Compiled Public Comments on

Request for Information: Critical resource gaps and opportunities to support Next Generation Sequencing (NGS) test development, validation, and data interpretation, including through the use of technologies such as artificial intelligence (AI)/machine learning (ML)

Guide Notice Number: NOT-OD-21-162

August 3, 2021 – November 1, 2021

Table of Contents

ID: 1764

Submit date: 10/5/2021

I am responding to this RFI: On behalf of an organization

Name: Ryan Dittamore

Name of Organization: Cordance Medical

Type of Organization: Industry (Biotech/Device/Pharmaceutical Company)

Role: Investigator/Researcher

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

Neuro-oncology

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**
   The incidence of brain metastasis continues to rise as we improve the therapeutic efficacy of agents in breast, lung, melanoma, kidney, and other solid tumor cancers. Collectively, with GBM and other gliomas, the ability to measure and track malignant brain tumor progression, MRD, and genomic heterogeneity is not clinically viable in patients. The last 5 years have seen strong pre-clincial data in the ability to open the blood-brain barrier (BBB) utilizing focused ultrasound and microbubbles to improve drug delivery across a number of small and large molecule therapeutics. This has led to multiple clinical trials underway in oncology, Alzheimer's and Parkinson's disease. One of these studies (NCT03616860) embedded liquid biopsy blood draws before and after opening the BBB in patients treated for GBM. The results demonstrated a 2.6X increase in cfDNA after opening the BBB, and a correlation of the cfDNA increase related to the tumor size & BBB opening, further the methylation analysis supported that the cfDNA visualized was malignant in nature (DOI: 10.1093/neuonc/noab057). Given the lack of clinical tissue or liquid biopsy options available in malignant brain tumors, the ability to non-invasively open the BBB provides a significant clinical opportunity to improve clinical decision making and therapeutic options for hundreds of thousands of US patients with gliomas, GBMs, or brain metastasis. Additionally, novel medical devices, such as our own Cordance device, which is painless, non-invasive, and is designed for a 30min outpatient BBB opening can enable a future where liquid biopsies may have a role in malignant brain tumors. To do so, clinical studies designed around liquid biopsy (rather than opportunistic studies) need to be developed and sponsored. Further exploration of integration and analysis between brain tumor tissue genomics, liquid biopsy, and imaging modalities needs investment. Finally, investment into studies and devices to understand the pre-analytical approaches to maximize circulating genomic biomarkers results. We look forward to an opportunity to the NIH taking a leadership role in expanding NGS towards patients with malignant brain tumors.
2. **Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)**
3. **Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)**

4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

Email: ryan.dittamore@cordancemedical.com

ID: 1768

Submit date: 2021-10-15

I am responding to this RFI: On behalf of myself

Role: Other

Role-Other: USG Contractor

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

Infectious Diseases

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**

Using modeling data methods that are more inclusive, geographically and demographically, of diverse populations to reduce disparities. Pak, T. R., & Kasarskis, A. (2015). How next-generation sequencing and multiscale data analysis will transform infectious disease management. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 61(11), 1695–1702. https://doi.org/10.1093/cid/civ670

2. **Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)**

The use of AI/ML will be a very big leap forward if done with the NGS mission outcomes in mind. Pak, T. R., & Kasarskis, A. (2015). How next-generation sequencing and multiscale data analysis will transform infectious disease management. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 61(11), 1695–1702. https://doi.org/10.1093/cid/civ670

3. **Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)**

Utilizing partnerships with academic, community, and international stakeholders to develop standards and best practices workshops. Pak, T. R., & Kasarskis, A. (2015). How next-generation sequencing and multiscale data analysis will transform infectious disease management. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 61(11), 1695–1702. https://doi.org/10.1093/cid/civ670

4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

More training and funding opportunities on data literacy at all levels is fundamentally instrumental for program support overall

ID: 1770

Submit date: 2021-10-25

I am responding to this RFI: On behalf of an organization

Name: Nathan Carrington

Name of Organization: Roche

Type of Organization: Industry (Biotech/Device/Pharmaceutical Company)

Role: Other

Role-Other: Regulatory Policy

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**

Roche recognizes the existing challenges related to NGS standardization and harmonization, and we support the concepts NIH and FDA have described here. Below, we outline some key aspects that are important to keep in mind when considering the practical implementation of the topics proposed in this section:

- We agree that the "ground truth" gap is frequently a limiting factor impeding high-quality research, development, validation, and regulatory science and that representative physical reference samples will aid in closing this gap. However, transparency will be needed with respect to the methods by which ground truth has been established with these samples and any potential biases and/or limitations that exist. It is important to avoid technology bias and ensure consistency across technological platforms when establishing ground truth or a gold standard. Further, cutting-edge technology (such as machine learning techniques) may find new genomic variants that the previous gold standard was unable to identify. Gold standards can be hard to determine due to heterozygosity and sampling biases and should therefore be distributional and connected with externally measured population statistics to ensure in-distribution comparisons. Such information will enable transparency and provide users performing ground truth comparisons with greater insights.
- Tools for data analysis, interpretation, and comparative assessments need to be flexible across platforms and, similar to the physical reference samples, provide transparency regarding limitations and biases. Further, consideration should be given regarding how such tools may interface with purpose-built tools established by NGS developers.
- Standardization with respect to infrastructure and genomic data management will facilitate data access and sharing. When establishing such an infrastructure, it is important to consider different formats and the ease of mapping between them. Pre-determined guidelines with appropriate policies/controls may be used that enable different data providers to exchange information on a pre-defined format. Incentivization strategies may be considered that encourage stakeholders to share data and solutions.

2. **Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)**

Roche agrees that access to robust, high-quality datasets and related information is one of the most significant challenges facing the routine use of AI/ML in research and development, including for the purposes of interpreting NGS data and multi-domain data. AI/ML techniques can provide innovative insights and solutions through the analysis of vast amounts of data, but this can only be achieved if the data are accessible. Current fragmentation in healthcare systems and the lack of use of standardized nomenclature and infrastructure prevents access to and use of large amounts of clinical data that could be used to advance patient and public health. Efforts are needed to ensure that the appropriate data standardization practices and infrastructure exist to enable unambiguous identification of health information within an interconnected healthcare ecosystem. Efforts such as FDA's SHIELD project (https://mdic.org/program/systemic-harmonization-and-interoperability-enhancement-for-lab-data-shield/) are critical for improving semantic interoperability and need greater recognition and implementation by laboratories, healthcare institutions, and the clinical community. We encourage NIH and FDA to consider mechanisms they can employ to ensure more widespread adoption of efforts such as SHIELD that will contribute to semantic interoperability of laboratory data, including NGS data and multi-domain data, that can be leveraged for AI/ML applications.

3. **Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)**
4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

Roche appreciates this information request from NIH and FDA and appreciates their consideration of our provided comments.

Email: nate.carrington@roche.com

ID: 1772

Submit date: 10/27/2021

I am responding to this RFI: On behalf of an organization

Name: Sasan Amini Ph.D.

Name of Organization: Clear Labs

Type of Organization: Industry (Biotech/Device/Pharmaceutical Company)

Role: Organizational Official

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

Next generation sequencing tests for infectious disease

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**

Please see the attached PDF file.

2. **Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)**

Please see the attached PDF file.

3. **Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)**

Please see the attached PDF file.

4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_ngs/uploads/IrJDHBxUAJ.pdf

Description: Comments submitted in response to the RFI.

Email: jennifer@ipolicysolutions.com

ID: 1773

Submit date: 10/29/2021

I am responding to this RFI: On behalf of an organization

Name: Daniela Kucz

Name of Organization: Palantir Technologies Inc.

Type of Organization: Other

Type of Organization-Other: Technology Company

Role: Member of the Public

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**

Palantir Technologies Inc. ("Palantir Technologies," collectively "we" or "our") is grateful to the NIH and FDA for the opportunity to provide information about existing gaps that are impeding Next Generation Sequencing (NGS) test and radiology tool development, validation, and data interpretation, including AI/ML techniques. Based on our experience with comparable efforts at both NIH and FDA, the biggest challenge to NGS test/radiology tool development and validation is the absence of a FAIR data infrastructure for clinical and translational research that reduces the burden of data sharing, integration, access, and use. The infrastructure should: 1. Enable multi-modal data integration and support large data scale 2. Facilitate resolution of data quality issues 3. Support data harmonization into an intuitive and dynamic data model 4. Protect data through technology-backed (permissions/access) controls 5. Provide users with rapid access to data and integrated analytical tools 6. Promote reproducibility by capturing user behavior and all versions of data and logic—back to the source 7. Improve data quality, breadth, and depth by connecting agencies to the point of care in a closed loop system We share lessons learned below with a focus on capabilities required for an effective and FAIR infrastructure (including tools). 1. Multi-Modal Data Integration. NIH and FDA require an infrastructure that can securely and directly sync structured and unstructured data inputs from hundreds of diverse data sources, regardless of data type (omic, EHR, imaging, etc.), scale, schema, or format—increasing the likelihood that users can access a sample size large enough to have sufficient statistical power to draw significant conclusions about relevant patient subsets, and to build cohorts representative of the U.S. population (including providing insight on underrepresented ethnic/racial groups). The sample size should also have sufficient generalizability. Further, to encourage data sharing and remove the burden on organizations contributing data, the infrastructure should be able to directly connect to source systems—as well as privacy preserving record linkage (PPRL) solutions—via bidirectional open APIs (including a FHIR-enabled API). Direct connection will not only remove reliance on intermediaries that aggregate and manipulate data and improve speed of access to and transparency into real-world data, but also offers more control and oversight to data owners (e.g., NGS test manufacturers who may have IP concerns regarding data sharing). Interoperability with other data commons or repositories via these open APIs will also reduce the likelihood of data duplication and the risk of overfitting an AI/ML model. Finally, the infrastructure

should easily accommodate new connections. 2. Data Quality Resolution. Real-world data is prone to gaps and inaccuracies, and requires continuous data quality identification, resolution, and validation. Traditionally, these processes are extremely manual and labor-intensive. NIH and FDA require an infrastructure with both code-based and point-and-click pipeline management tools, which facilitate close collaboration between data scientists and data subject matter experts (SMEs), and capabilities that enable users to pull data and propagate to downstream user artifacts on a near-real-time basis. To ensure data accuracy downstream, the NIH and FDA infrastructure should have automatic gating at the data connection and ingestion phase. Further, the infrastructure should include configurable templated checks on incoming data and metadata, which require validation before data can continue through a transformation pipeline. This decreases the chance of stale data, schema errors, and transmission issues even as data scale and complexity increase. The data infrastructure should also support data de-identification pipelines to protect data downstream. Finally, the NIH and FDA infrastructure should have a restricted central environment for data administrators, who can compare data quality across individual data providers prior to releasing it to authorized users. 3. Dynamic Data Harmonization. Data is only as useful as it is accessible to users, and NIH and FDA require an infrastructure that can link and bring together multi-modal data in an intuitive way—including facilitating PPRL-based linkage. A dynamic data model management system facilitates a broad range of users to engage with the same data foundation in common sense terms they can understand (e.g., "genome" or "patient")—empowering technical and non-technical SMEs with intuitive, data-driven tools and workflows. This data model management system should facilitate data characterization and standardization while maintaining interoperability between standards and data models (such as OMOP, TriNetX, PCORnet, or ACT/i2b2), given the lack of community consensus on data standards. 4. Configurable Data Access Controls for Compliance with Policies and Regulations. NIH and FDA require an infrastructure that can provision access to data for thousands of users across organizations and teams on a project-by-project basis. Further, this infrastructure should be backed by, and adhere to, all relevant policies to promote responsible data sharing and data use. To facilitate appropriate data access based on a user's intended operations, the infrastructure should feature a technology-backed data use and download request process that can be approved or denied upon human review. To promote data security and privacy across the infrastructure, all policies should propagate throughout without exception and be backed by audit logs verifying user behavior. 5. Rapid User Access to Workspaces with Both Data and Analytical Tools. The NIH and FDA infrastructure should provide rapid access to high-quality data—including reference datasets easily locatable via catalogs and/or tags. Additionally, this infrastructure should include a diverse selection of analytical tools—including for harmonization, analysis, interpretation, and comparative assessments—that will help researchers gain insight from a robust data foundation. To enable users with different skillsets and needs, the infrastructure should offer both natively integrated analytical tools and interoperability with third-party tools (e.g., open-source or custom-built). In particular, the native tooling should include: - Point-and-click analysis tools that allow less technical users to cohort and analyze data across genomic and clinical characteristics, generate publication-ready visualizations, and share their work and outputs with tools such as multi-modal dashboards. - Code-based analysis tools that allow users to use the latest R and Python libraries imported from GitHub and Conda to perform large-scale analysis backed by a distributed compute framework. To support reuse of work, the infrastructure should enable users to generate templatized code-based pipelines to allow others to perform these same analyses without code. - An AI/ML model management system to support more complex workflows (see Topic 2). 6. Validation and Reproducibility of Data and Analysis. When bringing

together complex data across a wide variety of sources, transparency and traceability are critical to fostering user trust, identifying attribution, and promoting the use—and reuse—of data by authorized users. Transparency and traceability enable users to understand and validate the origins of data and analysis, when it was last updated, and what transformations occurred—supporting a more efficient peer review process. Should any data quality or analytical issues arise, these features enable rapid root cause identification and subsequent resolution. The NIH and FDA infrastructure should therefore preserve transparency at all levels by tracking all actions, changes, and versions—enabling branching of not just logic but data. These capabilities enable: A. Administrators to centrally govern data and access controls, and audit user behavior when needed (assuming proper permissions) B. Users to trace data back to the source, understand work performed by others, revisit historical analyses, and test their own hypotheses and analyses in a private "sandbox" environment Once researchers have created code, datasets, or models that could be useful to other projects, they should have the ability to publish these to a centralized "knowledge store" that can support other projects. 7. Closed Loop System. Truly understanding patients requires leveraging information collected about them at the point of care. This is made challenging by the business and legal complexity of collaborating with health care organizations (HCOs) who house patient data to use it for secondary analyses. The NIH and FDA infrastructure should therefore streamline the process of collaborating with HCOs by: - Allowing HCOs to retain control over their data even as NIH and FDA bring their models and tools to that data, rather than the data egressing HCO systems - Continuously updating their datasets by maintaining provenance back to HCOs' point of care systems - Richly contextualizing feedback from NIH and FDA on data quality, breadth, and depth back to HCOs, who can then add existing data or even generate new data from biobanks in real time

2. Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)

To facilitate effective application of AI/ML to the interpretation of NGS, radiology, and other multi-domain data, we propose the following capabilities that will enable access to a powerful, trustworthy, and secure end-to-end environment for AI/ML application. Elements of this vision are already a reality at NIH and other agencies (see our response to Topic 3), and NIH and FDA should evaluate the feasibility and sustainability of leveraging existing resources. 1. Ability to curate, annotate, and pre-process data. As stated in response to Topic 1, data and the associated data infrastructure are the greatest determinants of an AI/ML effort's success or failure. AI/ML requires large, representative populations and high-quality curated and annotated data to enable the building of a generalizable model, which in turn requires the right data infrastructure. The NIH and FDA infrastructure should empower researchers to: - Easily curate and/or discover potential training datasets relevant to their use case through point-and-click and code-based curation tools and a transparent data catalog, as well as review metrics and metadata on those datasets - Collaborate on training data sets (including dataset annotation) in a user-friendly interface without compromising data lineage, integrity, or security - Facilitate data sequestration and cohorting to ensure the desired model is run on an appropriate selection of data - Branch a data set, modify it, and make those modifications available to the broader research community in a fully secure and transparent way. Branching datasets supports data sequestration for AI/ML, and users can discover, record, and mitigate bias and other data issues for the entire research community - Leverage external AI/ML tooling on top of the data asset through bidirectional open APIs and standard modeling libraries - Granularly secure data with low-friction, built-in access control tooling. Not all datasets should be fully accessible to all users (e.g., to protect sensitive information) and oversight is

required to assess the potential consequences of combining data sets 2. Ability to train, test, and retrain adaptive AI/ML models. The NIH and FDA infrastructure should provide researchers with the development environment, computational power, and tools required to train adaptive AI/ML models in a streamlined manner, or enable them to use their own tools, connected via open APIs. Users should be able to easily capture and share their work products (libraries, models, data modifications) back to the infrastructure, enabling knowledge to compound over time. A technical infrastructure designed to capture and share knowledge enables appropriate contextualization of existing data and research for faster onboarding to and use of the AI/ML environment. 3. Ability to monitor and evaluate real-world AI/ML model performance within the NIH and FDA infrastructure. Researchers should be able to compare models to baseline standards, view and capture metrics that show their AI/ML models' performance and capture performance against different evaluation datasets. Researchers should have the option to make this performance data discoverable so that others can learn from their work, while also providing provenance and securing their data so that it is only accessible by those with appropriate permissions. 4. Ability to contribute back to the NIH and FDA infrastructure. User friendly tools should be provided that incentivize researchers to contribute back to the broader knowledge base via a centralized and easily accessible platform. For example, researchers could make new and improved data pipelines, data annotations, model templates, analysis frameworks, etc. available to fellow researchers. Incentives such as additional compute allocations that encourage contributing knowledge back to the scientific community would compound the collective knowledge of the NIH and FDA infrastructure over time and streamline future research efforts. In turn, this ability to broaden the knowledge base will provide richer data sources to serve as the foundation for AI/ML applications.

3. Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)

N3C Data Enclave: One existing resource that fulfills the capabilities outlined above is the NIH National Center for Advanced Translational Sciences-backed (NCATS) National COVID Cohort Collaborative (N3C) Data Enclave, which was mentioned during the NIH and FDA virtual workshop as a viable repository to fill resource gaps. The N3C Data Enclave is backed by the Palantir Gotham Platform configured with Foundry ("Palantir Foundry"). Access to the N3C Data Enclave can be requested here: https://covid.cd2h.org/enclave. While the scope of research supported by the N3C Data Enclave is currently limited to COVID-19, the Enclave is dynamic—containing EHR data from across 65+ academic medical institutions for more than 8M patients (including 2.9M COVID-19-positive patients), across 9B rows of data. The COVID-19 data asset in N3C exclusively contains SME-validated data, enabling researchers to devote 100% of their time to conducting research instead of spending time doing data cleaning and harmonization. While the Enclave can support a wide variety of biomedical and other data types (including omics data), current data types include patient clinical data and outcomes (including lab results, medications, procedures, and visits) at a high level of characterization (including summary statistics, cohorts, and individual data) as well as DICOM imaging. It includes a library of 500+ reference sets for research concepts (e.g., lab measurements for a particular test or diagnoses for a particular condition), which have undergone rigorous review by leading clinicians. Enclave data continues to be expanded and updated on a weekly basis. Additionally, a proof-of-concept PPRL to TCIA has been completed and PPRL linkages to claims data as well as mortality data are in progress. Additionally, PPRL linkage to MIDRC radiology data is anticipated. Access Model. The N3C Data Enclave access model is purpose-based; namely, researchers receive access to data according to their research needs instead of receiving blanket access at a dataset level. Upon being granted access to the N3C Data Enclave, users—

logging in using their institutional credentials, as made possible by Palantir Foundry's integration with institutional authentication systems—must request access to specific N3C datasets at a given tier of information (e.g., synthetic, de-identified Safe Harbor, or HIPAA Limited Dataset data). Users make the request and provide justification using an in-platform configured form. The N3C Data Access Committee then receives automated notice of each request submission and approves or denies the request in the platform. Palantir Foundry powers true purpose-based access—and should the researcher require this same data for another project, they must submit another access request. Data Use. The N3C Data Enclave could support additional disease areas beyond COVID-19 if authorized to do so by NCATS and data contributors. Critically, Palantir Foundry's purpose-based access controls allow NCATS to enforce and audit conformance with Data Use Agreements. NIDAP: Another existing resource offering capabilities to close existing gaps is the NIH National Cancer Institute (NCI)-funded NIH Integrated Data Analysis Portal (NIDAP), also backed by Palantir Foundry. NIDAP provides NCI with a research-centric, open data infrastructure that connects systems across NCI (including High Performance Computing Data Management Environment [HPCDME], NIH's Biowulf High Performance Computing [HPC], BTRIS and LabMatrix clinical sources, lab-specific share drives, and imaging storage and analysis software). NIDAP is a key part of the data modernization strategy at NCI and is used by 50+ PI groups and 800+ users. NIDAP connects widely used systems across NCI and enables researchers to access and combine high-compute analysis, imaging, clinical, and genomic data for the first time. NIDAP enables the reuse and development of logic (e.g., bioinformatics pipelines) and data resources (e.g., aggregated genomic sequencing data), which serves as the connective tissue for NCI from data storage to analysis and publication. NIDAP supports a wide variety of primary data types, including: - Omics data (e.g., NGS and non-NGS, including WES, RNA-seq) - Radiology and pathology imaging data (e.g., MRI scans, tumor slides) - Clinical data (e.g., clinical procedures, clinical tests and measurements, clinical outcomes [survival/death]) Access Model. All data in NIDAP is access controlled based on NCI's direction. NIDAP propagates all NCI-configured data permissions throughout the platform, and across all user groups—enabling open access to general NCI data and tools while gating access to specific subgroups for select workflows and tools (e.g., a clinical research laboratory information management workflow for the Laboratory of Pathology or a patient-centric workflow for the Urologic Oncology Branch). These granular permissions secure all data at all times, including Personally Identifiable Information (PII) and Protected Health Information (PHI). These granular access controls enable researchers to securely share project-specific data with collaborators across the NCI, at other ICs, and (upon request) with extramural institutions. Data Scope and Characterization. Specimen data integrated in NIDAP stems from tens of thousands of subjects, and include associated whole exome, whole genome, bulk, and single-cell RNA sequencing, epigenetic (e.g., methylation), other omics data, MRI imaging data, pathology slides and imaging analyses, and tumor samples. Samples were collected in accordance with NCI-specified research protocols, and in-platform access controls configured with NCI automatically dictate which users have permission to access and use what data. The integrated data asset can be analyzed and characterized at the individual sample, source subject, or cohort/population level. Data is updated and integrated automatically, ranging from every five minutes to nightly depending on user needs, and continues to expand to support new types with plans to include microbiome, animal modeling, and high throughput natural products screening data. Data Linkage. NIDAP dynamically links sequencing, imaging, and other data modes together through the platform's configured ontology—a data model configured based on consultation with NCI SMEs. Data in NIDAP includes clinical outcomes (disease manifestation and progression, tumor growth rates, treatment response, etc.), all of which can be linked back to genomic

variants and other inputs through the ontology. NIDAP thus facilitates the performance of meta-analyses of multi-modal data (e.g., to identify all cancer patients in a certain age range with specific clinical outcomes and test values that also show positive expression of gene X) that would be otherwise be difficult or impossible. In addition to one-off analyses, NIDAP also facilitates multi-modal model creation and management on top of current data. MITRE mAbs DCP: A third resource that is a Closed Loop System (see response to Topic 1) is the MITRE monoclonal antibodies (mAbs) Data Collaboration Platform (DCP), also backed by Palantir Foundry. The MITRE Health FFRDC (on behalf of its sponsor, ASPR) is conducting an RWE study on the effectiveness of monoclonal antibody treatments for COVID-19 using the DCP, in coordination with four Health System partners. Of particular interest to MITRE and its sponsor is the effectiveness and impact of monoclonal antibody treatments in the face of SARS-CoV-2 viral evolution. The DCP facilitates inquiries on this topic through bidirectional communication between MITRE and its Health System partners: 1. MITRE analyzes deidentified clinical demographic and outcomes data from Health Systems to identify patients for which it would like COVID-19 RNA-Seq data. 2. MITRE communicates the cohort of patients of interest back to its Health System partners, using Palantir Foundry's capabilities to maintain provenance. 3. The Health System reidentifies those patients within their own commercial instance of Palantir Foundry (made available to the Health System under a Business Associate Agreement) and identifies corresponding COVID RNA samples to be sequenced. 4. The Health System ingests COVID RNA-Seq data into their commercial instance of Palantir Foundry, links it to the preexisting clinical data, deidentifies the entire dataset, and pushes it into the MITRE instance of Palantir Foundry.

4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

In our response above, we underscore that, in our experience, data—its quantity, its quality, and its appropriateness to the problem—is the greatest determinant of a data-related effort's outcomes. We strongly urge NIH and FDA to center its strategy—to enable NGS test development, validation, and data interpretation as well as radiological tool development and clinical data interpretation using AI/ML—around an infrastructure to support data quality.

Email: dkucz@palantir.com

ID: 1774

Submit date: 11/1/2021

I am responding to this RFI: On behalf of an organization

Name: Ezekiel Maier

Name of Organization: Booz Allen Hamilton

Type of Organization: Other

Type of Organization-Other: Consulting Firm

Role: Other

Role-Other: Support of Government AI/ML and Bioinformatics Initiatives

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

Advancing the use of artificial intelligence (AI) and machine learning (ML) for the analysis and interpretation of genomics and other health data

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**

Accelerating the use of artificial intelligence (AI) and machine learning (ML) for the interpretation and analysis of health data, particularly genomics data, can drive enormous benefit for human health and disease. AI and ML can help researchers extract actionable insights from health data, particularly if an analytics ecosystem encourages 1) use of interpretable technologies, 2) adoption of transparent procedures, and 3) opportunities for broad cross-pollination of embedded experts. However, gaps and open questions about accessibility of real-world data and the transparency of AI/ML models may pose significant challenges to realizing these opportunities. Below we highlight critical gaps, provide context, suggest mitigation strategies, and list relevant Booz Allen experience. Gap: Lack of readily available, real-world biomedical data for development of AI/ML models. Regulatory guidelines such as HIPAA (e.g., the Privacy Rule) aim to enforce controlled access to de-identified data. However, it is possible to re-identify and/or reveal sensitive health attributes from de-identified biomedical data. Therefore, these protective measures are insufficient to safeguard individual privacy. Moreover, uncertainty about regulatory guidelines and Institutional Review Board (IRB) review processes and paperwork that are critical for safe and ethical human subjects research, may thwart the pace and breadth of data sharing. Solution(s): To meet this challenge, NIH and FDA should take advantage of the opportunities provided by privacy-preserving data mining (PPDM) techniques, methods that quantify and protect the trade-off between data utility and subject privacy. For example, Federated Learning is a privacy-preserving data mining technique that allows data to be available in a collaborative, accessible environment, while remaining secure in its original server. Federated learning works by training a machine learning algorithm on multiple local datasets contained in local nodes without explicitly exchanging data samples. A similar approach known as model-to-data enables the release of synthetic data for AI/ML model development, while withholding sensitive data in a secure private computational environment for model evaluation. Finally, differential privacy is a method in which a small amount of noise is added to the data in order to

conceal the exact datapoints that comprise a specific dataset. Privacy-preserving data mining infrastructure development could enable 1) researchers to share more diverse datasets responsibly and widely, 2) regulators such as FDA to enhance the review of AI/ML-based software as a medical device without moving or exposing proprietary data used to train and validate the algorithms, and 3) allow central aggregation of data that could potentially accelerate AI/ML model training. Federated learning would minimize concerns about proprietary data breach by pharmaceutical firms, decrease overhead required to manage training and validation data submitted to the FDA, and ultimately accelerate regulatory review. For example, in the case of bioinformatics and advanced analytics tools developed and submitted for regulatory review, federated learning would enable FDA to better interpret sponsor-generated data by running the NGS companion diagnostic model. Moreover, federated learning or a model-to-data architecture could enable researchers to train AI/ML models on real-world biomedical data to produce more applicable and ethical analytical tools. Booz Allen Experience: Booz Allen has a long track record of supporting federal partners in the biomedical research community with responsible and private data sharing and AI/ML development. For example, in partnership with a military health agency, Booz Allen developed a secure and scalable cloud-based platform for genomic data management and analysis. Booz Allen evaluated the technology landscape of PPDM techniques, collating published information about efficacy, trade-offs, and feasibility of implementation. Currently, Booz Allen is supporting a federal biomedical research agency to examine and benchmark existing commercial off-the-shelf (COTS) privacy-preserving record linkage (PPRL) technologies to enable integration of different data types (e.g., clinical and genomics) in a more private way. Gap: Lack of robust systems and norms to maximize transparency. Transparency in human research creates trust, particularly between research participants and researchers. Proactive communication about what data is collected and how it will be used as well as giving research participants increased control of their data builds upon that trust and reciprocity helps to maintain that trust. If gaps in research and AI/ML modelling transparency, including the collection and reporting of methods, assumptions, and data, go unaddressed, findings can be gravely distorted through selection bias, in which subpopulations decline to participate in the research enterprise for fear of consequences. Improving and maintaining trust requires novel mechanisms to support dynamic consent of subjects, increased transparency of datasets (e.g., origins, modifications, and processing), and increased explainability of ML models (e.g., clarity, interpretability, validation-readiness). Solution(s): Technologies that support standardized curation of metadata and automated logging of data modifications, such as Apache Atlas or Apache Taverna, should be integrated into organizational workflows. Whenever possible, industry standards, data ontologies, and best practices for data formats should also be used. The Global Alliance for Genomics and Health (GA4GH), a technical standard-setting organization for genomic data, has published a structured metadata tool, Automatable Discovery and Access Matrix (ADA-M), and Data Use Ontology (DUO) standards as well as consent codes that allow users to tag datasets with usage restrictions. Both standards can be used to support dynamic consent, a mechanism to enable different types of mutable research participant or patient consent. Altogether, these policies will allow for clear and transparent data usage, labor-saving automation, and streamlined technical interoperability within the larger healthcare ecosystem. Booz Allen Experience: Through our support of the military health agency cloud-based genomics platform, we implemented the use of GA4GH consent codes, developed through GA4GH study of data use restrictions. These codes capture the conditions of data use and sharing as metadata. Mapping to GA4GH consent codes facilitates easier data access by relieving the burden of reviewing detailed consent documents per patient and study. The uploading researcher manually

assigns consent codes for the data they have uploaded based on the subject consent form. In addition, we organized and ran the BioCompute Object (BCO) app-a-thon on a federal cloud-based platform for advancing precision medicine and informing regulatory science to encourage the adoption of the FDA-recognized BioCompute specification for reproducible documentation of computational workflows.

2. Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)

Gap: Discoverability and Usability of Developed AI/ML Models It is challenging to locate and compare individual AI/ML applications for the analysis of omics, imaging, and health record data. Data scientists cannot use a tool without supporting documentation, such as versioning, performance metrics, and training approaches and data. However, the tools themselves are rarely packaged with the information required to make them useable. Results from AI/ML studies are often published in scientific journals, while the AI/ML tools that were used for discovery are often only available on individual laboratory websites rather than a centralized location. This approach makes the tools difficult to access, and performance difficult to reproduce, since there is no direct linkage of the established AI/ML model to the description and performance metadata provided in the publication. This challenge can result in a significant barrier to entry for scientists to begin using a tool, causing lost time searching for or recreating documentation or, even worse, not using a tool that would have faster and more accurate results. Solution(s): A public ModelOps and MLOps platform would enable centralized open sharing, use, and enhancement of AI/ML applications for analysis and interpretation of NGS and other real-world biomedical data. Such a platform would ease discovery, management, and deployment of containerized AI/ML algorithms. Within the platform AI/ML models should be paired with human readable/understandable metadata to describe the model, identify the process and datasets used to train and validate the model, metrics for model performance, and instructions for executing the model, including details such as hardware recommendations and data requirements. A cloud-based ModelOps platform also has the advantage of facilitating collaborative research within and across groups not only by providing a shared repository for tools and data, but also a standardized environment for execution and benchmarking. Booz Allen Experience: Booz Allen has demonstrated success in the establishment of AI/ML development platforms. One such platform called Modzy, which is described in greater detail in Topic 3, enhances developed AI/ML model portability, scalability, and discovery. In addition, Booz Allen supports community engagement on a federal cloud-based biomedical informatics platform, which offers more than 130 publicly available bioinformatics and AI/ML applications to empower the community of experts to advance precision medicine and inform regulatory science. Gap: Availability of Linked Multi-Modal Real-World Data Representative of Diverse Populations The need for accessible real-world data is critical because AI/ML modelling is powered by large volumes of diverse high-quality data. Linked omics, medical images, and health records, collected from the same human subjects is necessary to enable integrative advanced analytic analysis that can discover hidden patterns and novel biomarkers. These novel discoveries can lead to new diagnostics and treatments for improving human health. Moreover, linked data representative of diverse populations is needed to ensure AI/ML models are not overfit to better characterized populations. Currently, such available large-scale representative linked multi-modal real-world datasets are lacking in biomedical research community. One workaround approach often utilized by the community is to leverage genomics data from a reliable reference data set, like the 1000 Genomes Project, paired with generated synthetic phenotypic data. However, generated synthetic phenotypes are often not representative of the underlying molecular mechanisms.

Solution(s): Capabilities for generating synthetic healthcare record data have advanced significantly with generative adversarial networks (GANs). GANs, a deep learning algorithm, have been used to generate new synthetic data that maintains the statistical properties of the real-world data used for training. In addition to synthetic health record generation, recent publications have explored the use of GANs to generate synthetic human genomes. An article in PLOS Genetics (https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009303) found promise in "artificial genomes" that retained the complexity and relationships of the original data. GANs can be deployed on NHLBI TOPMed and NCI Data Commons data to ensure diverse sources and types of synthetic data. Continued advancement of GAN-based synthetic real-world data generation, and the availability public APIs for on-demand generation of linked multi-modal synthetic real-world data will accelerate AI/ML model development and benchmarking. Booz Allen Experience: Booz Allen has explored the approaches, tradeoffs, and applications of synthetic data generation in collaboration with the Veterans Health Administration (VHA) and National Institute of Standards and Technology, documented in a Towards Data Science post (https://towardsdatascience.com/synthetic-data-at-the-vha-8124989c7183). Booz Allen has also organized and run 4 public bioinformatics and AI/ML challenges on a federal cloud-based platform for advancing precision medicine and informing regulatory science that utilized synthetic data, including one modeling risk factors for COVID-19. In this challenge, participants developed AI/ML models to identify risk and protective factors for severe COVID-19 illness and predict COVID-19 outcomes for a large cohort of synthetic Veteran health records. During the post-challenge phase, Booz Allen and its government partners are evaluating the performance of winning models on two synthetic datasets generated by different tools, and on real de-identified Veteran health records. This analysis will highlight the strengths and limitations of training AI/ML models on synthetic health record data. Gap: AI/ML Training and Development Engagement and Education As AI/ML platforms and algorithms become increasingly available, far-reaching benefits to NGS analysis and interpretation could be achieved via an increase in AI/ML adoption activities and training for the biomedical informatics workforce. AI/ML is crucial to the extraction of large volumes of linked genomic, imaging, and electronic health record data, as well as the downstream integrative analysis, but knowledge of what tools are available, how to use them, the advantages and limitations of using popular tools, and appropriate interpretation of the results is still needed in the workforce to get widespread, meaningful benefit to the research community. Solution(s): We recommend the development of a publicly available flexible AI/ML knowledge assessments and training tailored to the specific needs and workflows of the biomedical informatics workforce. This should include personalized data science training, which emphasizes analytics skills and knowledge needed for specific roles, and events such as cross-cutting symposia, lunch seminars, online training/forums, and hackathons. These events provide the workforce with opportunities for continuous training and de-silo experts. In addition, advanced training should be offered to empower the use of cutting-edge AI algorithms, such as novel deep learning approaches, and specialized computing infrastructure, such as graphics processing units (GPU) and field-programmable gate arrays (FPGA). Use of this hardware can significantly accelerate model training and inference to make research possible that would previously have likely been considered too resource-intensive or time-consuming. For example, GPU accelerated deep learning models are being used to improve genetic variant calling (e.g., DeepVariant) and associate omics markers with features in medical images (e.g., radiogenomics). Wider exposure and knowledge of these advanced approaches will enable data scientists and researchers from disparate backgrounds to develop novel pipelines to better extract complex patterns from large data sets. Booz Allen Experience: Booz

Allen partnered with a federal biomedical library to design and implement a data science training program to build a workforce for data-driven research and health. Staff were trained on topics ranging from high-level overviews to 120-hour deep dives. Assessments of data science expertise were gathered prior to and after the trainings to identify impactful topics for continued learning and development. Booz Allen also developed a "Data Academy" for a federal public health agency, utilizing commercial offerings as well as our own novel content to deliver learning pathways via self-paced, on-demand e-learning. This was nominated for an agency-wide award. Booz Allen has also trained federal health clinicians on kidney disease prediction AI/ML models we developed to further patient health, patient care, and patient-centered outcomes research.

3. Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)

As discussed in topic 1, there is a lack of readily available, real-world biomedical data for development of AI/ML models. The lack of clear and effective standardized methods for de-identification of real-world data has limited the sharing and accessibility of linked omics, medical imaging, and health record data that is necessary to advance the application of AI/ML to the interpretation of biomedical data. Nvidia has open-sourced a standalone Python library, Nvidia Federated Learning Application Runtime Environment (NVFlare) (https://nvidia.github.io/NVFlare/), that provides a framework to easily deploy and operate a federated learning environment for AI/ML model development. The federated learning environment deployed by NVFlare ensures data is protected through at least three mechanisms: 1) secure client server communications by creating an explicit and constrained network among the constituent institutions, 2) using differential privacy techniques to exchange only a subset of model parameters during each update rather than sensitive data, and 3) using homomorphic encryption to allow computation on encrypted exchanged model parameters. This Python library is distributed under the Apache License 2.0 which enables commercial use, distribution, and modification. As strategic partners, Booz Allen and Nvidia's strong ties enable us to enhance and deploy NVFlare to advance the application of AI/ML to the interpretation of real-world biomedical data via federated learning. NIH and FDA can utilize the NVFlare Python library to deploy federated learning environments to empower research and enable regulatory review using sensitive biomedical data. As discussed in topic 2, discovery, reuse, and enhancement of developed AI/ML models for analyzing and interpreting omics, medical imaging, and electronic health record data is difficult because models and metadata are often managed separately and dispersed to different scientific journals and laboratory websites. To improve AI and ML model management, governance, and discoverability, Booz Allen developed Modzy (https://www.modzy.com), a platform to accelerate AI/ML adoption and deployment. Modzy offers 1) a common model marketplace of vetted, open-sourced AI/ML algorithms; 2) an AI/ML collaboration environment where algorithms can be rapidly shared and enriched across permissioned users; and 3) an AI/ML governance environment to seamlessly deploy, use, and monitor models. To improve discoverability, transparency, and reuse, models managed by Modzy are tagged with relevant metadata including a description, training and validation procedures and data, and performance metrics. An example AI model managed in the Modzy marketplace is DarwinAI's COVID-NET model, which is a convolutional neural network that uses chest X-rays to predict the likelihood of the patient being infected with COVID19 pneumonia, regular pneumonia, or being healthy. NIH and FDA can accelerate the application of AI/ML to the interpretation of NGS data and multi-domain data by making models and model metadata more discoverable, transparent, and reusable by providing Modzy as a solution for managing models.

4. Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)

Booz Allen Hamilton (Booz Allen) is pleased to submit this response to the National Institutes of Health (NIH) and U.S. Food and Drug Administration (FDA) for the Critical Resource Gaps and Opportunities to Support Next Generation Sequencing (NGS) Test Development Request for Information (RFI). Our response summarizes critical gaps and our capabilities for advancing the analysis of NGS data, including the use of AI/ML. Booz Allen is the largest provider of U.S. public sector AI services with approximately 30% of market share. In addition, we hold the largest AI-focused contracts in the Department of Health and Human Services (HHS) and Department of Defense (DoD). We have more than 120 AI engagements across the federal government, which includes 43+ AI engagements across civilian federal agencies. We hold 63 patents in AI, ML, and deep learning. In addition to our AI expertise, Booz Allen has a wealth of bioinformatics experience, having supported numerous large omics research programs and initiatives across the federal government. Since 2011, we have supported a national genomic medicine research program run by a federal healthcare agency. Over 850,000 research subjects have enrolled in this program which has discovered novel biomarkers of many diseases impacting veterans, including post-traumatic stress disorder (PTSD) and kidney disease, through the integrative analysis of genotypes, longitudinal health records, and behavioral health surveys. Booz Allen provides a broad range of support to the research program, including data standardization and harmonization, bioinformatics workflow development, and analytics. In support of a military health agency, Booz Allen developed a secure and scalable cloud-based platform for genomic data management and analysis. This platform provides capabilities for storing, processing, and sharing genomic data, and integrating health outcomes data in predictive models. Since 2017, Booz Allen has supported a secure, collaborative, cloud-based platform for advancing precision medicine and informing regulatory science. Our team was responsible for boosting engagement and increasing the scientific impact of the platform by organizing public crowdsourcing competitions which incentivize novel algorithmic development and provide an independent source for benchmarking tool performance. Finally, Booz Allen is performing several genome-wide association study analyses in support of a federal biomedical research agency. These analyses utilize genotyping, metabolomic, epigenetic, gene expression, and phenotypic data to identify novel biomarkers. As a partner of both NIH and FDA for more than 15 years, we bring the mission knowledge, expertise, and understanding required for addressing critical gaps to advance the use of AI/ML for the analysis and interpretation of genomics and other health data.

Email: Maier_Ezekiel@bah.com

ID: 1777

Submit date: 11/1/2021

I am responding to this RFI: On behalf of an organization

Name: Jeremiah McDole

Name of Organization: Bio-Rad

Type of Organization: Industry (Biotech/Device/Pharmaceutical Company)

Role: Other

Role-Other: Marketing

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**

In light of the currently stated situation, "The lack of well-characterized and widely available somatic and germline samples makes NGS test and methodology validation across laboratories difficult" we propose the launch of an NGS proficiency testing program. While the College of American Pathologists (CAP) has long offered valuable proficiency testing, the program we propose would be modeled after what is currently implemented by the American Society for Histocompatibility and Immunogenetics (ASHI) for applications such as chimerism and engraftment monitoring. Blinded samples distributed to participating laboratories would be characterized via ddPCR in order to establish a "ground truth" measurement of genetic target frequency given the superior sensitivity, precision, and reproducibility of this platform vs. NGS. Samples would be sent to labs for testing via their NGS system(s). NGS results received back from participating labs would be set against their respective ddPCR benchmarks. A combined readout of all de-identified lab results would generate a data range to provide labs a view of +/- (or NC) "drift" from the established benchmarks. This view would provide a lab with the opportunity to implement self-correction, if needed. This approach has merit given synthetic reference sample providers, such as SeraCare, utilize ddPCR to create their NGS reference materials due to the previously stated high sensitivity, precision and reproducibility of the ddPCR platform vs. NGS. The type of sample(s) to be distributed to labs for proficiency testing may be a complex issue. Solid tumor specimens, for example are often rare, precious, and importantly, heterogeneous. To better replicate the heterogeneity found within real clinical samples, synthetic samples could be modeled after one or more well characterized clinical samples. To create such a product, a company/ organization/ investigator could perform "high resolution" genetic mapping of multiple tissue slices/slides taken from a clinical sample using both NGS and ddPCR to capture breadth of mutations presence and depth of genetic mutation frequency. These genetic mutations and their precise frequencies can be recreated per clinical tissue slice/slide and then incorporated into the above stated proficiency testing schema. Such an exercise could, among other things, reproduce diagnostic "edge cases" and lead to a better understanding of how reliability labs can make correct calls when presented with these challenging samples. Depending on the outcome, data from these edge case testing results may point to the need for greater utilization of ddPCR for reflex testing.

2. **Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)**
3. **Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)**
4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

Email: Jeremiah_Mcdole@bio-rad.com

ID: 1779

Submit date: 11/1/2021

I am responding to this RFI: On behalf of an organization

Name: Dr. Giovanna Bucci

Name of Organization: Palo Alto Research Center

Type of Organization: Industry (Biotech/Device/Pharmaceutical Company)

Role: Investigator/Researcher

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**
2. **Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)**

The Palo Alto Research Center, Inc (PARC), in collaboration with the Massachusetts Institute of Technology (MIT) identifies an area of intense research interest in the field of single-molecule protein sequencing. Whole-proteome sequencing and profiling of the vast repertoire of cell types is key to enhance fundamental understanding of living systems. Advances in hybrid and domain-aware AI (e.g., where geometric and physical modeling meet machine learning and data analytics) have opened unprecedented opportunities in basic science and medical diagnostics/therapeutics. Since its first demonstration, nanopore sensing has dramatically advanced, ultimately achieving the goal of single-molecule DNA sequencing. This technique has the potential to serve as a generic tool for the analysis of biomolecules, including proteins. Nanopore-based protein sensing is in its infancy, facing challenges unique to proteins and proteomics. Proteins span a large range of sizes and have a stable three-dimensional folded structure. In contrast to nucleic acids, the backbones of peptides are not naturally charged, complicating the possibility of electrophoretic threading them into nanopores. In addition, proteins are composed of combinations of 20 different amino acids instead of 4 nucleobases, further complicating the task of relating the ionic current signals to the amino acid sequence. Long-term opportunities for high-throughput, single-molecule sequencing range from studying the causes of neurological diseases like Parkinson's and Alzheimer's to earlier diagnosis and more effective treatment of cancer. Protein nanopores have shown promise for identifying amino acids and post-translational modifications (PTMs). For example, Ouldali et al. recently showed that 13 of the 20 standard amino acids are distinguishable based on their current signals using an aerolysin nanopore [1]. The detection of PTMs, which serve as biomarkers of cell states and diseases [2-3], has also been achieved with nanopore sensors [4-8]. Controlling protein translocation through the sensor remains a significant challenge. Initial studies demonstrate a protein-tethered oligonucleotide being captured by the nanopore and unfolding because of the pulling force. The charged oligonucleotide respond to the electric field and drives the protein through the nanopore via electrophoresis. This method, however, generates exceedingly fast translocation events (

3. **Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)**

4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_ngs/uploads/GiYIsOtAsz.pdf

Email: gbucci@parc.com

ID: 1780

Submit date: 11/1/2021

I am responding to this RFI: On behalf of an organization

Name: Anant Madabhushi

Name of Organization: Case Western Reserve University

Type of Organization: Academic Institution

Type of Organization-Other:

Role: Investigator/Researcher

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

artificial intelligence, radiomics, pathomics, computational imaging, precision medicine

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**

Integration of quality control and assessment approaches into the data curation process rather than simply "dumping" a set of data for public use Reference patient cohorts to be able to test harmonization and correction approaches for each modality within radiology (MRI, CT) etc as well as for different acquisition types Making it easier to share curation and annotation efforts for public repo cohorts - maybe leverage OHIF etc and encourage investigators to have a one-click option to make their annotations or curated cohorts available within a large pool. Else each group using a dataset ends up reinventing/redoing the same wheel. Focus on making datasets and clinical variables available for non-cancer, non-HLB diseases as well. Need for individual institutes within NIH to make datasets generated though NIH funded projects publicly available

2. **Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)**
3. **Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)**

Need for multimodal, multiscale datasets - Cooperative Oncology groups to make clinical trial datasets (imaging) more easily available to the research community - Slide scanning of previously accrued clinical trial cohorts (cooperative oncology group conducted clinical trials)

4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

ID: 1783

Submit date: 11/1/2021

I am responding to this RFI: On behalf of an organization

Name: Jesse Tetreault

Name of Organization: Nvidia

Type of Organization: Industry (Biotech/Device/Pharmaceutical Company)

Role: Investigator/Researcher

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

Accelerated Computing; GPU Computing; Accelerated Genomics Analysis

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**
2. **Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)**
3. **Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)**

In clinical genomics, deep learning algorithms are used to process large and complex genomic datasets. The tools and infrastructure is targeted towards tasks that are impractical to perform using human intelligence and error prone when addressed with standard statistical approaches. The availability of large datasets for training like large functional genomics datasets, in conjunction with advances in AI algorithms and in the GPU systems used to train them, is driving a surge in productivity. A turnkey software tool designed for enabling clinical research in NGS is NVIDIA Clara Parabricks. It delivers powerful acceleration to primary, secondary, and tertiary analyses of genomic data. Based on GATK, Clara Parabricks gives unmatched secondary analysis performance and throughput. AI systems are being increasingly used in various fields within clinical diagnosis and NGS can benefit greatly from the recent advancements in AI and ML. One example can be the use of computer vision techniques for identification of functional regulatory elements i.e. recurrent motifs in DNA sequences in the human genome. This implementation is analogous to how pixel patterns are detected in images by convolutional neural networks. Many AI and ML techniques have also been adapted to address the steps involved in clinical genomic analysis - including variant calling, genome annotation, variant classification, and phenotype-to-genotype correspondence. Standard variant-calling tools are prone to systematic errors. NVIDIA Clara Parabricks has Google's DeepVariant tool that uses a deep neural network trained directly on read alignments which outperforms standard tools on some variant-calling tasks.

4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

Email: jtetreault@nvidia.com

ID: 1784

Submit date: 11/1/2021

I am responding to this RFI: On behalf of an organization

Name: Dr. Sookyung Kim

Name of Organization: Palo Alto Research Center, Inc. (PARC)

Type of Organization: Industry (Biotech/Device/Pharmaceutical Company)

Role: Investigator/Researcher

Domain of research most important to you or your organization (e.g. cognitive neuroscience, infectious epidemiology):

1. **Development of reference samples, tools and infrastructure for clinical and translational research using NGS (limit: 8000 characters)**
2. **Application of AI/ML to the interpretation of NGS data and multi-domain data (limit: 8000 characters)**

Precision medicine aims to design and optimize the pathway for diagnosis, therapeutic intervention, and prognosis using large multidimensional biological datasets that capture individual variability in genes, function, and environment. However, capturing the variability of an individual genetic structure requires an accurate assessment of 3D structure from available 1D protein sequences of individual genes. Specifically, predicting protein folding and structural genetic mutation are critical in the drug design for personalized treatment of cancer patients as there are 10,000-100,000 mutations that can be simulated. It is therefore important to maximize the efficiency of the treatment on an individual basis. A valuable way to address cancer treatment is to correlate the disease with the structural genetic mutation of each patient. Understanding the 3D structure of proteins also is a critical yet challenging task in the drug discovery process. When a drug molecule binds with the target protein, the form of their 3D structure plays a vital role in determining the binding affinity between the two. Another challenge arises because many protein-protein interactions involve conformational changes (structural changes) upon binding with its partner or even the coupled folding or refolding of disordered segments. Moreover, in the drug-target binding scenarios, the disordered protein segments are frequently involved in transient protein-protein interaction. In this case, both the structural prediction of the protein-protein complex and binding affinities become more complicated. Despite the progress of experimental crystallography in recent years, the experimental determination of the 3D structure of protein continues to be a challenging task. It remains impossible to experimentally determine all proteins in a cell. As 3D protein structures are not available, conventional structural-based drug design processes mostly rely on random high-throughput screening. Such methods scan protein datasets and collect multiple molecules to test their binding affinity with a given target protein based on structural similarity between molecules in the database and the target drug molecule. The hit rate for such methods is poor: e.g., for one thousand scans performed on a DNA-related dataset, the hit rate was found to be 0-0.01%. An important prerequisite in the process of structural drug discovery and precision medicine is the availability of an accurate and reliable modeling approach for the 3D protein structure and eventually the availability of co-crystal structures (i.e., the structure of the crystallized drug-target complex). The dramatic advent of

high-performance computing and artificial intelligence has opened doors for the new drug discovery paradigm and precision medicine. Data-driven structural prediction of protein structure using AI has the potential to make breakthrough impacts in the drug discovery process and could save the U.S. medical and pharmaceutical sectors up to $100 billion per year. AI-driven protein structural prediction is making breakthroughs and changing the paradigm of the drug discovery process, however, most supervised learning-based machine learning algorithms suffer from the lack of interpretability of models and clinical data transparency issues of transparency of clinical data. When it comes to something as critical as developing a new pharmaceutical, AI cannot be a black box that gives answers that cannot be verified or interrogated. The ML model must be developed to provide reliable accuracy and human-level explainability in making predictions. We also need transparency of the genetic data that the model uses (e.g., the source, distribution, and methods used to pre-process data). Therefore, the end-users of AI algorithms, such as healthcare professionals or physicians, should easily understand how AI tools operate and reach conclusions before they practice AI tools in clinical environments. Because AI and ML research is a breathtakingly fast-growing area, there should be strategic and government-wide programs to integrate AI with structure-based drug discovery, taking into account ethics, diversity, equity and inclusion, and clinical practice. NIH as an organization is in the best position to lead such a program. For the research prototyping of AI-based drug discovery, PARC is interested in developing data-efficient 3D protein structure prediction models using cutting-edge algorithm development from natural language, third-wave AI techniques, and with a robust, interpretable analysis system.

3. **Existing resources that could be leveraged to fill resource gaps (limit: 8000 characters)**
4. **Any general comments related to critical resource gaps and opportunities to support NGS test development and validation (limit: 8000 characters)**

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_ngs/uploads/bsGgmHfvSi.pdf

Description: PARC AI-enabled-drug-discovery

Email: sookim@parc.com