March 10, 2020

Office of Science and Technology Policy
Executive Office of the President
Eisenhower Executive Office Building
1650 Pennsylvania Avenue
Washington, DC 20504

**Re: Request for Public Comment on Draft Desirable Characteristics of Repositories for
Managing and Sharing Data Resulting from Federally Funded Research (85 FR 3085)**

Submitted electronically to: OpenScience@ostp.eop.gov

The Association of American Medical Colleges (AAMC) appreciates the opportunity to
comment on the White House Office of Science and Technology Policy (OSTP) request for
information on desirable characteristics of data repositories, as proposed by the National Science
and Technology Council's Subcommittee on Open Science. The AAMC is a not-for-profit
association representing all 155 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.

The AAMC strongly supports improved access to data resulting from federally funded research.
The development of consistent guidelines and clearly defined characteristics for repositories to
preserve and provide access to research data are critical in enabling academic institutions to
achieve this goal. AAMC encourages harmonizing these guidelines for investigators and
institutions across agencies as much as possible, while still allowing for flexibility to
accommodate different fields of research and agency objectives. We also agree that in some

instances, it is most effective for the agency to designate a specific repository for particular research initiatives or data types.

Additionally, as AAMC noted in recent comments[1] in response to the National Institutes of Health's draft data management and sharing policy, many institutions are planning on building and/or expanding their own repositories as agencies institute new requirements for researchers, and "without guidance from the agency on standards for data storage and discoverability… holding data in such disparate platforms and systems will place a significant technical burden on anyone who wants to access the data, thwarting the agency's laudable goals to increase and improve data reuse."

The AAMC is generally supportive of the proposed desirable characteristics of data repositories, many of which we note align with community-driven criteria proposed last year.[2] Given the rapidly developing importance of data in scientific research, these guidelines should be flexible enough to keep pace with technological advances, as well as the increasing volume and diversity of scientific data.

We strongly agree with the recommendation (C) that repositories assign datasets a "citable, persistent unique identifier (PUID), such as a digital object identifier (DOI) or accession number." Attaching a PUID to a dataset would not only support data discovery and research progress reporting, as noted by OSTP, but is a critical step in tracking data reuse, crediting investigators for their work, and ultimately developing a more comprehensive understanding of research outputs. We also note that the use of PUIDs has previously been suggested by several federal research funding agencies, including the National Science Foundation.[3]

A recently published initiative from AAMC and other research stakeholders to promote effective data sharing describes a path to connect researchers to their datasets, based on the use of PUIDs.[4] While the use of PUIDs for datasets is key, we recommend that the subcommittee consider specifying that repositories provide the option to attach additional unique identifiers to the dataset, including ORCID ID for investigators, and in the future, grant and/or organizational IDs.

---

[1] AAMC Response to NIH NOT-OD-20-013: "Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental Draft Guidance" (2020). https://www.aamc.org/system/files/2020-01/ocomm-ogr-AAMC%20Response%20to%20NIH%20draft%20data%20sharing%20policy.pdf

[2] Sansone, et al. Data Repository Selection: Criteria That Matter (2019). https://osf.io/m2bce/

[3] NSF 19-069: Dear Colleague Letter- Effective Practices for Data (2019). https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp

[4] Pierce, et al. Credit Data Generators for Data Reuse. *Nature* 570, 30-32 (2019). https://www.nature.com/articles/d41586-019-01715-4

Such connection between identifiers, beginning with the repositories, is necessary if the goal to fully and effectively track data reuse is to be realized.

AAMC agrees with the recommendation (E) that repositories should provide "maximally open access to datasets, as appropriate, consistent with legal and ethical limits to maintain privacy and confidentiality." We suggest that this recommendation also include providing access to metadata, in agreement with FAIR data principles.[5] We also recommend that the language in (E) replace the recommendation (F) that repositories should make datasets "accessible free of charge in a timely manner after submission," which does not seem to allow for restricted use cases.

We appreciate the subcommittee's recognition that repositories which store data from individuals require additional considerations in order to ensure adequate privacy and security, as well as controls on use and access, even when those data are considered de-identified. However, some of the proposed characteristics, including (A) restricting dataset access to appropriate uses consistent with original consent and (B) the need for a repository to enforce submitters' data use restrictions, while imperative considerations for human subjects data, may be outside of the traditional purview of a repository. We urge the subcommittee to consider specifically which of these recommendations are suitable for a list of recommended repository characteristics, and which would be better addressed under a separate agency policy or guidance and be the responsibility of the investigator depositing the data. Regardless of the mechanism, we agree that specific promises made to human subjects through consent documents about the use or sharing of research data should be honored and that repositories should facilitate, not create barriers to, the ability for investigators to ensure those promises are kept.

We strongly encourage, in addition to these guidelines on repository characteristics, the creation of a clearinghouse for federal research data policies and related resources, such as tools for metadata creation. Investigators may also find helpful a comprehensive list of agency-supported repositories, as is currently maintained by the National Library of Medicine,[6] as well as links to other commonly used repositories to store the results of federally funded research. In order for data to be successfully reused, it must not only be deposited in an appropriate repository, but also meet several other criteria, including adequate metadata, curation, and the use of common standards. Providing additional guidance on these topics is essential to meeting the end goal of effectively sharing the results of federally funded research.

---

[5] Wilkinson, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://www.nature.com/articles/sdata201618
[6] Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)- Data Sharing Resources (Accessed March 2, 2020). https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

The AAMC appreciates OSTP's efforts to seek input from stakeholders and looks forward to continued engagement as the federal government develops guidance relevant to data management and sharing. 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,
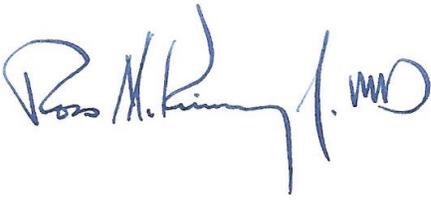
Ross E. McKinney, Jr., MD
Chief Scientific Officer