



**Association of
American Medical Colleges**
655 K Street, N.W., Suite 100, Washington, D.C. 20001-2399
T 202 828 0400 F 202 828 1125
www.aamc.org

December 10, 2018

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

Re: NOT-OD-19-014 “Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research”

The Association of American Medical Colleges (AAMC) appreciates the opportunity to comment on NIH’s request for information regarding proposed provisions for a draft data management and sharing policy. The AAMC is a not-for-profit association representing all 152 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 173,000 faculty members, 89,000 medical students, 129,000 resident physicians, and more than 60,000 graduate students and postdoctoral researchers in the biomedical sciences. These comments on NIH’s proposed provisions consider feedback provided by AAMC-member institutions on their data sharing practices as well as broader standards in the scientific community.

The AAMC supports efforts to increase sharing and re-use of scientific data generated through NIH-funded research and for the agency to develop a clearly defined policy on data sharing. While there has been ambiguity in this space that has impeded investigators from meeting sharing aspirations, there is also recognition that the benefits and utility of data sharing vary significantly across datasets. Any policy developed should reflect this understanding.

In addition to responding to the specific questions for which NIH has requested information, AAMC provides the following high-level recommendations as NIH develops its data sharing policy and the processes it will use to implement and enforce that policy:

- Given the constantly evolving nature of scientific data, we encourage NIH to ensure that the new policy be evaluated regularly and be sufficiently flexible to keep pace with rapidly changing technologies. **Plans to evaluate the impact of the policy should be described and implemented prior to its effective date** to align agency and community expectations about the metrics that will be evaluated. At a minimum, these metrics should include the percentage

of the awarded grants' budgets that are designated for data management and sharing activities and a mechanism to determine whether the shared data have been accessed or re-used.

- **The data sharing policy should be applied consistently regardless of which institute is providing the funding.** The proposed provisions allow for individual Institutes or Centers (ICs) to set differing requirements for data management and sharing, but we urge NIH to harmonize requirements and develop an agency-wide policy that takes into account data type and scientific discipline. Given that institutions and investigators receive funding from multiple ICs, a more consistent policy would facilitate and simplify the development of data management and sharing processes across grants and projects. Where IC-specific variations are necessary, this guidance should be clear and readily available to researchers.
- We recommend that NIH give **explicit guidance on how the new policy would “establish expectations for other NIH policies,”** particularly the NIH Policy on the Dissemination of NIH-funded Clinical Trial information and the NIH Genomic Data Sharing Policy.
- The proposed provisions suggest that NIH is intending to create a draft policy that lists only broad principles and requirements, and then asks investigators to develop and propose an application-specific data management and sharing plan. While the flexibility is appreciated, **if NIH does have unstated expectations for what types of data must be shared or how accessible that data should be, those expectations should be included in the policy.**

In furtherance of the fourth recommendation, we suggest that NIH provide more specific guidance about the agency's expectations about which data investigators should share and for how long that data should be available. Perpetual storage for all data is not feasible or helpful; it would be hugely expensive to hold all data to the same standard and also make the most useful data harder to find. If the policy is too broad, it will not achieve the desired goals of producing meaningful shared data, but an overly prescriptive policy might not be able to keep up with the state of the science. It is hard to know in advance what data will be valuable, although useful benchmarks for a middle ground include: any data that underlies a publication or the minimum dataset that is required to reproduce the analyses that leads to the conclusions of a research project, as well as data which are readily deposited into well described, curated, funded repositories.

Effective data sharing will also be aided by a greater understanding of the use of shared data, and we are pleased that the NIH is involved in the AAMC's "Credit for Data Sharing"¹ initiative, a collaborative project with the New England Journal of Medicine and the MRCT Center, which is working to promote a validated, systematic pathway to link datasets to publications, allow academic researchers to obtain credit for shared datasets, and incentivize and promote data sharing in accordance with FAIR (Findable, Accessible, Interoperable, Reusable) data principles.

¹ www.aamc.org/datasharing; <http://www.nejm.org/doi/full/10.1056/NEJMs1616595>

In response to the specific questions posed by NIH:

I. The definition of Scientific Data:

The AAMC generally agrees with the proposed definition that scientific data should not include preliminary analysis, lab notebooks, and other early outputs of the research process, and should primarily consist of “individual level and summary or aggregate data, as well as metadata.” NIH might consider including not only metadata in this definition but also code (e.g., SQL, R, Python) used to interact with the data.

We note that without the use of standards for data and metadata, data sharing will not be useful for secondary analysis or results replication and is unlikely to achieve the stated goal of increased scientific progress. Thus, AAMC urges the NIH to provide as much guidance as possible on the necessary elements for both data and metadata, and wherever possible, provide examples of accepted standards to increase the usability and interoperability of the data reported to NIH. For certain common data types, it would be helpful if NIH specified the metadata necessary for further analytic processing. We also recommend that NIH continue to invest in projects such as the NCATS Biomedical Data Translator² and PhenX Toolkit³, which harmonize metadata and promote cross-study comparisons and analyses.

The AAMC supports the usage of common data elements and discipline-specific schema for metadata; however, researchers have expressed that resources such as the NIH Common Data Element (CDE) portal can be difficult to navigate without significant knowledge of data science. Additionally, investigators who use a core for data analysis (sometimes at a different institution) may not have adequate expertise to identify the necessary metadata if it is outside of their primary research domain. We recommend that NIH explore ways to make it easier for researchers to identify the appropriate metadata for their data type, update the CDE portal so it is more user-friendly and continue to fund and incorporate lessons from community-based tools such as the NCBO BioPortal⁴.

II. The requirements for Data Management and Sharing plans

The development of a Data Management and Sharing Plan (“DMSP”) is a critical element in integrating data-specific considerations into the research lifecycle. The AAMC supports a requirement to include a DMSP as a part of applications and proposals for NIH-funded research, both for its role in promoting future data sharing as well as its utility as a planning and compliance document. We have also commented in this section about current challenges to carrying out proposed

² <https://ncats.nih.gov/translator>

³ <https://www.phenxtoolkit.org/>

⁴ <https://biportal.bioontology.org/>

elements of the DMSP which the NIH would have to address prior to putting a data sharing policy in place.

We agree with the current proposal that the DMSP be evaluated as an Additional Review Consideration but not factored into the overall impact score for extramural grants. Particularly in the early stages of policy implementation, the focus should be on education and guidance for investigators on how to correctly formulate a DMSP. Once a plan is agreed upon and approved by NIH staff and grantees, compliance with the DMSP could be integrated into award terms and conditions. Formal inclusion of the DMSP in the annual Research Performance Progress Report would likely facilitate adherence to the plan within the proposed timeframe, as well as an opportunity to address any challenges or barriers that have emerged.

The AAMC encourages the NIH to harmonize its DMSP requirements with other federal science agencies whenever appropriate. We also recommend that NIH provide a template or recommendations as well as sample DMSPs which would clearly state the necessary components of a plan as well as the appropriate level of detail. If NIH has specific expectations for investigators with regard to data management, it is important that the DMSP be comprehensive enough to define these requirements at the start of the research funding period.

With regard to data preservation and storage, the timeline for how long data should be maintained will depend on several factors, including the potential long-term utility of the data. Most institutions currently use grant funding in real-time for data collection and curation. It is unclear who bears the responsibility for funding the continued curation and storage of data after a grant has ended. One solution might be to require the researcher to save data for a fixed period of time, determined by the initial grant so that it varies appropriately with the science. For fields where the mandated time is very long (e.g. >5 years), NIH should consider a federally funded repository, possibly run by the National Library of Medicine.

Additionally, as it can be very expensive for institutions to retain data locally, cloud-based platforms from NIH would create long-term capability for data storage and help alleviate some of this financial burden. We appreciate and encourage NIH's ongoing effort to expand resources in this space, including the development of the NIH Data Commons as well as the STRIDES Initiative to make use of commercial cloud computing.

In relation to issues of data discoverability and access, the proposed provisions suggest that the data should follow the FAIR principles⁵, and we also commend to NIH for consideration the "FAIR-TLC" principles⁶ proposed by the Monarch Initiative, TransMed NCATS Data Translator projects, and International Society for Biocuration, which posit that in addition to being FAIR, data should

⁵ <https://www.nature.com/articles/sdata201618>

⁶ <https://zenodo.org/record/203295#.XAgMEuInaUk>

also be traceable, licensed, and connected. From AAMC's work on data sharing, we have found that if datasets are not assigned a persistent identifier when shared, it is impossible to fully track data usage, and subsequently, appropriately credit researchers for their contributions.

We also recommend that NIH provide a list of the desired characteristics of a repository, using standards for discoverability and archiving developed by the research data community and groups such as the Research Data Alliance⁷ and FORCE11⁸, as well as specific recommendations for common data types (as in the case of DbGaP for genomic data). In terms of a timeline for when the data needs to be made accessible after the research concludes, we believe it will create significant confusion and lack of consistency if this element is proposed anew by the investigator for each grant and recommend instead that NIH suggest in the policy a specific length of time within which the data should be made accessible, with the option for the researcher to request a longer timeframe and provide justification.

Under the section on data sharing agreements and licensing, the proposal states that "NIH encourages terms that provide for the broadest use of data resulting from NIH-funded or -supported research, consistent with privacy, security, informed consent, and proprietary issues." Institutions have shared with AAMC that the lack of standardization in licensing agreements greatly complicates and hinders inter-institutional sharing of NIH-funded data. This process could be facilitated if NIH were to provide a template licensing agreement that lays out sharing terms for data as it has previously done for biological materials.

To better recognize the unique aspects of research with human subjects and to distinguish research with data derived from biospecimens, we urge NIH to revise the proposed provisions' approach to requirements for data sharing, access, and privacy as they relate to these types of research. We appreciate NIH's recognition that data sharing aspects of a proposal (including standards for data and metadata collection) need to be developed by an applicant in parallel with informed consent considerations but urge that the data sharing policy itself not appear to dictate informed consent expectations or requirements. We recommend that the draft policy instead remind institutions and investigators that where informed consent is required, the processes of developing the DMSP and the informed consent language must be coordinated instead of suggesting that the data sharing considerations drive the promises made to subjects in the informed consent document. We further recommend that NIH publish guidance to ensure that Program Officers and awardees alike are attuned to the set of issues that will dictate these elements of the DMSP.

The proposed provisions include the statement: "Data may be shared across institutions and repositories to maximize utility, and informed consent should permit broad sharing wherever possible." While an understandable ideal, the language in the informed consent may have been

⁷ <https://www.rd-alliance.org/>

⁸ <https://www.force11.org/>

developed and administered long before the application (in the case of existing data), may not be required at all (in the case of unidentified biospecimens), or may be substantially modified by the IRB after the grant application to account for local context, vulnerable populations, or other considerations. For research that involves human subjects, it will be critical for awardees to connect the informed consent development and institutional review board (IRB) approval process with the negotiated elements of the DMSP to ensure that objectives of the DMSP are realistic, given any constraints of state or federal regulations, institutional policies, or the nature of the data being collected. Even when the research will not involve direct interaction with human subjects or with biospecimens, the policy should explicitly acknowledge the limitations that may be placed on data sharing when using data derived from participants or biosamples from certain populations, such as Native American tribes.

On the issue of identifiability, it is important to recognize that de-identification of data from human subjects is neither simple nor inexpensive and requires additional resources, as further discussed below. Institutions have also expressed the need for allowances with respect to data sources that are not easily de-identifiable (for example, certain images or results from clinical equipment that do not support de-identification) and would like the NIH to provide additional guidance on how to assure appropriate use of research data made publicly available. While the policy frequently references “researchers and the broader public,” AAMC emphasizes that these are two very different end-users, particularly in the case of clinical data, which are more likely to be shared under a controlled-access model.

III. The optimal timing, including phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy.

Compliance with a new data management and sharing policy may require significant changes to the research process and additional time, infrastructure, and resources both on the part of the institution and the individual investigator. AAMC encourages NIH to include a proposed implementation timeline in the forthcoming draft policy to provide stakeholders with an opportunity to comment and suggests that, at a minimum, the policy should be applicable to applications submitted one year from the publication of the final policy. For further insight into the needs of the extramural community in this space, we also refer NIH to the “Accelerating Public Access to Research Data” initiative⁹ from the Association of American Universities (AAU) and Association of Public and Land-grant Universities (APLU), which is working with institutions on strategies to facilitate data sharing and management.

The proposal states that “reasonable costs” associated with data management and sharing could be requested under an award budget, and we believe this is essential for investigators to be able to comply with a data sharing policy. As this number may vary widely depending on the data type and

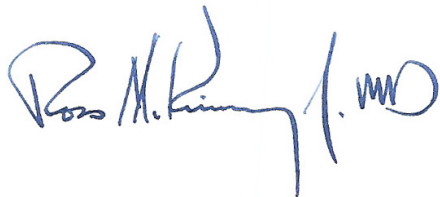
⁹ <https://www.aau.edu/key-issues/aau-aplu-public-access-working-group-report-and-recommendations>

volume, we would like to request additional estimates from NIH on allowable expenses. It would also be helpful if the cost of data sharing was included as a standard line item in grant budgets, as it is now for certain types of research. We realize that NIH has funding mechanisms outside of the Research Project Grant, such as training or infrastructure grants for libraries and core facilities, that could supplement funding for data sharing efforts. While this is helpful for institutions, we would encourage the NIH to disseminate any lessons learned from these funding mechanisms to the broader extramural community, so that the benefit of this investment is distributed beyond the individual institution.

We are in an era of data-centric approaches to understanding biomedical problems, and the AAMC shares the NIH's goal of increased scientific data sharing. However, we would emphasize that a policy alone will not be sufficient to reach this objective. The agency must ensure that it is providing adequate training, education, and guidance, increasing available financial resources, and leading the development of tools and infrastructure in order to enable and facilitate policy implementation. Currently, there are not enough data scientists or informaticists to curate data in the way it will be required by the NIH. While this issue can be partially addressed in the long-term through efforts in training and curriculum changes, a different solution is needed for the implementation timeframe of this policy. Scientists need to be able to identify high-value data and appropriately annotate that data before it is shared. NIH's corresponding investment in research and development will create the toolbox that makes this possible.

The AAMC is very appreciative that NIH is engaging stakeholders at this early stage of the policy process and looks forward to continued engagement on this issue as the data sharing and management draft policy and other guidance are developed. Please feel free to contact me or my colleagues Anurupa Dev, PhD, Lead Specialist for Science Policy (adev@aamc.org) and Heather Pierce, JD, MPH, Senior Director for Science Policy and Regulatory Counsel (hpierce@aamc.org) with any questions about these comments.

Sincerely,

A handwritten signature in blue ink, appearing to read "Ross E. McKinney, Jr., MD". The signature is stylized and cursive.

Ross E. McKinney, Jr., MD
Chief Scientific Officer