



November 1, 2017

Patricia Flatley Brennan, RN, PhD,  
Director  
National Library of Medicine  
National Institutes of Health

Submitted electronically at: [NLMEPInfo@mail.nih.gov](mailto:NLMEPInfo@mail.nih.gov)

Re: Notice Number: NOT-LM-17-006, Request for Information (RFI): Next-Generation Data Science Challenges in Health and Biomedicine

Dear Dr. Brennan:

The American Medical Informatics Association (AMIA) appreciates the opportunity to submit comments regarding the National Library of Medicine's (NLM) Request for Information (RFI) on Next-Generation Data Science Challenges in Health and Biomedicine. AMIA is the professional home for more than 5,400 informatics professionals, representing researchers, front-line clinicians, public health experts, and educators who bring meaning to data, manage information and generate new knowledge across the health and research enterprise.

The NLM is an indispensable and critical component of the NIH. The research it funds, the training it provides, and the infrastructure, tools and resources it makes publicly available are foundational to biomedical informatics and broadly applicable to domain-specific research across the NIH. **AMIA fully supports NLM as it endeavors to become the “intellectual and programmatic epicenter for data science at the NIH,” as articulated in the June 2015 report by the Advisory Committee to the NIH Director.**<sup>1</sup> AMIA believes that the NLM is uniquely positioned to foster data science competencies, develop, or fund, data science tools / services, and otherwise be the pan-NIH home for data science.

This RFI sought “input on new data science research initiatives that could address key challenges,” and “...promising data science research directions in health and biomedicine.” As such, we have organized our response by describing key challenges and key opportunities for data science in health and biomedicine across the three focal areas. Below, we offer pragmatic research ideas that should enable better use of data in health and biomedicine, as well as promising research projects or concepts.

#### Promising directions for new data science research in the context of health and biomedicine

Health and biomedicine are two domains that are undergoing rapid digitization, supported by evolving IT and data ecosystems. This trend has generated numerous standards and approaches

---

<sup>1</sup> <https://acd.od.nih.gov/documents/reports/Report-NLM-06112015-ACD.pdf>

November 1, 2017

meant to make traditional clinical data computable, and it has enabled various supplemental uses of data generated at the bench and at the bedside. However, we must now make progress on the systematic and strategic use of data in health and biomedicine.

**To do this, AMIA recommends NLM focus research on the basic science of data standards, including development of granular data specifications to enable a “periodic table of elements,” approach to biomedical data standards.** Such an approach would enable the combination and substitution of discrete data elements for specific use cases, such as quality measures, and facilitate data re-use more readily than is the case today. A science of data standards focus must also include development of metadata in health and biomedicine, especially for data traceability, provenance, and accuracy. AMIA members note that NLM could consider advancing the annotation of data or development of metadata by increasing the numbers of and funding amounts for Administrative Supplements grants that focus on both informatics and data science issues.<sup>2</sup> Additional areas of focus that we consider basic research for the betterment of data science include improved methodologies for:

- Data capture from patients and providers across traditional and emerging care settings;
- Data storage (e.g., data formats)
- Data use (including methodologies that enable documentation of how data have been processed) and interactions (e.g. web-based annotation and visualization),
- Measuring data accuracy, and
- Auditing metadata and terminology qualities.

Associated research could focus on:

- Further development of standards for data elements in specific disease areas (e.g., extensions for standards like SNOMED, LOINC, or RxNorm), as well as in nursing and other clinical areas.
- Ways to determine when coding variation (e.g., coding the same natural language phrases into SNOMED by different coders) is appropriate and when it is inappropriate, and support for harmonizing variations in coding.
- Development of standard imaging recognition algorithms for radiology, pathology, and other image based health disciplines.

Dedicated research in these areas could then underpin and contribute to more complex directions for data science, such as:

1. Research on ways to utilize data from democratized sources, including wearable devices, mHealth applications, online journaling and social media tools, geospatial sensors, genomics, and Internet of Things sensors, among others.<sup>3</sup>

---

<sup>2</sup> An example is <https://grants.nih.gov/grants/guide/pa-files/PA-17-090.html>

<sup>3</sup> Examples include the Mobile Sensor Data-to-Knowledge (<https://md2k.org/>) and Mobilize Center (<http://mobilize.stanford.edu/>), both at Stanford University

November 1, 2017

2. Research into *in silico* workflow generation based on existing literature for new workflows or new ways to produce knowledge.<sup>4</sup>
3. Research into how to manage / control systematic biases in “big data” when studying large cohorts of patients.<sup>5</sup>

Promising directions for new initiatives relating to open science and research reproducibility.

Open science and research reproducibility are two important and related trends. By creating a more democratized and decentralized research enterprise, important discoveries and collaborations not previously possible are now commonplace. Meanwhile, research reproducibility represents one of the most existential challenges for research today, perhaps even more so in this new environment of open science. **AMIA recommends that NLM approach these related issues by funding research that can foster trust and assurance in the scientific process.** Three areas that NLM should focus, include:

1. Development of metrics to evaluate how NIH-funded research data are shared;
2. Development of metrics to evaluate the quality and fit-for-purpose of digital repositories; and
3. Development of citation policies more reflective of the open science environment.

The NLM could lead NIH-wide efforts to improve data sharing through (1) making Data Sharing Plans a scorable element of grant applications subject to the existing policy, and (2) developing guidance or best-practices on Data Sharing Plans. Making data sharing plans scorable would enable those projects that prioritize systematic and strategic data sharing through use of standards and accepted best-practice to garner higher scores. Subsequently, best-practices could be developed and disseminated to improve data sharing across the NIH portfolio, which would greatly improve research rigor, transparency, traceability, and reproducibility. Additionally, NLM could play a role in developing guidelines and best practices around data discoverability, including the use of model annotations, metadata schemas focused on a given domain (e.g. imaging) and minimal metadata expectations for NIH-funded research.

In prior comments to the NIH, AMIA strongly supported development of metrics to evaluate the quality and fit-for-purpose of digital repositories.<sup>6</sup> As knowledgebases and deposition repositories collect new and novel datatypes, there will be an urgent need to have a framework indicating the relative strengths and weaknesses of such efforts. AMIA recommends that NLM work in tandem with other NIH Institutes and Centers to publish an evaluation framework for deposition repositories and knowledgebases in a timely fashion, inclusive of dimensions related to data quality

---

<sup>4</sup> One potential application of this research would be to automate efforts like FPIN’s Priority Updates from the Research Literature (PURLs) Surveillance System (<http://fpin.site-ym.com/page/WhatarePURLs>)

<sup>5</sup> Weber G., Adams W., Bernstam E., et al. Biases introduced by filtering electronic health records for patients with “complete data”, *Journal of the American Medical Informatics Association*, Volume 24, Issue 6, 1 November 2017, Pages 1134–1141, <https://doi.org/10.1093/jamia/ocx071>

<sup>6</sup> Request for Information (RFI): Metrics to Assess Value of Biomedical Digital Repositories

November 1, 2017

and completeness.<sup>7</sup> One example for how NIH could operationalize the scoring of data sharing plans, could be to score according to priority data types (e.g. sharing data for autism or rare diseases) and data quality / usability, similar to the 5-star deployment scheme for Open Data.<sup>8</sup> As data science and deep learning tools proliferate, this concept should be extended to include metadata registries and common data element (CDE) repositories, as well as standard reference data sets, which could include documentation of the rationale that underlies permissible value sets.<sup>9</sup>

Finally, given NLM's history with PubMed, it is well-positioned to lead conversations over how to cite data and software developed by grantees in an open science environment. AMIA members point to real-time open science projects that have developed citable reference, which add to the diversity of solutions available. Through a project called "Thinklab," researchers can get feedback on their grant proposals, participate in open peer review, and even lead entirely open research projects.<sup>10</sup> An example of an open science project developed using Thinklab is "Rephetio: Repurposing drugs on a hetnet," where each of the discussions that were part of that project is developed as a citable reference.<sup>11,12</sup> While we acknowledge that such citations do not have the same rigor as peer-reviewed publication it may provide another avenue for recognition for researchers. We also point to the concept of "nanopublications" as means to disseminate individual data as independent publications with or without an accompanying research article.<sup>13</sup> Because nanopublications can be attributed and cited, they provide incentives for researchers to make their data available in standard formats that drive data accessibility and interoperability. Another trend worth noting includes open sharing of original datasets through interactive publication.<sup>14</sup>

Increasingly, the research community understands that reproducibility requires all software used to collect, transform and analyze the data must be publicly available for inspection, modification, and reuse, along with the data.<sup>15</sup> We encourage NLM to become more sensitive to open source concepts relevant to ensuring scientific software practices support reproducibility such as the value of (1) using well-defined, standard open source licenses approved by the Open Source Initiative (OSI)<sup>16</sup>

---

<sup>7</sup> Ibid.

<sup>8</sup> <http://5stardata.info/en/>

<sup>9</sup> An example of an existing standard is ISO/IEC 11179 "Information technology – Metadata registries (MDR) which underlies NCI CDE and AHRQ USHIK. This standard should be modified to include specification of such rationales.

<sup>10</sup> <https://thinklab.com/about>

<sup>11</sup> <https://thinklab.com/p/rephetio>

<sup>12</sup> <https://thinklab.com/p/rephetio/discussion>

<sup>13</sup> <http://nanopub.org/wordpress/>

<sup>14</sup> Ackerman, J. The Educational Value of Truly Interactive Science Publishing. *J. Electronic Pub.*, 18(2), 2015, DOI: <http://dx.doi.org/10.3998/3336451.0018.201>

<sup>15</sup> Ince, D., et al., "The case for open computer programs." *Nature. February 2012.* **482**, S485-488. doi:10.1038/nature10836

<sup>16</sup> <https://opensource.org/licenses>

November 1, 2017

and (2) open source best practices of a public code repository (e.g., on GitHub<sup>17</sup> or Bitbucket<sup>18</sup>) and a public issue tracker.

### Promising directions for workforce development and new partnerships

Development of data science tools and methodologies will do little if the workforce cannot leverage, or does not understand, the basics of how to collect, analyze, and apply data to health and biomedical problems. **AMIA recommends NLM build on its leadership in informatics education and training by (1) ensuring that basic informatics training is supported broadly across health and biomedical domains; (2) encouraging cross-cutting and multidisciplinary programs at the post-graduate and undergraduate levels; and (3) initiating new research on how scientists document their data analysis and information methods.**

AMIA believes a three-pronged approach<sup>19</sup> is needed to educate and train a modern healthcare workforce, including:

- Basic “informatics literacy” for all health professionals that is part of medical education, biomedical research, and public health training to give clinicians the skills needed to collect and analyze information and apply it in their practice;
- Intensive applied informatics training to improve leadership and expertise in applying informatics principles to healthcare problems; and
- Support for education professionals who will advance the science and train the next generation of informatics professionals in this developing and dynamic field of study.

AMIA is encouraged by private sector efforts to develop cross-cutting and multidisciplinary education programs, such as computer science and information science with health and biomedical sciences,<sup>20</sup> in addition to recent public-sector initiatives, such as the CTSA Program Data to Health Initiative.<sup>21</sup> AMIA recommends NLM seek ways, through funding or other means, to support graduate and undergraduate programs tailored to health data science, clinical predictive analytics, and health informatics through incorporation of computer science and statistics curriculum. One possible area of collaboration could be with the National Science Foundation and their graduate research fellowship program, which could help integrate traditional and emerging STEM fields with the biomedical and life sciences.<sup>22</sup> AMIA members see a glaring need to provide basic medical / health literacy for engineers, statisticians, and image scientists whose skills will be necessary to meet

---

<sup>17</sup> <https://github.com/>

<sup>18</sup> <https://bitbucket.org/>

<sup>19</sup> Perlin, J., Baker D., et al. “Information Technology Interoperability and Use for Better Care and Evidence: A Vital Direction for Health and Health Care,” National Academy of Medicine. September 19, 2016

<sup>20</sup> An example is a new joint PhD degree program offered by Clemson University and the Medical University of South Carolina in Biomedical Data Science and Informatics (BDSI) <https://www.cs.clemson.edu/bdsi/>

<sup>21</sup> <https://ncats.nih.gov/files/council-concept-CD2H-initiative-9-2016.pdf>

<sup>22</sup> <https://www.nsfgrfp.org/>

November 1, 2017

the challenges presented by data science and big data in health and biomedicine. The NSF may be one such partner towards this aim of developing and delivering data science and informatics curricula.

Lastly, AMIA suggests that NLM fund analogous studies to the 1991 model study on biotechnology awareness<sup>23, 24</sup> to determine ways in which modern information and communications technology have impacted how scientists document their data analysis and information methods. Such a study would provide a great service to education and training programs by providing a view into pervasive methods and inform ways to optimize data analysis and information methods given advances in data science.

\* \* \*

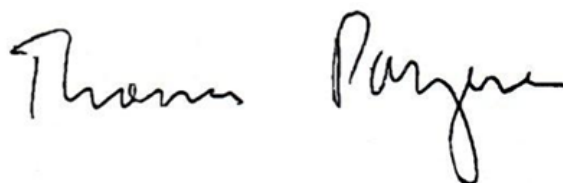
We appreciate the NLM's work in this important area, and we are excited about the possibilities data science can bring to health and biomedicine. We look forward to working closely with the NLM to bring the expertise of health informatics professionals to our exciting and shifting paradigm for health and healthcare.

Thank you for considering our comments. Should you have questions about these comments or require additional information, please contact Jeffery Smith, Vice President of Public Policy at [jsmith@amia.org](mailto:jsmith@amia.org) or (301) 657-1291. We look forward to continued partnership and dialogue.

Sincerely,



Douglas B. Fridsma, MD, PhD, FACP,  
FACMI  
President and CEO  
AMIA



Thomas H. Payne, MD, FACP, FACMI  
AMIA Board Chair  
Medical Director, IT Services, UW Medicine  
University of Washington

---

<sup>23</sup> Grefsheim S., Franklin J., Cunningham D. Biotechnology awareness study, Part 1: Where scientists get their information. Bull Med Libr Assoc. 1991 January; 79(1): 36–44.

<http://pubmedcentralcanada.ca/pmcc/articles/PMC225482/>

<sup>24</sup> Cunningham D., Grefsheim S., et al. Biotechnology awareness study, Part 2: Meeting the information needs of biotechnologists. Bull Med Libr Assoc. 1991 January; 79(1): 36–44.

<http://pubmedcentralcanada.ca/pmcc/articles/PMC225484/>