Docket No: OMB-2019-0003

Alliance for AI in Healthcare (AAIH)
1340 Smith Ave, Suite 400
Baltimore, MD 21209
Tele: (410) 779-1245

August 9th, 2019

Subject: AAIH Comments on "Identifying Priority Access or Quality Improvements for Federal Data and Models for Artificial Intelligence Research and Development (R&D), and Testing: Request for Information"

Dear OMB and Executive Office,

The **Alliance for Artificial Intelligence in Healthcare (AAIH)** welcomes the opportunity to comment on the Office of Management and Budget and the Executive Office's Request for Information: **Identifying Priority Access or Quality Improvements for Federal Data and Models for Artificial Intelligence Research and Development (R&D), and Testing** (**Docket No: OMB-2019-0003).**

The **AAIH is a 501(c4) not-for-profit** formed in January 2019 as the **global advocacy organization** for the advancement and use of Artificial Intelligence (AI) in healthcare applications. The AAIH is a membership-based organization with **over 25 participating entities** (https://www.theaaih.org/members), who either **develop or utilize AI** in biomedical R&D, devices and diagnostic systems, and clinical applications. Our membership ranges from large diagnostics and drug developers, to leading-edge start-ups, and includes research and medical institutions, technology providers, etc. This RFI mentions the need to understand best practices for quality improvement as it relates to AI data access and models for testing in the United States federal government. We echo and comment on this need through our primary constituency of **industry and academia** member organizations.

AAIH members work **collectively to identify common challenges** and develop strategic priorities to **address industry-wide concerns.** Our organization generates work products through the collaborative efforts of our six standing committees, each focused on separate, broad areas of interest (Education, Federal Engagement, Investment, Communications, Industry Performance and Data Analytics, **Technology and Standards Development**).

We believe that especially in the **highly regulated arena of healthcare**, the establishment, utilization, and sustained improvement of standards in the AI space requires a multipronged and coordinated approach across the healthcare continuum. This is the basis of our philosophy of a **member-driven approach** to a **Public-Private Partnership**, which will be critical to optimizing federal data access and models for collaborative development.

We also realize the linkages and implications of **international policy** in **normalizing AI standards** among emerging governmental strategies. To that end, led by our **Federal Engagement and Regulatory Affairs Committee**, we recently hosted an interactive workshop with industry, academia, and Federal Agency perspectives (**NIH, DOE, NCI, FDA,** etc.) around a need for and potential areas for **Public Private Partnership**, and are continuing to engage our member organizations, collaborators, and stakeholders in other **jurisdictions to foster cross-border and cross-sector harmonization**.

Please see our comments on the RFI around quality improvements for Federal Data and Models in AI in the following pages, and we look forward to continued engagement and collaboration with the Executive Office at this critical juncture in the development of American leadership in AI.

Respectfully submitted *on behalf of AAIH members*,

AAIH TSDC (Technology & Standards Development Committee):
Annastasiah Mudiwa Mhaka, PhD (Co-Founder, AAIH)
Aaron Chang (Strategic and Technical Advisor, AAIH)
info@theaaih.org
www.theaaih.org

Responses to Specific Questions

What Federal data and models are you seeking to use that are private and not at all available to the public? Describe the agency that has the data and what, if any, attempts you are aware of that have been made to increase access to the data or model. What types of AI R&D and testing would be accelerated with increased access to this data?

AAIH members see use in several types of federal data including certain Center for Medicare & Medicaid Services (CMS) claims data, FDA data, NIH data, multimodal data, and biological datasets. Without available and curated proprietary and public datasets, the efficiency of AI applications in areas including drug design and clinical care including telehealth is limited.

Greater accessibility and interoperability of currently disparate patient trajectory data (ICD codes, CPT codes, notes, lab results, costs, etc.) would accelerate the development of predictive AI models for patient outcomes, support AI-based decision systems, manage costs, help triage patients, etc. CMS is the agency that has the data, and several of the FHIR-based interoperability initiatives from ONC and HHS are a step in the right direction.

Separately, claims history at CMS for Medicare reimbursement for telehealth services would be invaluable in assessing whether, as the Congressional Budget Office claims, expanded reimbursement for telemedicine will break the bank or whether, instead, reimbursement will be cost-neutral or even cost-saving.

The FDA has data in the FDA Adverse Event Reporting System (FAERS) database, but it is not optimally formatted and curated to the point of being barely usable. This database needs to be improved and the usefulness of this data for modeling and other data science uses needs to be emphasized during the reformatting process. The FDA has also gathered an enormous amount of data from numerous clinical trials, including why they were terminated, and observed adverse drug reactions (ADRs). This should be made readily available.

In general, the Gene Expression Omnibus (GEO) database serves as a good repository for RNAseq and microarray datasets post publication, but GEO is not widely used across NIH. Furthermore, validation datasets (e.g., Western Blots that target specific genes and proteins) are not publicly available. Validation datasets are often reported through tables and graphs in papers, but those datasets are not publicly available. This practice undermines the ability of outside researchers to validate this research. Access to the raw data enables reproducibility.

Additional data that would allow for the use of AI to facilitate more efficient and effective drug discovery, candidate selection, and drug development include: National Center for Translational Sciences (NCATS) biochemical screening data; toxicology reports from US government-sponsored preclinical studies; full clinical reports on US government-sponsored clinical trials, including blood chemistry and biochemistry, images, genotype, etc.; biophysical models of membrane permeability and tissue distribution; ADME models and associated PBPK models; translational ADME models, preclinical to clinical; molecular initiating event data associated with toxicological adverse events; and target-probe molecular pairs identified in US government-sponsored research.

What Federal data and models are you seeking to use that are restricted to the public, i.e., the data asset is available under certain use restrictions? What types of AI R&D and testing would be accelerated with increased access to this data?

Federal agencies that recruit, train and employ/deploy specialized populations, for example NASA, currently restrict access to potentially useful datasets. We believe that privacy can be protected while making these datasets more widely available. NASA tracks the astronaut population before, during and after spaceflights, sometimes for decades after spaceflights. "N of 1" studies could be enabled by releasing some of these datasets to the public.

In addition, to the extent that some of the information described in the answer to the previous question is available in a restricted way, drug discovery, candidate selection, and drug development would be accelerated by more open and complete access.

What Federal data and models are you seeking to use that are available to the public with no use restrictions, but which have technical issues inhibiting data access? Specifically, what are the technical issues (e.g., is it too big to be downloaded, is it not optimally formatted)? What types of AI R&D and testing would be accelerated with increased access to this data?

The FDA Adverse Event Reporting System (FAERS) is not optimally formatted and curated. Improvements to this database would enhance automated, AI-driven post market quality control and quality improvement initiatives for drugs and therapeutics.

In addition, and in general metadata needs to be supplied and accompany all datasets collected, disseminated and enabled by the US Government. Metadata include how the data were originally generated, for what purpose they were generated and who was the audience and original consumers of the data.

What are the key gaps in data and model availability that are slowing progress in AI R&D and testing? Which areas of AI R&D and testing are most impacted?

In terms of data, published experiments often do not include full data disclosures. The type of data that should be provided is a complete set of metadata that describes the experiment in as much detail as possible. For example, sharing raw experiment files without describing the experimental conditions and the intent of the experiment diminishes the quality and viability of the data.

The data problem affects AI model development. One example that hinders research is the fact that non-disclosed data can stand in the way of comparing one model against another model used in federally funded research. These issues also contribute to the reproducibility problem in academic medicine. The NIH, for example, could require or financially support data disclosure for NIH-sponsored research to promote a culture of data transparency in health research.

Other data gaps include toxicology data emanating from private organizations and tissue distribution, both linked to molecular structure; data and reports on failed clinical trials and metadata, and known target-probe-molecule combinations. Were these data made available, drug discoverers and developers would be able to utilize AI to create better drug candidates that are more likely to have strong safety profiles and to demonstrate clinical effectiveness.

As agencies review their data and models, what are the most important characteristics they should consider? Stated differently, what characteristics of data sets or models make them well-suited for AI R&D?

The old adage of garbage in garbage out lies at the core of why well-curated, annotated, and featurized data is essential for quality AI R&D. Several aspects of ideal datasets and models are as follows:

- **Large datasets**
  - The larger, more inclusive, and evenly sampled the training set, the higher quality the model. A better training set will ensure that the model, when evaluated on the test set is interpolating and not extrapolating (i.e. the training set should contain examples similar in feature space to the test set).
- **Multiple dimensions or measurements**
  - Test set data selection methodology and data redundancy. Randomly generated test sets lead to overestimated practical utility when training data has many related entities. For example, two medical images generated for the same patient could be split between the training and testing sets in a manner that provides a prediction in the evaluation criteria that is not representative of real-world data.
- **Large overlap of measurements between samples**
- **Well populated metadata, even in free text**
  - The data need to be released with extensive metadata, such as; a data format with a clear explanation of the data fields and clear explanations of how the data were generated and why.
- **Clear data access rules and restrictions**
- **Clear and appropriate consents for use**
- **Proper data blinding for privacy or other concerns**
- **Model Evaluation Metrics**
  - A critical review of the experimentation process that was put in place to evaluate the models, including feature space analysis would help attenuate these issues. Datasets with low clustering of feature space are ideal in this practice.
  - Agencies should consider developing standardized test sets, e.g., molecules that are known to FDA to have a given effect on humans.

Which models are most important for agencies to focus on, and why?

Several types of models play an important role in moving the healthcare sector forward. These include public health models, root cause analysis models, and high-quality models in NIH and FDA's initiatives.

Public Health Models

The US spent $3.5T last year on healthcare and we are projected to reach $5T by 2025. Without AI and data to support and enhance medical discovery and clinical workflows, our healthcare system is in real danger of collapsing.

Root Cause Analysis Models

Models that can uncover root cause are very useful. Probabilistic programming and explainable models could be of most benefit to this domain. Agencies typically are not required to employ these models, but the public and the wider research community would benefit if they did.

Quality-Assured Models in Research

Flawed models are prolific in academic research as there is often a rush to publish. The FDA, mainly when approving devices and drugs, but also the NIH should remain vigilant of these issues when awarding grants to academic researchers in the field. Best practices embedded in grant awards could help reward fundamental best-practices and better training for subsequent generations of trainees who will be developing commercial AI products in the near future.

Models for Drug Development

Examples of models that FDA or other agencies could generate with data they have that would most quickly facilitate drug development include therapeutic window algorithms and translational animal preclinical to human adverse event toxicology prediction models.

What characteristics should the Federal Government consider to increase a data set or model's utility for AI R&D (e.g., documentation, provenance, metadata)?

Please see responses above to the question around data and model characteristics amenable to AI R&D.

In addition, searchable and curated documentation of metadata is preferred, and is rarely found in our experience. Metadata describing conditionality of data acquisition, including time, location, experimental conditions, etc. and metadata describing any non-trivial data relationships (ie, acquired from the same patient) can aid in preventing biased interpretation and avoidable errors in model development. Data featurization would also increase a data set's utility.

What data ownership, intellectual property, or data sharing considerations should be included in federally-funded agreements (including, but not limited to, federal contracts and grants) that results in the production of data for R&D?

Data sharing and privacy considerations should take a high priority. Existing GDPR and HIPAA rules are good examples as to how to enable safe and respectful access to patient/personal data.

We argue that data collected on federally funded R&D should be made publicly available. This will not prevent organizations and others from leveraging their work, investing their own resources into it, and protecting what emerges from their investments. However, their initial work, if publicly funded, must be shared with the public, with all the caveats concerning privacy, etc. Data and models should be released, but the algorithms should be protected, if the organization would like to do so.

Furthermore, while it is generally considered that publicly funded data or algorithms should be made free of charge to other academics, there are often licenses imposed on the use for private organizations in an effort to draw more funding into the non-profit research institutes. These fees need to be thoughtfully structured to prevent the stifling of innovation from early stage private sector.

What research questions and applications are you trying to solve with AI, that require specific types and/or quantities of Federal data and models, and how might the Federal Government reduce barriers to discovery and access?

AAIH companies are addressing a variety of research questions and applications which require specific types of Federal data and models. A couple of examples follow:

- Question: What would be the impact upon the Federal budget of liberalizing the rules governing Medicare reimbursement for distance care, such as by eliminating the harsh originating site requirements?
- Application: Predicting patient resistance or response to therapies utilizing models that describe the use and deployment of therapies
- Application: Reducing the failure rate of drugs in the clinic. Models that we develop, we would happily share with the FDA for their use in evaluating new IND applications.
- Application: Reducing the time it takes to go from target-probe-molecule to lead molecule to clinic to under 1 year.

As mentioned earlier, policy should dictate hard limits for license fees for data generated through public funding. These fees should also have a pre-determined cost structure that scale with the size of the organization.

Accelerating the application of AI can be enabled with pre-trained models (e.g., ResNet trained on ImageNet) that facilitate transfer learning. What research questions and applications would benefit most from the transfer learning?

Some applications that could benefit from transfer learning include:

- Measuring objective phenotypes from complex images
- Interpreting device stream measurements
- Converting free text to useful metadata
- Rare disease research