AAMC

Association of
American Medical Colleges
655 K Street, N.W., Suite 100, Washington, D.C. 20001-2399
T 202 828 0400
www.aamc.org

January 19, 2017

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

RE: NIH Request for Information: Strategies for NIH Data Management, Sharing, and Citation (NOT-OD-17-015)

The Association of American Medical Colleges (AAMC) is pleased to have this opportunity to offer comments related to data management and sharing strategies and priorities for the NIH. The AAMC is a not-for-profit association representing all 147 accredited U.S. medical schools, nearly 400 major teaching hospitals and health systems, and more than 80 academic and scientific societies. Through these institutions and organizations, the AAMC represents nearly 160,000 faculty members, 83,000 medical students, 115,000 resident physicians, and thousands of graduate students and postdoctoral trainees in the biomedical sciences.

The AAMC has long supported data sharing in basic and clinical studies, and has embraced efforts to maximize the use of data resources. We appreciate that NIH has asked the research community itself for information as it formulates broad strategies for building data resources. The following suggestions for key elements in development of a data sharing strategy and in data citation are drawn from research leaders at our member institutions. The AAMC has also helped publicize and disseminate the RFI to encourage researchers and organizations to respond directly.

RFI Section 1: Data Sharing Strategy Development:

(1) Highest priority types of data to be shared and the value of sharing such data.

- In our discussions, there was no real sorting of priorities for the types of data to be shared. Ideas ranged across basic, translational, and clinical research, as well as health services and population data, and were not confined to specific fields or studies. The point most commonly and emphatically made by investigators and research leaders is the necessity to capture the totality of information required to make data useful, including documentation of context, limitations, and other metadata. Data are seldom useful absent such curation. Useful data storage needs to include relevant software or analysis code in the resource as well as the raw data. Imaging studies require documentation regarding acquisition, imaging modalities and patient parameters, in addition to other study information. These types of complex data packages will optimize the utility of the information and facilitate reproducing studies.

- Negative data are especially valuable to post in repositories. While many data are not published because they negate or are inconclusive about posited research questions, they become particularly valuable in meta-analysis. The advantages to sharing negative data, especially in clinical studies, have been noted elsewhere, and are consistent with the notion that science advances one failure at a time.
- Ultimately, discussants noted, we never know what data will be useful, or how it might be used in future, given unpredictable changes in science, and in the technologies that make use of data.

(2) The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

- The length of time for making data available for secondary research purposes would be indefinite. It will often exceed the length of time for research projects or grants themselves, and may well exceed time that key personnel remain at an institution. This has implications, noted below, for stewardship of data, and intrinsic cost.
- The nature of studies will also affect this calculation. Consider, for example, long-term longitudinal studies, where data may accumulate and be shared through the life of the project. In many other studies, the data will be posted a certain time after initial publication (our constituents preferred a calendar year). Original investigators should have sufficient time for analyzing data before making publicly available, on timeframes that may vary by type of study.
- Discussants also noted that we never know what data will be of value in the future, given unpredictable advances in science and technology. An animating vision to guide NIH might be the use of shared data resources to support machine learning and specialized algorithms for searching, synthesizing and analyzing information.

(3) Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

- Related to the need to curate and document data and relevant software, cost was the central concern raised by researchers. For data sharing to advance, research sponsors and institutions must commit resources.
- It is not clear that the public or political leaders, who increasingly support or call for data sharing (and other “transparency”) appreciate the additional burden and cost of creating usable shared data resources. Perhaps this is because of the ease with which other types of information can be so easily shared. Submitting data and documentation to repositories ensures preservation and accessibility of data for the research team. It also increases citation of work, increases visibility, and opportunities for new scientific collaborations. On the other hand, data sharing reduces investigator advantages when applying for grants and limits protection of publication opportunities for the research team, students and colleagues. The research community is coming to appreciate the opportunity costs and expense to society and science of not sharing data from publicly or

privately financed research. That realization is ultimately the impetus for continued progress.

- There is a need for both generalist data repositories, for a wide variety of data, and specific repositories. These should continue to be developed and supported by NIH, in addition to institutions or other organizations. The overall utility of these efforts is related to the commitment to standards and providing support. Investigators we spoke with also favored opportunities that facilitate creation of study specific repositories. The best designs and standards emerge from the research communities themselves.
- Unfortunately, establishing many free-standing data repositories will likely limit the utility of the data: what you can't find, you can't use. In addition, inconsistent formats may further degrade utility. Thus, there is a conundrum: one central repository with common standards (perhaps for clinical data from studies) might be possible and have real utility, but it would be likely to be too constrained for data obtained in non-standard ways.

(4) Any other relevant issues respondents recognize as important for NIH to consider.

- A principal concern for data repositories is cybersecurity. Not only is security an issue for clinical data—where the privacy of human research participants, patients, etc.,--must be protected, but is also an issue for non-human and other types of data as well, which may be subject to theft, sabotage, or alteration. New policies, such as strong legal protections, standards, and data-use agreements, as well as new technologies will help address concerns. Particular attention is being paid to blockchain technology, a data ledger used for Bitcoin, as a means to enforcing privacy and agreements, for example.
- Lead time for sharing data includes - formatting data; describing scope of consent and data usage; preparing documentation and data dictionaries; and obtaining required institutional approval. Given researcher and institutional commitments that must take place for data sharing to occur. The research community needs to create clear parameters and pathways that make the process easy and consistent. In addition, utilization of these approaches needs to be evaluated and investments should be balanced and proportional to utilization and impact.
- While we tend to describe repositories as centralized resources, they can also be decentralized, federated and structured according to many types of arrangements (consistent with the "bottom-up" approach preferred by most discussants). NIH should examine various existing models, and build incrementally from those models.

Section II: Data and Software Citation in Research Performance Progress Reports (RPPRs) and research grant applications.

(1) The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing.

- Citation and credit for generating and sharing data is fundamentally important as an incentive, to help recognize and advance productive investigators. Use of standard object codes and links for citation will help in recognition. Blockchain, mentioned above, can also be used to track who has accessed or made use of data. For some types of research, data generators may be viewed as research collaborators (or even authors) on a study. But as shared data resources become more routine and commoditized, data may be cited like other sources or references.

- While DOI is useful, NIH should encourage biomedical informaticists to develop alternative methods for standard identification (the Biomedical informatics subgroup of CTSA consortia may be helpful.) NIH and the research community should also consider developing a global tracking system for secondary publications from shared data sets.
- Effective citation will help improve the rigor and reproducibility of studies, including the increased availability of negative data.
- Citation will also have an impact on efforts to improve research integrity.
- Discussants noted that it is necessary to change the current system, but urged accepting that such changes will take time, and encouraged the creation of an easy, consistent format that is not up to interpretation.

Other topics important for NIH to consider:

- Several constituents have proposed that secondary research conducted with patient-level data should be independently reviewed for scientific merit as a condition of access. This point emphasizes again protection of risk to research subject confidentiality where identifiable data necessary for analysis, or where there is potential for re-identification.

The AAMC appreciates the opportunity to comment to the NIH on this issue and would be happy to provide any further information moving forward. Please contact me or my colleague, Stephen Heinig, Director of Science Policy, (sheinig@aamc.org, 202-828-0488) with any questions about these comments.

Sincerely,

Ross M. Kinney J. MD

Ross E. McKinney, Jr, MD
Chief Scientific Officer