade-offs. This could include examples of trade-off decisions that researchers might commonly face and preferences or balancing criteria that the NIH would use in evaluating not only what decisions have been made but how those decisions have been made.

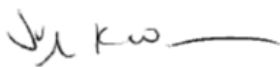
3. The draft guidance permits “an entry of ‘to be determined’” if there is “justification provided along with a timeline or appropriate milestone at which a determination will be made.” The policy needs additional details articulating what types of justifications would be sufficiently compelling, and this should be rare.

Other Relevant Considerations to this DRAFT Policy Proposal

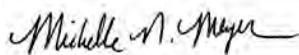
N/A.

We appreciate the opportunity to provide this input and look forward to further development of this policy and supplemental guidances.

Best regards,



Jennifer K. Wagner, J.D., Ph.D.



Michelle N. Meyer, Ph.D., J.D.

Submission ID: 1444

Date: 1/11/2020

Name: Ruth O'Hara, PhD

Name of Organization: Stanford University

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: All data listed above

Type of Organization: University

Type of Organization - Other:

Role: Other

Role - Other:

Domain of Research Most Important to You or Your Organization:

I am the Senior Associate Dean for Research in the Stanford University School of Medicine and all the types of data listed above come under my oversight.

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

RFI NIH 01102020.doc

Description:

Response to RFI from Stanford School of Medicine Senior Associate Dean of Research

Jan 10, 2020

To whom it may concern

Stanford leadership very much appreciates the opportunity to respond to NOT-OD-19-014 "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research.

We are very much aligned with the response to your RFI, dated January 10th, 2020, provided by the Council on Governmental Relations (COGR).

Specifically, we support making data from federally-funded research accessible both to the public and others in the research community to accelerate scientific discovery and by making data more open to scrutiny and re-analysis. We fully support the submission of data management and sharing plans at JIIT with programmatic review, rather than the data management and sharing plans being considered as part of the overall impact score for extramural support. We are in agreement with the COGR recommendation that NIH allow additional data management costs to be added to the budget at JIIT based on the final negotiated data management and sharing plan. We are aligned with all other COGR recommendations in response to this RFI.

As the field of data management and sharing is rapidly evolving, and the nature of data is highly heterogeneous, we would also welcome opportunities to partner with NIH and other relevant stakeholders to develop resources and tools to better facilitate data management and sharing.

Please feel free to contact us if you have any questions.

Sincerely,



Ruth O'Hara, Ph.D.
Senior Associate Dean for Research,
Stanford University School of Medicine,
Professor,
Department of Psychiatry and Behavioral Sciences,
Stanford University School of Medicine,
Phone: (650) 493-5000 ex.63620
e-mail: roh@stanford.edu

Submission ID: 1445

Date: 1/11/2020

Name: Felice J Levine

Name of Organization: American Educational Research Association

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: behavioral, social, and education science data

Type of Organization: Professional Org/Association

Type of Organization - Other:

Role: Institutional Official

Role - Other:

Domain of Research Most Important to You or Your Organization:

education, health, and wellbeing; developmental cognitive, learning, and social emotional process

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

I. Purpose: AERA strongly supports the purpose of this policy. We particular applaud the purpose of sharing data not just to test the validity of research findings but also to stimulate analyses of hard-to-generate data and linked data as well as examining new questions or hypotheses at the "frontiers of discovery."

Section II: Definitions:

II. Definitions: We have no recommendations for this section. We concur with the definitions as set forth.

Section III: Scope:

III. Scope: AERA agrees that the draft NIH policy should apply to all research funded by the agency. We believe data sharing consonant with human subjects protection is a critical dimension of sound science, whether or not NIH funded. We recognize that the scope of NIH policy reaches only to research it supports; we support data sharing applying to all such NIH-funded research.

Section IV: Effective Date(s):

IV. Effective Dates: While understanding that the effective date will depend on the final release of the guidance, we would encourage NIH to make a final policy effective no later than a year after it is issued. This would allow time for NIH to produce any additional supplemental materials and templates to facilitate data preservation and sharing, and for institutions to support their faculty and research teams in the application process.

Data sharing is not a new concept or topic. Federal science agencies were encouraged as far back as 1985 to support and encourage data sharing in the National Research Council Panel Report on Sharing Research Data. Therefore, NIH should move ahead and consider implementation of a final policy within a year after adoption.

AERA, among others in the education and learning sciences, are working to promote a culture of research transparency, of which data sharing is a major component. We are also aware of general concerns from the field about data sharing, specifically regarding certain forms of qualitative data and involving vulnerable populations. NIH can provide training on best practices through its responsible conduct of research training and in webinars prior to the effective date that can address questions and concerns from NIH grant applicants across methodologies and forms of data.

Section V: Requirements:

V. Requirements: The requirements set forth in this section are acceptable especially when the Supplemental Draft Guidance is considered as part of the Requirements. Some of the information in the Supplemental Material might be incorporated more explicitly under Requirements. Sharing data in restricted access form, for example, is very important but resides only in the Supplemental material. The Supplemental material also makes reference to making data available as "soon as practicable," but without any general requirement as to what that might look like.

Section VI: Data Management and Sharing Plans:

VI. Data Management and Sharing Plans: We encourage NIH to consider altering the sentence, "NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public." We agree that there should not be a set time limit for preserving data and strongly support the fact that there is no language to require destruction of data after a specific length of time. However, the statement as currently written could lead NIH-supported researchers to only preserve their data for a short amount of time, limiting the use of such data for replication and reproducibility studies or to forge new research directions using extant data.

AERA also appreciates aligning any limitations on sharing data with privacy and confidentiality reasons and with approvals from Institutional Review Boards, in addition to legal and ethical factors. We also emphasize, however, that concerns about privacy and confidentiality can often be addressed through data management plans and mechanisms to make data available in restricted form with safeguards against inadvertent disclosure.

Section VII: Compliance and Enforcement:

VII. Compliance and Enforcement: AERA supports the inclusion of a term and condition on the overall commitment to preserve and share data in the Notice of Award. We ask NIH to consider providing flexibility in case there are changes to the specifics of a data sharing and management plan (e.g., using a different repository than indicated in the grant proposal).

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

AERA strongly supports this draft guidance. Early career scholars and researchers at institutions that may not have the necessary infrastructure in place for data sharing have indicated costs as a potential barrier to sharing data. Allowing grant applicants to include reasonable costs for preparing their data to share and to cover expenses associated with storing data in trusted repositories would encourage data sharing.

In addition to including proposed costs for data sharing in grant applications, AERA also recommends encouraging grantees to report their actual costs. Data on the necessary costs for data management and sharing can be helpful to the broader scientific community in the initial planning of a research project.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

In general, AERA is pleased that NIH is encouraging potential grantees to describe the data they expect to collect through the course of a research grant, along with the software used to analyze data, a description on how data will be archived, code, and potential limitations to sharing data. These elements are important for NIH grantees to consider at the outset of their work and also will help support replicability and reproducibility of NIH-funded research.

Other Considerations Relevant to this DRAFT Policy Proposal:

AERA offers a couple of items for consideration in the section on Data Preservation, Access, and Associated Timelines as NIH continues developing this guidance:

- We recommend amending the first bullet point to explicitly encourage grant applicants to include the repository(ies) they plan to use for data preservation and to state how they will store and manage data if they will not use an existing repository. Repositories are important in safely securing data and are critical to the preservation, management, and discovery of data.

- In the last bullet point, we recommend encouraging study preregistration as a mechanism to share aspects of data analysis or to note the types of data that are being collected when data are not ready to be shared from a funded grant.

Attachment:

AERA Comments - Draft NIH Data Management and Sharing Policy_1-10-20_FINAL.pdf

Description:

AERA PDF Comment Letter



January 10, 2020

Dr. Carrie D. Wolinetz
 Acting Chief of Staff & Associate Director for Science Policy
 National Institutes of Health
 6705 Rockledge Drive, Suite 750
 Bethesda, MD 20892

Re: Draft NIH Policy for Data Management and Sharing

Dear Dr. Wolinetz,

On behalf of the American Educational Research Association (AERA), I want to thank you for the opportunity to comment on the draft NIH Policy for Data Management and Sharing and the accompanying supplemental guidance.

AERA is the major national scientific association of 25,000 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. Many of our members receive funding from the National Institutes of Health (NIH) for fundamental research in areas such as understanding learning processes and the intersection of health and education outcomes. AERA has long been committed to data sharing as set forth in the *AERA Code of Ethics* (revised 2011) and in the *Standards for Reporting on Empirical Social Science Research in AERA Publications* (2006).

We commend NIH for developing this draft plan as part of the agency's commitment to open science and as a steward of the federal investment in scientific research. NIH has been at the forefront of promoting the sharing and use of scientific data, and we appreciate the work that NIH is undertaking to continue building a culture of data sharing consistent with the FAIR (findable, accessible, interoperable, and reusable) principles.

The following responses were also submitted to the corresponding fields on the [RFI website](#):

Draft NIH Policy for Data Management and Sharing

I. Purpose: AERA strongly supports the purpose of this policy. We particular applaud the purpose of sharing data not just to test the validity of research findings but also to stimulate analyses of hard-to-generate data and linked data as well as examining new questions or hypotheses at the "frontiers of discovery."

1430 K Street, NW • Washington, DC 20005 • (202) 238-3200

Facsimile (202) 238-3250 • <http://www.aera.net>

II. Definitions: We have no recommendations for this section. We concur with the definitions as set forth.

III. Scope: AERA agrees that the draft NIH policy should apply to all research funded by the agency. We believe data sharing consonant with human subjects protection is a critical dimension of sound science, whether or not NIH funded. We recognize that the scope of NIH policy reaches only to research it supports; we support data sharing applying to all such NIH-funded research.

IV. Effective Dates: While understanding that the effective date will depend on the final release of the guidance, we would encourage NIH to make a final policy effective no later than a year after it is issued. This would allow time for NIH to produce any additional supplemental materials and templates to facilitate data preservation and sharing, and for institutions to support their faculty and research teams in the application process.

Data sharing is not a new concept or topic. Federal science agencies were encouraged as far back as 1985 to support and encourage data sharing in the National Research Council Panel Report on *Sharing Research Data*. Therefore, NIH should move ahead and consider implementation of a final policy within a year after adoption.

AERA, among others in the education and learning sciences, are working to promote a culture of research transparency, of which data sharing is a major component. We are also aware of general concerns from the field about data sharing, specifically regarding certain forms of qualitative data and involving vulnerable populations. NIH can provide training on best practices through its responsible conduct of research training and in webinars prior to the effective date that can address questions and concerns from NIH grant applicants across methodologies and forms of data.

V. Requirements: The requirements set forth in this section are acceptable especially when the Supplemental Draft Guidance is considered as part of the Requirements. Some of the information in the Supplemental Material might be incorporated more explicitly under Requirements. Sharing data in restricted access form, for example, is very important but resides only in the Supplemental material. The Supplemental material also makes reference to making data available as “soon as practicable,” but without any general requirement as to what that might look like.

VI. Data Management and Sharing Plans: We encourage NIH to consider altering the sentence, “NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public.” We agree that there should not be a set time limit for preserving data and strongly support the fact that there is no language to require destruction of data after a specific length of time. However, the statement as currently written could lead NIH-supported researchers to only preserve their data for a short amount of time, limiting the use

of such data for replication and reproducibility studies or to forge new research directions using extant data.

AERA also appreciates aligning any limitations on sharing data with privacy and confidentiality reasons and with approvals from Institutional Review Boards, in addition to legal and ethical factors. We also emphasize, however, that concerns about privacy and confidentiality can often be addressed through data management plans and mechanisms to make data available in restricted form with safeguards against inadvertent disclosure.

VII. Compliance and Enforcement: AERA supports the inclusion of a term and condition on the overall commitment to preserve and share data in the Notice of Award. We ask NIH to consider providing flexibility in case there are changes to the specifics of a data sharing and management plan (e.g., using a different repository than indicated in the grant proposal).

Supplemental Draft Guidance: Allowable Costs for Data Management and Sharing

AERA strongly supports this draft guidance. Early career scholars and researchers at institutions that may not have the necessary infrastructure in place for data sharing have indicated costs as a potential barrier to sharing data. Allowing grant applicants to include reasonable costs for preparing their data to share and to cover expenses associated with storing data in trusted repositories would encourage data sharing.

In addition to including proposed costs for data sharing in grant applications, AERA also recommends encouraging grantees to report their actual costs. Data on the necessary costs for data management and sharing can be helpful to the broader scientific community in the initial planning of a research project.

Supplemental Draft Guidance: Elements of a NIH Data Management and Sharing Plan

In general, AERA is pleased that NIH is encouraging potential grantees to describe the data they expect to collect through the course of a research grant, along with the software used to analyze data, a description on how data will be archived, code, and potential limitations to sharing data. These elements are important for NIH grantees to consider at the outset of their work and also will help support replicability and reproducibility of NIH-funded research.

AERA offers a couple of items for consideration in the section on Data Preservation, Access, and Associated Timelines as NIH continues developing this guidance:

- We recommend amending the first bullet point to explicitly encourage grant applicants to include the repository(ies) they plan to use for data preservation and to state how they will store and manage data if they will not use an existing repository. Repositories are important in safely securing data and are critical to the preservation, management, and discovery of data.

- In the last bullet point, we recommend encouraging study preregistration as a mechanism to share aspects of data analysis or to note the types of data that are being collected when data are not ready to be shared from a funded grant.

Thank you once again for the opportunity to comment. Please do not hesitate to call upon AERA if we can be helpful in the further development of this important policy and guidance.

Sincerely,

A handwritten signature in black ink, appearing to read "Felice J. Levine". The signature is fluid and cursive, with the first name being the most prominent.

Felice J. Levine, PhD
Executive Director
flevine@aera.net
202-238-3201

Submission ID: 1446

Date: 1/11/2020

Name: James Love

Name of Organization: KEI

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Economic

Type of Organization: Patient Advocacy Organization

Type of Organization - Other:

Role: Biothecist/Social Science Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Costs of clinical trials

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

KEI Comments Regarding DRAFT NIH Policy for Data Management and Sharing.pdf

Description:

KEI comments on economic data on R&D costs



1621 Connecticut Avenue NW
Suite 500
Washington, DC 20009
www.keionline.org

January 10, 2020

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive
Suite 750
Bethesda, MD 20892
Via Online Submission

Re: Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Knowledge Ecology International (KEI) welcomes the opportunity to comment on the DRAFT National Institutes of Health (NIH) Policy for Data Management and Sharing.

KEI is a nonprofit non-governmental organization with offices in Washington, DC and Geneva, Switzerland that focuses on improving the management of knowledge resources in search of better outcomes in the public interest. KEI's primary areas of work are policies concerning equitable access to knowledge and medicines in the US and globally. Through these efforts, KEI has sought to promote open source data and the sharing of information to promote increased innovation and improved outcomes for patients and consumers. As a part of our work in increasing affordable, sustainable access to medicines, KEI has advocated for increased transparency of data in the area of medical R&D costs, in order to better inform future research and policies to ensure the public has access to the life-saving treatments they need.

While primarily concerning the data management and sharing of scientific data generated from NIH-funded research, KEI would like to offer comments regarding improved reporting and sharing of economic and cost data associated with US taxpayer-funded research.

As noted in section *I. Purpose* of the DRAFT NIH Policy for Data Management and Sharing ("Draft Policy"), the NIH has a "longstanding commitment to making the results and outputs of the research that it funds and conducts available to the public." One such output is data concerning the costs of conducting research, including human subject clinical trials, and the costs of manufacturing of medical technologies.

As a part of the submission of the required Data Management and Sharing Plan as outlined in the Draft Policy, researchers with NIH-funded or conducted research projects should explain the expected costs to be incurred in the course of their project, including in particular, the expected costs of clinical trials, and the sharing of that cost among various sources of funding, including the NIH grants or contracts, funds from other federal and other government agencies, private charities, industry and reimbursements from government or private health insurance plans.

In the *Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan* section, KEI would recommend that the Elements of a Plan should include (in addition to those already proposed in the DRAFT Guidance) data related to the costs of the research project such as:

1. Expected/Actual Total Cost of the Research Project;
2. NIH Grant Numbers associated with the Research Project (if applicable); and
3. Any Clinical Trials conducted in connection with the research.

If there are clinical trials associated with research for which the Data Management and Sharing Plan is submitted, KEI recommends the following data regarding the trial be included in the Plan:

1. NCT Number
2. Trial Sponsor
3. Phase
4. Enrollment
5. Trial Start/End Dates
6. Expected/Actual Cost of the Clinical Trial
7. Expected/Actual Per Patient Cost of the Clinical Trial
8. The expected contribution to the trial cost by:
 - a. the NIH grant or contract,
 - b. any other federal agency,
 - c. any other non-federal government agency,
 - d. any charities,
 - e. Industry,
 - f. And health plans that provide reimbursements of trial related expenses, including those required to do so under PHS Act section 2709(a).[1]

It will also be helpful to have the collection and sharing of data on manufacturing costs for medical technologies.

The information will contribute directly to policy researchers exploring how to design or reform incentives for biomedical research and development, in the evaluation the reasonableness of prices or incentives, and also give researchers, academic, non-profit and industry, useful information on the budgets required to bring inventions to practical application.

We would be happy to further discuss the suggestions and answer any questions related to these issues. Thank you for the opportunity to offer comments on the NIH Policy for Datamanagement and Sharing.

James Love
Knowledge Ecology International
1621 Connecticut Avenue, Suite 500
Washington, DC 20009
<https://keionline.org>
james.love@keionline.org

Footnote:

[1]. “PHS Act section 2709(a), as added by the Affordable Care Act, states that if a group health plan or health insurance issuer in the group and individual health insurance market provides coverage to a qualified individual (as defined under PHS Act section 2709(b)), then such plan or issuer: (1) may not deny the qualified individual participation in an approved clinical trial with respect to the treatment of cancer or another life-threatening disease or condition; (2) may not deny (or limit or impose additional conditions on) the coverage of routine patient costs for items and services furnished in connection with participation in the trial; and (3) may not discriminate against the individual on the basis of the individual’s participation in the trial.

A qualified individual under PHS Act section 2709(b) is generally a participant or beneficiary who is eligible to participate in an approved clinical trial according to the trial protocol with respect to the treatment of cancer or another life-threatening disease or condition; and either: (1) the referring health care professional is a participating provider and has concluded that the individual’s participation in such trial would be appropriate; or (2) the participant or beneficiary provides medical and scientific information establishing that the individual’s participation in such trial would be appropriate.”

https://www.cms.gov/CCIIO/Resources/Fact-Sheets-and-FAQs/aca_implementation_faqs15 ;
see also: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4876354/>

Submission ID: 1447

Date: 1/11/2020

Name: Melissa Haendel and Julie McMurry

Name of Organization: Monarch Initiative

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Translational

Type of Organization: University

Type of Organization - Other:

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

rare diseases, genotype-phenotype data, data integration, semantic engineering, science of science, public health

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

While the proposed policy is an improvement over baseline, it still lacks the specificity and enforceability that would ensure that resources are not only shared but also useful. For example, while the policy includes requirement of sharing plans as part of JIT, there is no description as to the process for their review and approval. Further, there really is not much guidance into what constitutes the most critical elements of quality resource sharing and reuse, which should be the attributes against which the Plan will be evaluated.

The proposed policy attempts to describe the importance of resource sharing to society; it might be more compelling if it drew more from the experiences garnered from numerous previous large NIH collaborative data sharing programs or from community recommendations such as from the Biden Moonshot Blue Ribbon panel on data sharing. What is needed isn't more box checking, but rather to fundamentally change the research landscape, elevating the practice of resource sharing to its rightful place alongside the usual suspects of hypotheses, reagents, and results. The overall shared goal will be to create a network effect - or the "data ecosystem" as described in the blue ribbon recommendations.

Moonshot data sharing policy:

<https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/funding/public-access-policy>

Moonshot blue ribbon panel recommendations: <https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/enhanced-data-sharing-working-group-report.pdf>

Section II: Definitions:

Enhance the definition of "Data Management." As described above, this definition does not include the fact that data management must be an iterative and integral part of ongoing research requiring specialized skills to ensure that the quality of shared data is fundamentally useful and reusable where possible.

Enhance the definition of "Data Sharing". Specifics should be included such as access mechanisms (API, data downloads, etc.), archival plans and persistence (DOIs or other persistent identifiers), standards and data harmonization (checklists, models, ontologies), and licensing (use rights, flow-through terms).

Add specific definitions for standards and data and resource licensing.

Add definition for software, standards, clinical instruments, and other resource types. The "Plan" should be a data and resource sharing plan.

Section III: Scope:

Scope should make clear that the policy continues to apply for scientific resources produced by funding in whole or in part from NIH after the NIH funding period is over.

Scope should make clear that these requirements apply not just to research project grants and contracts, but most other forms of requests for support that will lead to the creation of scientific data and resources. This includes cooperative agreements, career grants, fellowships, scholarships, and training grants.

Scope should make it clear that these requirements apply not just to primary data but also to derived data and to other products of research including software, standards, protocols, reagents, devices etc. Each of these should be defined in the definitions. Further, it should be noted in the Plan how such things will interact - for example, often specific data types may be inaccessible without specific software tools.

Scope should include a statement about human subjects data sharing. While obviously human subjects privacy rules and ethics must be adhered to, this needs to become less of a barrier to data sharing properly. Improved recommendations and processes for sharing human subjects data must be provided and incentivized by NIH. Further, there is a fundamental right of patients to share their data with proper informed consent. Within the scope, it must be clear that harm to patients can come from both sharing and not sharing their data and proper sharing and risk mitigation must both be taken into account.

Section IV: Effective Date(s):

NIH might consider a phased rollout:

Month 2: Resource Sharing Plans are strongly advised for all new submissions

Month 6: Resource Sharing Plans are required for all new submissions

Month 12: Resource Sharing Plans are incorporated into scored review components for all new submissions. Instructions to reviewers are updated to reflect this.

In time, it would be optimal to transition from a template to some kind of formal registration of proposed products; structured representations would improve not only enforceability but also potentially discoverability.

Regardless of the timeline, the specific roll out plans and responsibilities must be clearly detailed to ensure quality compliance while retaining good will.

Section V: Requirements:

It is not entirely clear what the requirements section is aiming to address (that all grants must have one? We suggest the following two aspects be considered:

The data sharing plan MUST be part of the review criteria. Review panels must include data science expertise in standards/reusability/dissemination/archiving. Further, review and

endorsement of the plan during JIT should be necessarily be performed by experts in data sharing, data standardization, and data archiving.

We applaud the notion that "The Plan will become a Term and Condition of the Notice of Award." However, this aforementioned expert review process should similarly occur during grant reporting periods in order to determine adherence to the plan and make recommendations for updates to the plan as the work proceeds. There is not enough detail about what the process will be for determining whether or not the Terms and Conditions have been met.

Section VI: Data Management and Sharing Plans:

We advise that the policy specifically mandate that Data and Resource sharing plans:

Be required for all proposals (regardless of amount or nature of the award);

Be incorporated into the review (Guidance to reviewers on how to score review criteria such as significance and approach should include review of the Data Management and Sharing Plan);

Include descriptions of any genuine ethical or legal constraints that would preclude or limit the sharing of data, if applicable;

Have required sections corresponding to all expected products (see sections below)

Require applicants to describe their proposed use of existing resources, standards, etc.

We have published in Zenodo (<https://doi.org/10.5281/zenodo.3604521>) our most recently submitted sharing plan in the hopes that it could serve as an example of a format that could prompt applicants to be more strategic and thoughtful about the downstream use of their work.

We also advise that all progress reports for funded projects include a description of progress toward the sharing plans described in the proposal.

The plan should also delineate the different rationales for sharing data. Some data is designed to be directly reusable in downstream applications, whereas other data is shared to promote reproducibility and transparency. These two rationales exist in a spectrum. However, the process used to make data reusable in downstream applications differs from simpler

reproducibility endeavors - a more robust curation, metadata, quality assurance, identifier, and persistence strategy needs to be employed in addition to methodological documentation and cost management for storage or processing.

Finally, there does not seem to be a recognition regarding the need to record provenance or attribution of the data or other artifacts. It is critically important to know how the data was created and where it came from in the context of data reuse. Further, the use of credit systems such as the newly developed Contribution Role Ontology and Contributor Attribution Model can incentivize sharing and recognize team contributions in addition to adding to the provenance. See <https://contributor-attribution-model.readthedocs.io/en/latest/>

Section VII: Compliance and Enforcement:

As currently written, the compliance and enforcement does not appear to be much stronger than the status quo. As described above, there should be an opportunity for investigators at the time of JIT to improve and finalize their Data and Resource Sharing plan in response to expert reviews. The revised Plans should be made public alongside abstracts, with contact information regarding what to do to report violations or to simply make suggestions. Often times one community is not aware of standardized best practices in another. Therefore a dedicated data sharing navigator may be warranted to help investigators, especially in the early phases of this policy rollout. Finally, the consequences of a serious violation need to be made clear and should be equally applied regardless of stature or institution. The sharing of taxpayer-funded data and resources should be mandatory and the consequences for not doing so should be as significant as data fabrication or other scientific misconduct. Similar to not releasing funds until public access policy has been demonstrated, so should adherence to resource sharing plans.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Quality preparation of data for actual reuse takes time and money. Researchers should not only have to justify the process and expense, but the budget and / or budget justification should have a section identifying what requested resources would be utilized for this purpose. Just as they would with other budget categories, reviewers should consider whether proposed allowances are appropriate and adequate to ensure compliance with the policy. The type of expertise required to curate, provision quality metadata, harmonize data or otherwise standardize and disseminate data is its own specialized skill. Given that most researchers do not recognize this a priori, it should be a recommendation that funds be included for such activities and/or staff with these expertise throughout the life of the proposed work and beyond. Costs that can be accrued during the funding period that enable data preservation post-funding period should be allowed and recommended where needed.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Useful products of NIH funded research extend well beyond data alone. Sections should be included for each of the following and the scope of each section should be defined together with an example. To ensure that the policy can be enforced without unnecessary delays, implementation could start out as a data sharing document template; however, ultimately it would be better to develop a simple system dedicated to capturing and tracking research outputs, and for assessing compliance downstream.

Type of resource (eg. Software (including algorithms), Data (including images), Standards (including terminologies, data exchange formats), Protocols, Reagents (including Antibodies, cell lines, animals), Devices)

What the proposed funding covers:

--Creation of a new resource (eg. generation of Primary Data, or development of brand new software)

--Enhancement (eg annotation/curation of existing data, improvement of software)

--Derivation (secondary use of existing data)

Title of resource

Problem that the resource addresses

Proposed repository and / or registration venue of resource

Existing standards proposed to be used

URLs where resource is described if applicable

Proposed or existing license of resource

Legal and ethical implications, if any

While it is worthwhile for the policy to include information regarding data types are worthwhile (e.g. rationale for which data will be preserved, what level of processing will be performed, etc.), the policy lacks the actual recommendations for describing data standardization.

While "Identifying metadata" is described, this is vague and unenforceable. The plans for standardizing data must be included.

Resource sharing plans must identify the standards that applicants propose to use in order to describe and exchange data. If new standards or ontologies are needed, the lack of existing

ones must be documented and communities to advise or contribute to new development should be indicated. Plans to perform compliance testing against standards should be provided.

There are no recommendations regarding identifiers. Quality provisioning and use of identifiers is a key component of making data FAIR (in all categories of FAIR). We highly recommend that the NIH provide guidance for identifiers and require including this in the Plan. We have published identifier best practices based on our extensive efforts to reuse data from public knowledgebases and databases. <https://doi.org/10.1371/journal.pbio.2001414>

While we share the desire to make the plans easily digested, we feel it is unreasonable to limit the plans to two pages. The majority of data sharing plans that we (as experts in data reuse and data sharing), feel obligated to supply to adequately cover our plans - provided at the same level of granularity as the rest of the proposed work - are often longer than 2 pages (see our published example at <https://doi.org/10.5281/zenodo.3604521>) (see our published example at <https://doi.org/10.5281/zenodo.3604521>). A structured document format or dedicated platform would make it less onerous for applicants to prepare while also making it easier for reviewers to assess, and for compliance officers to follow up.

Other Considerations Relevant to this DRAFT Policy Proposal:

In designing grant review processes, it is really important to recognize that the scientific workforce is not limited to those who themselves have been awarded funds. Peer review for data science and data sharing aspects of research is often not assessable by domain experts found on such panels. We highly recommend that the grant and Plan review processes leverage experts in these areas, which is really no different than soliciting reviews from domain experts for a given research topic.

While the idea of FAIR has created community awareness that is to be applauded, more needs to be done to realize those goals. From the perspective of those of us in the business of reusing large numbers of data sources and aiming ourselves to make data more reusable, one needs to be cautious about "FAIR-washing." The policy recommendations should therefore focus on the important, specific, and enforceable practices that truly make resources more Findable, Accessible, Interoperable, and Reusable. For example, we refer the reader to our earlier FAIR-TLC response to RFI NOT-OD-16-133 Metrics to Assess Value of Biomedical Digital Repositories. <https://doi.org/10.5281/zenodo.203295>

Attachment:

Description:

Submission ID: 1448

Date: 1/11/2020

Name: Kristi Holmes

Name of Organization: CTSA Program Center for Data to Health (CD2H)

Type of Data of Primary Interest: Other

Type of Data of Primary Interest - Other: Translational science, including Genomic, Clinical, Imaging, Basic Biomedical, Qualitative, Other

Type of Organization: Other

Type of Organization - Other: NIH-supported coordinating center

Role: Scientific Researcher

Role - Other:

Domain of Research Most Important to You or Your Organization:

Translational medicine, informatics, data science

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Recommendation Ia: Clearly communicate the NIH Policy for Data Management and Sharing in the context of downstream users of the data

In order for the research community and broader public to make effective use of research data, the providers of this data must take into account the needs of these secondary consumers. One such need is timeliness in sharing research data. Reproducing and replicating research studies can be more readily achieved if the data are released as soon as possible upon a project's completion, and after all human participants' privacy rights and identifiable information have been sufficiently considered and protected in a dataset. Throughout the Draft Policy for Data Management and Sharing, no recommendation is given for either the minimum or maximum amount of time that has elapsed since a study's completion at which the study's data must be shared. A general recommendation such as those promulgated by the AHRQ and NOAA could be made, requiring data sharing either upon publication or 24-36 months after initial data collection. As section VI of the policy states, not all data collected from human subjects will be eligible for sharing, as can be outlined in the Data Management and Sharing Plan of individual awards. However the existence of such exceptions would not preclude a general recommendation for a data or milestone-based deadline for sharing of research data for the majority of funded studies, barring only those studies for which data cannot, for legal, ethical, or privacy reasons, be shared.

Recommendation 1b: Clearly define and describe best practices of data management and sharing and infrastructure that is recommended to be made available for this purpose at the institution level (also see Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan)

This section notes that the NIH "emphasizes the importance of good data management practices," as such practices will enable reproducibility of studies with the original data by other researchers. However beyond describing in the Data Management and Sharing Plan how data from a research study will be managed, no guidance is provided for best practices in data management. Data management will be most effective and comprehensively practiced when it is a collaboration between principal investigators, their affiliated organizations, and research funders. To ensure shared data meets a basic standard for sharing and reproducibility, the NIH can espouse guidelines and best practices in research data management, offering resources and consultations for those embarking on these efforts for the first time. The NIH can also provide clear recommendations for making data FAIR (findable, accessible, interoperable, and reusable). For instance, a basic level of metadata descriptors for shared data can be required, which allows for findability within repositories; a requirement for a DOI for datasets deposited in repositories for sharing can ensure accessibility; usage of controlled vocabularies for metadata description in repositories, for interoperability, can be required and assisted with links, guidance and recommendations; and clear community standards for data description and guidance on licences to assign to funded datasets can help NIH-funded data assets to be reusable. Within many extramurally-funded organizations, infrastructure already exists to support researchers in data management, making data FAIR, and sharing data. This infrastructure frequently involves academic health centers' health sciences libraries, leveraging the expertise of data librarians and data scientists who provide trainings, consultations, and assistance in sharing data through tools such as institutional repositories, which themselves are assets often administered and maintained by libraries. Further, making data reusable often requires the application and development of standards. This requires specialized expertise, which should be documented, where relevant, as part of the Data Management and Sharing plan and how such activities will be supported within the full course of the proposed research..

Additionally, recommendations about the temporal aspect of data management should be addressed: how long is the data expected to be available, and what are the expectations of NIH regarding sustainability of datasets and the dependencies that make it FAIR (Ex: stable repository, software integral for effective reuse, maintenance of ontologies).

In order to ensure data management/sharing best practices across the institution, NIH should encourage institutions to foster and support collaborations between investigators, libraries and research computing departments.

Section II: Definitions:

Recommendation IIa: Make a comprehensive glossary of data management and sharing available.

Best practices in data management and resource sharing impact a broad range of stakeholders and are necessarily found across a wide range of disciplinary boundaries, including computer science, information science, digital scholarship, economics, ethics, linguistics, philosophy, and the many different disciplines where data are generated and need to be managed and shared. Given the broad influence of different disciplines on the data generated as well as on the infrastructure needed to manage and share it, a glossary of terms will help to clarify terms and their definitions and ensure that collaborators have a shared understanding of concepts. The data glossary can be something quite simple with terms and definitions, such as <https://www.data.gov/glossary> and should be supplemented with source materials as a primer for self-paced learning and reference. Likewise, science is becoming ever more interdisciplinary, libraries are the one place where a global view of research data can be obtained, and who have the knowledge to manage data at both the disciplinary and general level. Libraries can play a role with orientation, training, and technical support.

Recommendation IIb: Modify and supplement terms and definitions as follows:

Data Management and Sharing Plan (Plan): A plan describing how scientific data will be managed, preserved, and shared with others (e.g., researchers, institutions, the broader public), as appropriate.

*No substantive changes recommended; delete "as appropriate".

Data Management: The process of validating, organizing, securing, maintaining, and processing scientific data, and of determining which scientific data to preserve.

*The definition for data management definition should also include the concept of "curating" to reflect the ongoing and active process of data management throughout the data lifecycle. The University of Illinois' Graduate School of Library and Information Science defines data curation as "the active and ongoing management of data through its life cycle of interest and usefulness."

Data Sharing: The act of making scientific data available for use by others (e.g., researchers, institutions, the broader public).

*No changes recommended

Metadata: Data describing scientific data that provide additional information to make such scientific data more understandable (e.g., date, independent sample and variable description, outcome measures, and any intermediate, descriptive, or phenotypic observational variables).

*The metadata definition should be supplemented to reflect the role that metadata can play in supporting FAIR -- Metadata: Data describing scientific data that provide additional information to make such scientific data more findable, accessible, interoperable, reproducible (FAIR), and understandable (e.g., date, independent sample and variable description, outcome measures, and any intermediate, descriptive, or phenotypic observational variables).

Scientific Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. NIH expects that reasonable efforts will be made to digitize all scientific data.

*It is important that the definition of Scientific Data remain worded as it is "...regardless of whether the data are used to support scholarly publications" to clearly communicate that the published paper should not be considered the sole driving force in how we consider Scientific Data and its generation, curation, analysis, dissemination, and preservation - especially given the evolving landscape of scholarly communication and options for dissemination of scholarship and knowledge.

*The recommendation to digitize all scientific data should likely be defined separately, along with community best practices for digitization to guide users. Support materials should include a consideration of format, description, metadata, preservation, rights, etc. and any hardware and software considerations as needed as to maximize the chances for survival and continued accessibility of digitized content well into the future. Support should also take into account person-hours needed to digitize materials not born-digital.

*Data, well-documented code, instructions for running analyses, workflows, and any other materials used to produce a publication and its figures/tables should be packaged and versioned so that the publication is computationally reproducible. There are many existing tools that can make this possible including reproducibility-friendly programming environments (e.g., jupyter, R Studio), software containers, and virtual machines.

Section III: Scope:

Recommendation III: Accommodate and support access of research objects beyond data.

Considering the great diversity of work across the biomedical research spectrum, we recommend that the scope is augmented to accommodate a recommendation from NIH that the wide range of research objects created during NIH funded research (beyond those that impact the production, curation, preservation, and dissemination, and downstream use of data e.g., protocols, training or engagement materials, technical reports, supplemental materials, survey instruments, etc.) are accommodated in data management and sharing plans more broadly. This will better enhance the visibility of research efforts, promote people and their expertise, support attribution of their work, aid the discovery and accessibility of datasets and other digital objects by the international scientific community, and support open and FAIR science. Investigators should be encouraged to make any resources generated with support from public funds freely accessible and repurposable by the public.

Section IV: Effective Date(s):

No recommendation.

Section V: Requirements:

Recommendation V: Clearly define the language and expectations around cost, responsible parties, and free or otherwise supported resources that would allow an investigator to comply with the Policy.

Clearly addressing questions such as: Is there a cap? Can costs be built in as direct or are they considered as part of the indirect? Can dedicated staff members be hired to the team for data management?

NIH funding is already highly weighted towards R1 institutions and previous NIH grant awardees, often leaving out smaller institutions that are doing excellent work in underserved and diverse populations, and by researchers from underserved and diverse backgrounds. Requiring a resource-heavy DMP without explicit understanding that this activity will be

supported in some way by the NIH will place yet another barrier to funding important and ground-breaking research that is already underfunded and difficult to discover because of lack of resources.

Section VI: Data Management and Sharing Plans:

Recommendation VIa: Substitute text to strengthen statement regarding factors that may impact one's ability to comply with the Policy (2nd paragraph, penultimate sentence):

Replace "Plans should include consideration of these factors, when applicable, in describing the approach to data management and data sharing." with "When such factors are present they must be clearly explained and sufficiently detailed to make it clear to the NIH that data sharing is not possible in these cases."

Recommendation VIb: NIH should provide the infrastructure, including established repositories for preserving and sharing of scientific data and, if need be, provide assistance in using these tools.

A requirement of researchers by the NIH to utilize established repositories for preserving and sharing scientific data requires a significant investment of time and research on the part of the investigator to identify and vet such repositories. In the absence of a data repository administered by the NIH itself, best practice recommendations for the selection of appropriate and established repositories for storing and sharing scientific data should be outlined. Recommendations can follow established guidelines from publishers for identifying and vetting repositories, such as those provided by the Public Library of Science (PLOS) [<https://blogs.plos.org/everyone/2018/03/01/criteria-for-recommended-data-repositories/>] and Data Repository Selection: Criteria That Matter [<https://osf.io/m2bce/>], blog at <https://blog.datacite.org/data-repository-selection-which-criteria-matter/>]. These requirements take into account both data preservation considerations and open access considerations. Recommendations of community-supported and vetted resources for identifying repositories, such as those listed at [FAIRsharing.org](https://fairsharing.org), would also be helpful.

In addition, the policy can more clearly outline a vision for data sharing compliance. What are the good, better, and best levels of data sharing that the NIH envisions for its funded datasets? Is full open access (deposit-enabled) desired, or creating a catalog record with a data access option to contact the data owner for access? Are any and all points in between acceptable? Encouraging, but not requiring, a basic level of sharing will likely lead to most sharing statements outlining reasons that the data cannot be shared.

An established repository can also take the form of a local institutional repository that has been robustly developed to meet Internet security protocols and to enable data-sharing on multiple levels. Allowance for interoperability through supported metadata-exchange channels can allow the use of local repositories with an option to transfer and migrate data to community-based or public repositories as needed. The Confederation of Open Access Repositories (COAR) has shepherded a highly relevant and forward-looking initiative that defines behaviors and technologies required for Next Generation Repositories [<http://ngr.coar-repositories.org/>]. This effort is focused on user needs and informed by best practices and standards as a foundation for a distributed, globally-networked infrastructure for scholarly communication, discovery, and innovation.

We wholeheartedly support the assertion from others in the community that support from NIH in the form of interoperability resources/human assistance would be helpful if researchers want to deposit to local or subject repositories and have their metadata and deposited files migrate easily to any NIH-sponsored tool.

Section VII: Compliance and Enforcement:

No recommendation.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Recommendation:

- *Allowable costs should include funds for data management and sharing
- *Make tools, training, and support available and discoverable to individuals and organizations, facilitating more efficient and effective good data practices in support of data management and sharing.
- *Consider making infrastructure or supplemental awards available at the institutional level to build or supplement local capacity. Awards could prioritize collaborative teams across key stakeholders in libraries, IT and research computing, and other research-oriented collaborative groups on campus.
- *Finally, a great wealth of NIH-funded resources exist in CD2H and across many other efforts that can help to support research data management and data sharing. Support discovery efforts to make these resources, services, and expertise themselves as Findable, Accessible, Interoperable, and Reproducible as possible.

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

In response to Response to RFI NOT-OD-19-014, we shared the following Requirements for Data Management and Sharing Plans [<http://doi.org/10.5281/zenodo.2302593>]; the Monarch

Initiative (<https://monarchinitiative.org/>) Data and Resource Sharing Plan has been uploaded as an example plan.

The basic components of a DMP. Resources are finite; perfunctory sharing of poorly-documented data does not achieve the objectives of the spirit of FAIR. However, we also recognize that not all data can be held to the same standard. Wherever practicable, data should follow good data practice; for example, data should be published together with methods and factors known to impact reliability. The full "chain-of-custody" should be documented from the generation of data through its safeguarding and analysis. Not only does this provenance help the users of data assess its veracity--but it also helps ensure attribution, whether to scientists or to members of the public.

Data management and sharing plans should describe a reliable, consistent, and well understood mechanism for recording data and its associated metadata. Supporting "knowledge services" should be described that provide the appropriate shared vocabularies, ontologies and reference information for documenting data. Knowledge services should include information about units, representational forms, scientific methods, domain knowledge, organizations, researchers and ontologies that define the methods and knowledge of the domain itself. This service must be made freely available, reliable and subject to independent verifiability and community correction. A description of the search and retrieval services should also be required, to allow researchers to query, discover and utilize data. The retrieval service must include both the data and its associated pedigree (see attribution, provenance, and reproducibility below) and must provide a mechanism that allows the provenance of derived and enhanced data to be re-entered into the data management plan faithfully and traceably.

Assessment of quality.

Data Management and Sharing plans are required for any grant proposals over 500K direct costs/year; however, making them required and scored for all grants is one of the single most impactful changes that the NIH could make.

Moreover, we believe that this section should be scored as part of the review criteria. Because most people on NIH review panels are experts in the given specific area of science and not in data management or open science, it would be beneficial for NIH to seek out this expertise for the review panels. It should be noted that often the professional profile of open science experts expands outside the usual background of more typical investigators with R01 funding, and as such, a different approach and criteria for selecting these experts should be created. This may be true for other areas of the review panel as well, but our RFI response is focused exclusively on the Data Management and Sharing plans.

Execution of the data management plans. Throughout the life of a funding award, the data provider/investigator should be evaluated for adherence to the data management and sharing plans by external expert reviewers. Data management and sharing plans should be versioned and updated as the science, community technology, and standards evolve. To encourage investigators to outline effective data management and sharing plans, funders should incentivize grant applicants to include funding allocations within their grant applications to establish partnerships with teams developing standards or data management systems. This also requires an increase in the overall award to reflect the added burden of data management. Otherwise the data management and sharing plan becomes an unfunded mandate. As well, there should be consequences for project leaders who do not abide by the spirit of their proposed plans.

Considerations regarding data reusability. We urge NIH to consider attributes of data management and sharing that most limit the actual reusability of data. While the FAIR principles have helped to promote awareness, these don't provide guidance on how to actually make data more reusable and therefore useful. We have written extensively about factors that maximize data reusability, specifically for data integrators and third party users. For example, in response to RFI NOT-OD-16-133 (Metrics to Assess Value of Biomedical Digital Repositories), we wrote about FAIR-TLC, where the T is Traceability, L is licensure, and C is connectedness. <https://doi.org/10.5281/zenodo.203295>. There are numerous specific recommendations in this RFI that are relevant here, but we therefore refer the reader to that other document. Finally, data should be openly available for use, in both human and machine readable formats.

Data licensing and data use agreements. We recommend that the NIH mandate that all publicly funded sources of data, knowledge, or tools be documented with a clearly defined and preferably standard licence. The (Re)usable Data Project has evaluated the licensing of 56 NIH publicly funded data resources (including some NIH sources), and has illuminated the fundamental barriers to data science due to a lack of ability to mash-up and redistribute these data. A preprint describing the findings is here: *A Measure of Open Data: A Metric and Analysis of Reusable Data Practices in Biomedical Data Resources*. This inability to redistribute seemingly "open" data sources limits their potential impact, not only for discovery, but also in for patient care. This problem is so significant, the community wrote a letter to Francis Collins, "Request for Community partnership in data resource licensing planning".

A change in the Data management and sharing plan seems ideal timing for coordinated improvements to licensing across NIH funded resources. We believe that licensing is one of the most important and overlooked requirements needed to make data reuse a reality. A license MUST be required in all DMPs.

Identifiers. All genes, proteins, variants, phenotypes, diseases, chemicals, species, biosamples etc. should be referenced using appropriate vocabularies/ontologies/reference data. The scientists that generate and publish data are those best positioned to properly annotate it; they should do so using community best practices. Provisioning and management of identifiers should be detailed in the data management and sharing plan; a three-year community coordination effort led to some agreed-upon best practices published in PLoS Biology [<https://doi.org/10.1371/journal.pbio.2001414>] Persons should be referenced by an identifier, with ORCID being the most prominent and most integrated person-level identifier in use. Moreover, organizations should also be referred to by a persistent identifier [<https://ror.community/>]. Resolvers (examples are N2T.net and identifiers.org) should be used to persistently redirect identifiers where they are made public, and details to avoid link-rot should be included in the management plan. Documentation of identifier provisioning, schema, and examples are often lacking, and we would recommend that this be a required component.

Other Considerations Relevant to this DRAFT Policy Proposal:

Attribution, provenance, and reproducibility. Reproducible science depends in part on knowing what has been specifically performed and by whom. Not only does this hamper reproducibility and evidence for scientific conclusions, it also disincentivizes diverse types of contributions. A data management plan should also document how it plans to include attribution, provenance of the data, and to address any reproducibility aspects (including identification of primary resources, see <https://doi.org/10.7717/peerj.148>). The goals should be to (1) to give credit where credit is due for everyone - regardless of their discipline, title, or contribution, regardless of whether it fits into narrowly defined construct of success; (2) to enable linking of all research outputs created during a study together, enabling access to data and results (even negative results), tools, and ideas which often remain invisible because they do not appear in a published manuscript, as access helps drive discovery; (3) facilitate re-use of the information in a variety of ways by all stakeholders (individuals, publishers, scholarly organizations, data repositories, and funding agencies). For example, relationships between people and their products/activities can be used to track research trends; to understand and leverage influences or projects; to promote collaboration and team formation; as recommender systems for scholarly products or methodologies; and to present a complete record of results and research outputs. Fundamentally, the data about the contributions that scholars make should be as open as the data and resources themselves if we really aim to incentivize sharing and open science. The CD2H has completed a number of activities to aid in supporting improved attribution, including the development of a data model and implementation into existing research systems, both with community and stakeholder input and coordination.

Attachment:

Data and Resource Sharing Plan - Monarch R24 2019-10-25.pdf

Description:

Monarch Initiative (<https://monarchinitiative.org/>)_Data and Resource Sharing Plan (example plan)

DATA AND RESOURCE SHARING PLAN

While a formal resource sharing plan is required due to the requested level of support (>\$500K), resource sharing has **voluntarily been at the heart of our highly collaborative phenomics work since its inception.**

The Monarch platform is comprised of multiple open-source, open-access components: **data, ontologies and other standards, and software tools, and algorithms.** We want researchers to be able to easily discover, search, explore, and download our work. We will continue to provide appropriate locations for the research community to access the resources we generate, along with documentation, links to the software, and other information that will help improve the accessibility and reusability of our products. Additionally, we provide communication channels for the community, such as a help desk, mailing lists, video tutorials, social media feeds, public presentations, and workshops, to facilitate two-way communication between project members and outside researchers, as well as enabling users to engage with each other. Use of GitHub issue trackers enables community members to submit questions, bug reports and suggestions that will be visible not only to project members but to other users as well, consistent with our commitment to transparency and community development.

Our resource sharing plan adheres to the NIH Grant Policy on Sharing of Unique Research Resources including the Sharing of Biomedical Research Resources Principles and Guidelines for Recipients of NIH Grants and Contracts issued in December 1999 (grants.nih.gov/grants/policy/data_sharing/). We are committed to enhancing the value of research and furthering the advancement of public knowledge. All resources developed by the project will be made available to the scientific community under the terms specified by NIH policy. Both software and resources will be released with licenses that enable further reuse, modification, and redistribution by downstream users and developers. Throughout the development process, key releases of software or data will be deposited in Zenodo, generating an associated DOI. We will evaluate our data sharing according to the rigorous rubrics we worked to develop[3–5].

Software

We always make a **standard and stable version of each piece of code available** for public download. The data and source code generated by this project will continue to be **stored in GitHub**. GitHub records basic metadata such as dates when changes were made, and by whom, and commit messages from developers. We also include documentation for users and software developers, in easy to view sites such as Read The Docs (readthedocs.org), to facilitate reuse of the software. In GitHub progress is trackable as code is committed and pushed and stable releases are created. Where practicable and useful, we will also continue to make our software accessible via REST APIs (see also data access methods below).

Our software licensing strategy:

In establishing a policy for software releases, we are guided by the philosophy that the software we produce should find the widest audience for dissemination and that should there be no hurdles for any legitimate investigator to access the software tools. All software written for this project will be released under the most liberal license possible given institutional limitations, with the default being the BSD 3-Clause License (opensource.org/licenses/BSD-3-Clause).

1. Allows all intellectual properties (IP) generated from the proposed research, including code developed to implement algorithms, data transformation scripts, and other functionalities of the system to be **freely available to everyone at no cost**. This includes researchers and educators in academic institutions, non-profit organizations and government laboratories, as well as for-profit companies;
2. **Allows all community members to modify** the software and share their changes under the same licensing model;

3. Makes it possible for anyone (not just members of our collaboration) to **freely commercialize advanced or customized versions** of the software or incorporate modified versions of portions of our code into commercial software packages, thereby increasing dissemination; and
4. Requires **licensing information to be included** with all distributed software products, including those modified by others.

We will encourage those who fork, modify or add to our code to submit it to GitHub to be considered for inclusion into our overall codebase.

Major components of the existing Monarch software stack and where to find them

Component	Description	Website or Documentation	Repository address
Dipper	data ingest pipeline	dipper.readthedocs.io	github.com/monarch-initiative/dipper
Monarch portal	User interface	monarchinitiative.org	github.com/monarch-initiative/monarch-app
k-BOOM	Algorithm to auto-generate mappings between the elements of different disease terminology sources.	2019 BiorXiv [6]	github.com/monarch-initiative/kboom
OwlSim	Ontology-based profile matching	berkeleybop.org/software/owlsim	github.com/monarch-initiative/owlsim-v3
Phenol	Phenotype Ontology Library	phenol.readthedocs.io	github.com/monarch-initiative/phenol
HPO Case Annotator	Next-generation biocuration app for annotating cases and PhenoPackets.	N/A	github.com/monarch-initiative/HpoCaseAnnotator
PhenoteFX	A Java app that is designed to help create and maintain Human Phenotype Ontology annotation files.	phenotefx.readthedocs.io	github.com/monarch-initiative/PhenoteFX
Exomiser	A tool to annotate and prioritize exome variants	exomiser.github.io/Exomiser/general	github.com/exomiser/Exomiser
Web HIPPO	Deriving insight from the medical literature by fuzzy semantic searches over diseases and phenotypes.	hippo.monarchinitiative.org	github.com/monarch-initiative/web-hippo
Phenogrid	Phenogrid is a Javascript component that visualizes semantic similarity calculations provided by OWLSim, as provided through APIs from the Monarch Initiative.	(Component in monarchinitiative.org portal)	github.com/monarch-initiative/phenogrid

Timeline: All of the resources and products that are generated, whether software, data, or standards, will be released early and often, and at least quarterly under a standard open license where permitted by the original sources. Versions of each will have stable persistent URLs so that they can be readily identified for reference in publications.

Data

The underlying data in Monarch are derived from multiple external sources[1,2] (<https://monarchinitiative.org/about/sources>) the use and secondary use of which is governed by the corresponding original license for each source. As we have described in detail, this is not without complex licensing implications for data integrators and their users [3]. Although it is not possible for Monarch (or for any aggregator) to legally provide unified data sources under a unified license, we will continue to make our data available via multiple mechanisms and we will abide by any and all sharing restrictions placed by the original provider including but not limited to appropriate citation of all the data sources that we use.

Access method	Description	Available formats	URL
API	Programmatic access to associations of individual entities	JSON	api.monarchinitiative.org/api
Downloads	Downloads of single datasets or the whole Monarch graph, including the proposed uPheno-compliant cross-organism mapping files.	RDF (whole datasets), TAB (subsets) as well as our Neo4j database, Solr index are also publicly available in this archive.	data.monarchinitiative.org and archive.monarchinitiative.org/latest
Interface	Ontology-aware, browsable, searchable knowledge graph	Download any result in TAB format	Via monarchinitiative.org
Evaluation data for algorithms	Downloads of publicly available solved clinical cases [proposed herein]	Download as phenopackets	[proposed herein]

This plan reflects that data sharing is an essential aspect of responsible scientific conduct and, in particular, that data gathered with NIH funding should be made publicly available in a time period that assures appropriate confidentiality until the data are accepted for publication, but that also gives the research community and the public at large prompt access to potentially important findings. **That time period is defined as a maximum of 1 year from the final generation of data. This estimate is based on our experience with large studies to balance data quality and rigor with maximizing data access and fostering data re-use.**

HIPAA and Human Subjects

The data sets we propose to use here are publicly available and thus do not generate any privacy or confidentiality concerns. However, because Monarch sits in a complex translational landscape we have -- for good measure -- included a separate Human Subjects section.

NIH Data Commons

The work of the proposed project will not directly generate any data suitable for deposit in the NIH Commons; however, the resources that we will create are highly relevant to that project. In time, the Monarch resources will help to prospectively unify the Data Commons content as the platform comes into fruition. To this end, we have successfully piloted the use of Mondo in the context of semantic search for discovery within the NHLBI STAGE Data Commons. Our goal is to make the Monarch data as widely discoverable and reusable as possible while maintaining the provenance and attribution.

Standards for data annotation and exchange

Input and output data will be annotated with ontology terms from an appropriate OBO Foundry ontology, where available. Those annotations will be stored in a CSV file. All data will be made available in non-proprietary formats.

Data and exchange standards

Standard	Description	GitHub Repo	Website or documentation	License
Phenopackets	An open standard for sharing disease and phenotype information will improve our ability to understand, diagnose, and treat both rare and common diseases.	github.com/phenopackets	phenopackets.org	BSD3
uPheno ontology pattern templates	Cross-modality, cross-species design patterns	github.com/obophenotype/upheno	ebi.ac.uk/ols/ontologies/upheno	CC0 1.0 Universal
BioLink-Model	A high-level data model for representing biological and biomedical knowledge.	github.com/biolink/biolink-model	biolink.github.io/biolink-model	CC0
Monarch API	A lightweight API for exchange of information that sits on top of the Monarch knowledge graph	github.com/biolink/biolink-api	github.com/biolink/biolink-api/blob/master/README.md	BSD3

Ontologies

Ontology	Website	GitHub Repo	License
HPO	hpo.jax.org and ebi.ac.uk/ols/ontologies/hp	github.com/obophenotype/human-phenotype-ontology	Custom no derivatives license (The most similar standard license is CC-BY ND)
SEPIO	obofoundry.org/ontology/seprio and ebi.ac.uk/ols/ontologies/seprio	github.com/monarch-initiative/SEPIO-ontology	CC-BY 3.0
Mondo	monarch-initiative.github.io/mondo and ebi.ac.uk/ols/ontologies/mondo	https://github.com/monarch-initiative/mondo	CC-BY 3.0
MAXO	N/A	https://github.com/monarch-initiative/MAXO	CC-BY 3.0
uPheno	obofoundry.org/ontology/upheno and ebi.ac.uk/ols/ontologies/upheno	github.com/obophenotype/upheno	CC0 1.0 Universal
GENO	obofoundry.org/ontology/geno and ebi.ac.uk/ols/ontologies/geno	https://github.com/monarch-initiative/GENO-ontology	CC-BY 3.0

Protocol and Reagent Sharing Plan

The proposed project will generate only digital (software and data) resources, not material resources. No protocols, organisms, or reagents will be used or generated.

SHARING PLAN LITERATURE CITED

1. McMurry JA, Köhler S, Washington NL, Balhoff JP, Borromeo C, Brush M, et al. Navigating the Phenotype Frontier: The Monarch Initiative. *Genetics*. 2016;203: 1491–1495. doi:10.1534/genetics.116.188870
2. Monarch Initiative Explorer. [cited 22 Oct 2019]. Available: <http://beta.monarchinitiative.org/about/data-sources>
3. Carbon S, Champieux R, McMurry JA, Winfree L, Wyatt LR, Haendel MA. An analysis and metric of reusable data licensing practices for biomedical resources. *PLoS One*. 2019;14: e0213090. doi:10.1371/journal.pone.0213090
4. Haendel M, Su A, McMurry J. FAIR-TLC: Metrics to Assess Value of Biomedical Digital Repositories: Response to RFI NOT-OD-16-133. 2016. doi:10.5281/zenodo.203295
5. (Re)usable Data Project. [cited 23 Oct 2019]. Available: <http://reusabledata.org>
6. Mungall CJ, Koehler S, Robinson P, Holmes I, Haendel M. k-BOOM: A Bayesian approach to ontology structure inference, with applications in disease ontology construction. *bioRxiv*. 2019. p. 048843. doi:10.1101/048843

Submission ID: 1450

Date: 01/13/20

Name: Peter Sorger, Laura Maliszewski, Catherine Luria

Name of Organization: Harvard Medical School

Type of Data of Primary Interest:

Type of Data of Primary Interest - Other:

Type of Organization:

Type of Organization - Other:

Role:

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

Harvard-LSP_Response_to_DRAFT_NIH_Policy_for_Data_Management_and_Sharing-F(1).pdf

Description:

Attachment submitted via email

Response to “*DRAFT NIH Policy for Data Management and Sharing*”

- Peter Sorger (Otto Kraye Professor of Systems Biology at Harvard Medical School, Founding Director of the HMS Laboratory of Systems Pharmacology, Head of the Harvard Program in Therapeutic Science)
- Laura Maliszewski (Executive Director of the Harvard Program in Therapeutic Science and the HMS Laboratory of Systems Pharmacology)
- Catherine Luria* (Scientific Program Manager, Harvard Program in Therapeutic Science and the HMS Laboratory of Systems Pharmacology) *Please contact catherine_luria@hms.harvard.edu for further information.

Over the past five years we have been involved in multiple large NIH/NCI grants that involve data sharing and management activities (including the NIH LINCS and IDG programs and the NCI HTAN program). We therefore have considerable experience in implementing such activities from the perspective of practicing scientists and NIH grantees. We have commented on this topic in a perspective in *Science Translational Medicine*¹ and written multiple papers attempting to improve the reproducibility of one important type of data: preclinical assays of drug response²⁻⁴.

Overall, we are highly supportive of the development of a more robust set of policies and associated infrastructure for data management and sharing that conform with FAIR principles. NIH’s effort in seeking and incorporating feedback from the scientific community on the October 2018 *Proposed Provisions for a Draft NIH Data Management and Sharing Policy* has resulted in what we view as positive changes. The statement that “costs associated with data management and data sharing may be allowable under the budget for the proposed project” and the associated *Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing* acknowledges the resources required to comply with FAIR principles, a critical point that was not addressed in the previous draft.

However, some of the concerns that we expressed in response to the *Proposed Provisions* remain. We believe that the implementation of policies under the current draft as well as successful, consistent adherence to data management plans proposed by individual investigators can only be accomplished as a part of a multi-part multi-year strategy that also includes: (i) education – including education of graduate students and fellows (ii) development of infrastructure for validating, storing and disseminating diverse types of data (iii) much more substantial investment in innovation and in computational tools and approaches (iv) incentives for timely and useful data deposition as opposed to simple mandates and penalties associated with specific data types.

First, we feel strongly that NIH should provide education and training materials to principal investigators before the requirements described in the November 2019 draft document are implemented. NIH has developed a very flexible draft document to accommodate a wide range

of research areas and approaches, but more guidance is needed to ensure that PIs provide reasonable data plans that can be implemented successfully once a grant is funded. Controlled vocabulary for data types and how data will be shared (e.g. individual vs. aggregated vs. summarized) should be provided along with guidance on minimum standards for the types of materials that should accompany scientific data, including standardized formats for study protocols and information about data collection instrumentation, software and code.

It is also essential to recognize that for most of non-genomic data generated in basic and translational research, there are no generally accepted standards or established repositories. The annotation and re-use of heterogeneous data arising from perturbational studies (the vast majority of the mechanism-oriented research in the NIH portfolio) is fundamentally different from storing and disseminating a single type of data on a steady-state sample (e.g. a genome sequence). Formats and reporting standards have not yet been developed for most types of microscopy data, the many variants of mass spectrometry, multiplex immuno-assays on cell and tissues lysates or on components of the microenvironment and emerging data types such as spatial transcriptomics or multiplex imaging among others. New standards will be required to adequately annotate experiments in which these types of data are collected over time following genetic or drug-mediated perturbation of a system.

It is not sufficient to establish ontologies for these data: tools must be developed to annotate data according to relevant standards, to impose uniform vocabularies and to validate annotations. While the current draft policy addresses the possibility of budgeting for data management, substantial investment in software development and hardening is required to implement FAIR data standards across all data types. The apps we all enjoy using outside of our scientific lives (e.g. Google Maps) have involved a much higher level of refinement than any of the code we use for storage and annotation of scientific data. cBioPortal and Cytoscape are two examples of well-developed code – both required large teams and many years of investment. In addition to the extensive work required for data types (and even file formats) for which no tools currently exist, existing infrastructure must be more actively supported. For example, the OMERO image management standard that we helped to develop over a decade at MIT (now in wide-spread use) has never received any NIH support despite multiple attempts. The entire development team was moved from the US to the UK, where it is now headed by Jason Swedlow with EU/UK funding.

The process of accurately measuring a natural phenomenon (and also a human-engineered device) is the subject of the field of metrology/measurement science which is closely related in the case of biomedicine to analytical chemistry. There has been little or no significant investment in metrology by the NIH and the fundamental tenants of analytical chemistry are unfamiliar to most biomedical investigators (with noteworthy exceptions, such as R-factors in crystallography). Much has been made of insufficient statistical training by our community but we believe that absence of metrology is equally problematic: if the data are no good to begin with, more rigorous statistics will not save us. We cannot expect assays to be reproducible if we spend little or no effort studying sources of variation and irreproducibility. Importantly, this situation could be easily rectified by the creation of RFA/RFPs focused on measurement

science, error modeling etc. Given the relatively small size of the community having the necessary expertise, we believe that a multi-year commitment that will grow the number of instigators is required. Given the breadth of the topics, we suggest that a two-step review similar to that used for DP2 and similar interdisciplinary programs be considered.

Other considerations:

- Assessment of data management plans: Plans will be assessed as part of Just-in-Time for extramural awards. While this saves effort on the part of proposal developers who will ultimately not be funded as well as reviewers, it prevents data management plans from being assessed by a team of scientific reviewers, which is concerning.
- Data archiving: Draft guidance on allowable cost states that recurring fees for sharing and preserving data in existing repositories may be included in proposal budgets. NIH does not explicitly address whether data storage will continue to be funded after the grants which supported data production have ended, how long data must be preserved, and how these decisions might be revised if a particular type of data is no longer of interest to the community due to technological advancements. The possibility of retiring some types of data (e.g. RAW files or image-based screening data) after a predetermined period must also be considered.
- Standards development: We hope that NIH will catalyze community groups to develop community-based standards for data types for which no standards exist already. The task of developing these standards is too large for individual grantees and standards developed by small groups are unlikely to result in FAIR data. In some cases, innovative machine-human collaborations (e.g. AI) are likely to be required. Standardization efforts should be international; particularly in the area of pre-clinical, basic research data, EMBL/EBI is well ahead of anything in the US.
- Standards dissemination: Before the proposed data management and data sharing requirements are implemented, existing data standards should be more actively supported and disseminated. Many existing standards are currently difficult to locate and sometimes poorly documented, meaning that research groups struggle to find and correctly implement existing standards. The Common Data Element Resource Portal suggested by NIH in *Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan* is largely limited to disease-specific clinical studies and is not necessarily relevant to many basic research projects. A more broadly applicable and more easily searchable resource than the Common Data Element Resource Portal should be in place to provide information about existing data standards for NIH investigators. For data types for which no NIH data repository exists, a list of accepted non-NIH repositories will be required; these will need persistent unique identifiers for deposited data. For example, the PRIDE database maintained by EMBL/EBI is becoming the standard for deposition of proteomics data. NIH should support and mirror this and similar types of repositories, by analogy with mirrored genomics databases. If existing standards are not easy to find, some groups may “reinvent the wheel” and develop new, redundant standards, which again, ultimately reduce data FAIRness.
- Software tools: As noted above, a standard is useless without the software infrastructure needed to implement and validate it. In our experience, this often

requires some ability in scripting – we teach all of our trainees basic Python coding skills. It is for this reason that data annotation and education and closely interrelated.

REFERENCES

1. AlQuraishi M, Sorger PK. Reproducibility will only come with data liberation. *Sci Transl Med*. 2016 May 18;8(339):339ed7. PMID: PMC5084089
2. Hafner M, Niepel M, Chung M, Sorger PK. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat Methods*. 2016 Jun;13(6):521–7. PMID: PMC4887336.
3. Niepel M, Hafner M, Williams EH, Chung M, Barrette AM, Stern AD, Hu B, Gray JW, Birtwistle MR, Heiser LM, Sorger PK. A multi-center study on factors influencing the reproducibility of in vitro drug-response studies. *Cell Systems*. 2019.24 July 2019, Pages 35-48.e5 PMID: PMC6700527
4. Hafner M, Niepel M, Sorger PK. Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics. *Nat Biotechnol*. 2017 Jun 7;35(6):500–502. PMID: PMC5668135

Submission ID: 1451

Date: 01/13/20

Name: Kirk Francis, Kitcki A. Carroll

Name of Organization: The United South and Eastern Tribes Sovereignty Protection Fund (USET SPF)

Type of Data of Primary Interest:

Type of Data of Primary Interest - Other:

Type of Organization:

Type of Organization - Other:

Role:

Role - Other:

Domain of Research Most Important to You or Your Organization:

DRAFT NIH Policy for Data Management and Sharing

Section I: Purpose:

Section II: Definitions:

Section III: Scope:

Section IV: Effective Date(s):

Section V: Requirements:

Section VI: Data Management and Sharing Plans:

Section VII: Compliance and Enforcement:

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing:

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan:

Other Considerations Relevant to this DRAFT Policy Proposal:

Attachment:

USET SPF Comments to NIH_Draft Policy for Data Management and Sharing and Supplemental Guidance FINAL 1_10_20.pdf

Description:

Attachment submitted via email

Transmitted via
SciencePolicy@mail.nih.gov

January 10, 2020

Andrea Jackson-Dipina, Dr.PH
Director of the Division of Scientific Data Sharing Policy
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Dear Dr. Jackson-Dipina,

The United South and Eastern Tribes Sovereignty Protection Fund (USET SPF) is pleased to offer comments to the National Institutes of Health (NIH) regarding the agency's *Request for Public Comments on a Draft NIH Policy for Data Management and Sharing and Supplemental Draft Guidance*. USET SPF recognizes that sharing data among the scientific community is imperative for scientific discovery and advancement. However, as NIH advances its policy regarding data management and sharing, the agency must recognize the historical relationship between scientific study and Tribal Nations, where researchers committed ethical violations against our communities and our people. We underscore that NIH must seek to prevent these violations from ever occurring again by ensuring all NIH policies are reflective of the federal government's obligation to honor, protect and uphold Tribal sovereignty by requiring explicit consent from Tribal Nations.

USET SPF is a non-profit, inter-tribal organization representing 30 federally recognized Tribal Nations from the Canadian border to the Everglades and across the Gulf of Mexico¹. Both individually, as well as collectively through USET SPF, our member Tribal Nations work to improve health care services for American Indians. Our member Tribal Nations operate in the Nashville Area of the Indian Health Service, which contains 36 IHS and Tribal health care facilities. Our patients receive health care services both directly at IHS facilities, as well as in Tribally-operated facilities under contracts with IHS pursuant to the Indian Self-Determination and Education Assistance Act (ISDEAA), P.L. 93-638.

¹ USET SPF member Tribal Nations include: Alabama-Coushatta Tribe of Texas (TX), Aroostook Band of Micmac Indians (ME), Catawba Indian Nation (SC), Cayuga Nation (NY), Chickahominy Indian Tribe (VA), Chickahominy Indian Tribe–Eastern Division (VA), Chitimacha Tribe of Louisiana (LA), Coushatta Tribe of Louisiana (LA), Eastern Band of Cherokee Indians (NC), Houlton Band of Maliseet Indians (ME), Jena Band of Choctaw Indians (LA), Mashantucket Pequot Indian Tribe (CT), Mashpee Wampanoag Tribe (MA), Miccosukee Tribe of Indians of Florida (FL), Mississippi Band of Choctaw Indians (MS), Mohegan Tribe of Indians of Connecticut (CT), Narragansett Indian Tribe (RI), Oneida Indian Nation (NY), Pamunkey Indian Tribe (VA), Passamaquoddy Tribe at Indian Township (ME), Passamaquoddy Tribe at Pleasant Point (ME), Penobscot Indian Nation (ME), Poarch Band of Creek Indians (AL), Rappahannock Tribe (VA), Saint Regis Mohawk Tribe (NY), Seminole Tribe of Florida (FL), Seneca Nation of Indians (NY), Shinnecock Indian Nation (NY), Tunica-Biloxi Tribe of Louisiana (LA), and the Wampanoag Tribe of Gay Head (Aquinnah) (MA).

Native people and Tribal communities continue to face negative impacts from previously unauthorized and unpermitted use of genomic data without Tribal Nation informed consent (*Arizona Board of Regents v. Havasupai Tribe*). Despite Tribal efforts to require informed consent regarding the use of Tribal data, NIH has continued to advance certain initiatives, including a Tribal Consultation Policy, without engaging in meaningful consultation with Tribal Nations. In August 2018, USET SPF provided comments to NIH regarding the agency's inadequate Tribal consultation on three initiatives, including proposed provisions for the Draft NIH Data Management and Sharing Policy. In our comments, we note NIH's ineffective and insufficient consultation practices with Tribal Nations which are in violation of the U.S. Department of Health and Human Services (HHS) Tribal Consultation Policy. While we recognize some improvement with the addition of clear deadlines and a request for broader guidance on research with our population, we remain focused on the results of these efforts. As stated in past communications, we expect NIH to engage in consultation with Tribal Nations in a transparent and meaningful manner to resolve outstanding concerns from Indian Country to ensure sovereignty is upheld and past abuses never happen again. This includes taking active steps to implement the recommendations and guidance of Tribal Nations. In addition to our below recommendations, we continue to underscore that NIH must consult with Tribal Nations on an ongoing basis regarding the agency's research, data sharing, and data management policies to ensure the privacy of Tribal Nation communities, as well as American Indian and Alaska Native (AI/AN) individuals.

Draft NIH Policy for Data Management

Within the Draft NIH Policy for Data Management, the 'Effective Date' seems to include only research to be conducted in the future. Because of the historical research abuses outlined above, USET SPF believes that ALL projects, current and future, be required to submit a Data Management Plan. There is an opportunity to ensure that data currently being collected and utilized is protected. Under the 'Compliance and Enforcement' section, USET SPF insists that an oversight mechanism, specific to Tribal Nation data, designed consultation with Tribal Nations, be included. This mechanism would detail Tribal Nation data protection best practices, procedures, ensure researcher compliance, and recommend consequences for violations.

Supplemental Draft Guidance: Elements of an NIH Data Management and Sharing Plan

USET SPF appreciates NIH's forethought of requiring a data management and sharing plan as part of all research funded or conducted in whole or in part by the NIH. As part of the federal trust obligation to federally recognized Tribal Nations, NIH has a duty to honor, protect and uphold Tribal Nation sovereignty in its efforts 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.' Therefore, USET SPF's recommendations that all submitted data management and sharing plans require an element entitled 'Tribal Nation(s) Population' that shows, first and foremost, evidence of Tribal Nation consent for data sharing and collection. This element should be designed through ongoing consultation with Tribal Nations.

No Tribal Nation data should be included in any level of access without explicit Tribal Nation consent. The consent mechanism varies from Tribal Nation to Tribal Nation and may take the form of Tribal Nation Council resolutions, signed memorandums of understanding with a designated Tribal Nation leader, etc. In addition to documented Tribal Nation consent, the plan must address additional considerations between the researcher and the Tribal Nation such as:

- Data ownership and sovereignty;
- Publication requirements and Tribal Nation consent procedures;
- Community risks and benefits of the research and any potential data sharing

- Specimen use, storage, and destruction policy;
- Work product ownership and sovereignty;
- Data use provisions for future studies; and
- Data sharing and use provisions for NIH-maintained databases (i.e., genomics).

USET SPF notes that the above list is not exhaustive and that NIH must seek formal Tribal consultation on these recommendations, as well as on any future draft Data Management and Sharing Plan requirements. Much as 'The Belmont Report' and the National Research Act of 1974 have resulted in human subject protection as standard practice among researchers, USET SPF believes that such a required element for all NIH-funded research proposals will integrate Tribal Nation protection and sovereignty concerns into common research practice. USET SPF reminds NIH of its trust obligation to ensure that Tribal Nations can protect our citizens and data, and this obligation supersedes any data sharing interests.

NIH Institutional Review Boards

In addition to the comments above regarding data sharing, USET SPF strongly recommends Native representation on all NIH Institutional Review Boards (IRB) reviewing studies that include AI/AN people. Without this unique perspective advising the IRB, Native people and communities will not be adequately protected from research abuse.

Conclusion

Based on NIH's previous practices, USET SPF continues to be deeply concerned that the final policy may not contain necessary protocols for integrating Tribal Nation protection and sovereignty concerns into common research practice. As we seek to protect, regulate, and maintain ownership over the data of our citizens and Nations, NIH has a legal and moral trust obligation to uphold the sovereign status of Tribal Nations. Because data management and sharing policies have significant implications for Tribal governments and their citizens, NIH must seek formal Tribal Consultation on this and all other issues. We look forward to the opportunity to partner with NIH to ensure research conducted and data collected in our communities and with the Native population is done in a way that reflects our sovereign status and seeks to reconcile our difficult history with the scientific community. From study development and design, approval, ethical review, data collection/analysis, result interpretation, and reporting research within the Native community cannot be ethically conducted without Native representation and consent at all stages of the research process. Should you have any questions or require further information, please contact Ms. Liz Malerba, USET SPF Director of Policy and Legislative Affairs, at LMalerba@usetinc.org or 202-624-3550.

Sincerely,



Kirk Francis
President

Kitcki A. Carroll
Executive Director