AcademyHealth Response

NIH Proposed Provisions for a Future Draft Data Management and Sharing Policy

December 10, 2018

AcademyHealth consulted with a committee of members and thought leaders in order to offer the below response to an NIH request for comments on their Proposed Provisions for a Data Management and Sharing Policy. We thank and acknowledge Dr. Adam Wilcox, who chaired the committee, and whose assistance and expertise were invaluable to the compilation and formulate of this response.

As a member-based organization that serves the research community, AcademyHealth is deeply interested in the production and use of evidence to improve health and the performance of the health system. Our work, and that of our more than 4,000 individual and organizational members, is directly impacted by policies regarding the collection, use, governance and sharing of data to facilitate research. As such, AcademyHealth applauds the National Institutes of Health (NIH) for adopting a plan to increase access to scientific publications and data in 2015, and we are encouraged by the current effort to consider a new data management and sharing policy in support of that plan.

We commend the NIH for offering these proposed provisions, which provide a helpful foundation for considering these issues, and suggest that more definition is needed both to ensure that researchers understand NIH expectations and that the overall vision to encourage responsible data management and sharing is realized. In this spirit, that we are grateful for the opportunity to provide comments on the three primary topics of interest for the NIH.

The Definition of Scientific Data

The definition of scientific data offered in the proposed provisions is most helpful when considered through the lens of data sharing, to test the validity of research findings. Seen in the context of the Science Data Lifecycle Model (https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle), the definition appears to be focused on data at the *Analyze* stage - since preliminary analyses, which may be used for data processing and preparation, are not included. However, for the goals of testing validity, combining data sets, and exploring new frontiers, the definition may need to be expanded to include data in the *Process* stage. As more data are available in different forms for research and analysis, data for analysis is increasingly dependent on processing activities. This is important because some findings should be tested beyond the final analysis and more with data prior to being selected and filtered. In addition, data sets from different studies may be difficult to combine effectively without applying consistent processes for preparation.

An illustrative example of the value of pre-processed data for sharing is when phenotypes are computed from other data elements, such as with observational studies using data extracted from electronic health records. This is an increasingly common type of research performed and used by AcademyHealth members. Sharing these data sets is less valuable for validation, combination and exploration if it only includes the final computed value. Another example when scientific data are used with machine learning. Common approaches of applied machine learning include a process of feature selection or engineering, where characteristics of multiple data variables are evaluated together to select or derive a smaller set of variables that become the focus of the machine learning computation. Under the current definition of scientific data used, the feature selection or engineering stage could be interpreted as preliminary analyses and excluded from the policy, yet these data would be critical to sharing goals.

Another type of data that is increasingly important for inclusion in the definition of scientific data is synthetic data. Data resources that are created through random generation processes or disruption of data elements in known data sets are often used in simulation and modeling research. These are important in the scientific process. Testing the validity and accuracy of such research will require that these data are shared with other researchers.

We recognize that expanding the current definition to include pre-processed data for sharing may complicate other provisions in this policy. However, this may be necessary to better enable the policy to achieve its defined goals for sharing. Further, practices of data management may be identified that can allow an expanded scope of data sharing while limiting the challenges for data management and governance that would be introduced.

The definition of scientific data also needs consideration of and clarity around requirements regarding data that are generated through primary data collection vs. existing data sets. Some existing data sets used in health services research are licensed or have data use agreements that might restrict broader sharing. Provision #6, “Data Sharing Agreements, Licensing, and Intellectual Property” under “Requirements for Data Management and Sharing Plans” shows awareness of licensing and data use agreements, but could be easily interpreted as only applying to those that might be defined by the supported research. It would also be preferable to explicitly include these data in the definition of scientific data, and the provisions should provide guidance on how they should be considered in this policy. Inappropriate consideration of these existing data sets under this policy could lead to either misuse or reduced use of these data. Such data should not be simply excluded from the policy requirements. Methods for data processing or analysis could be shared, thereby providing pathways for validation and promoting exploration.

The concept of data management, which is used in definitions supporting the definition of scientific data, itself needs better definition in the provisions. Above we have applied existing models of data lifecycles for data management to help interpret the definition of scientific data. If the policy is more explicit about data management and the stages of the data lifecycle to which the policy pertains, it can better define scope and guide compliance.

Finally, AcademyHealth supports the emphasis on digitizing scientific data that is made in the definition, as digital data are generally more easily shared. We recommend expansion to include use of standards where possible as well. Data shared in digital formats may be more easily stored and distributed, but their actual use will be more limited when the data are represented in

non-standard formats. Data for sharing should be stored and represented in a way that they can be both distributable and interoperable.

The Requirements for Data Management and Sharing Plans

AcademyHealth believes the expanded requirements for data management and sharing plans can advance data sharing by requiring more detailed consideration and commitment by scientists to share data. Our members have observed that while some consideration of data sharing has been required in prior proposals for NIH-supported research, it has been difficult for researchers to develop plans that support the goals of data sharing. Often, plans are specified that do not promote open data for science, but rather provide a minimal commitment to make some degree of data sharing possible. We also believe that the support of and incentives for investigators to develop and implement acceptable data management and sharing plans will be critical to the success of the policy.

We recognize that these new requirements will require expanded efforts by investigators and their institutions for adherence, especially during early stages before best practices are established and systems for open research data are optimized. NIH will need to closely evaluate the research funding investigators and institutions need to facilitate open research data, and provide adequate support for these activities. NIH has applied open sharing to various projects, typically those with large multi-institutional programs where ground rules are equally applied to all participants. Experience with these projects will be useful in considering what additional support may be necessary. NIH should also evaluate costs and methods of engagement for additional data preparation by investigators providing scientific data that may be needed beyond an award period. NIH should ensure incentives are appropriate for investigators who agree to share data and should support and fund storage and sharing as part of the grant process.

Beyond these initial projects where open sharing has been applied, we believe that many investigators and institutions are limited in their experience and ability for developing appropriate data sharing plans. This can make the initial implementation of the policy difficult and may lead to inconsistent implementation while practices in effective data sharing evolve. Therefore, NIH should prioritize and recommend timelines for measures that reflect data sharing performance and institutional compliance, from plan development to actual data sharing. Either the provisions should specifically define the prioritization and timelines, or the provisions should reference guidelines that may be adapted over time. Of these options, the use of guidelines may be more flexible and better adapt to changing knowledge of best practices.

NIH should also fund additional research on research support operations and technical platforms that could improve the efficiency of developing and implementing an appropriate data management and sharing plan. NIH should also support research and training on how to best manage and share data under this new policy. Such support will be important in accelerating the discovery best practices.

Portions of the draft policy provisions pertaining to the acceptance of appropriate data management and sharing plans prior to awarding funding are reasonable, as long as the

requirements for the plans are specified in advance in funding guidelines, and appropriate support is provided to investigators in developing plans. Investigator adherence to and advance of best practices in data sharing may be best achieved if the appropriateness of plans is considered more directly in scoring for funding, rather than just as a condition for funding. This could be implemented as a specific question during peer review or may be more effective as a component of the other dimensions for scoring (e.g., Approach).

Specific expectations for data management and sharing plans can improve the overall data management of proposals, which is an ancillary benefit to the provisions. Clarity of expectations and best practices is also needed regarding informed consent. Participants in research studies have a right to know how data will be shared and should be informed during the consent process. The provisions also need to consider how protected health information should be considered in sharing plans. Appropriately managing privacy requires better guidance, as asking for broad consent for subsequent data use places the burden of protecting privacy on research participants who must decide up front whether or not to consent to any potential uses. This is an important risk, because if data sharing requires protected health information sharing, consent may be more difficult to obtain from participants, particularly those who may have reasonable concerns about use of their data too broadly. Elements of good data stewardship and fair information practice principles should be included as well.

The current provisions suggest that data management and sharing plans could have a two-page limit. We appreciate the intent of the suggestion to make the plans concise; however, such a limitation may be difficult practice without precise guidance in how different research elements should be addressed. For example, typical data use agreements are well beyond the recommended page limit, even when including only those sections most relevant to the proposed policy. Without clear expectations, the recommended page limit may not be useful.

The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

As mentioned above, flexibility is needed in implementing these provisions, as they can represent a significant burden to investigators and institutions until standard practices for data management and sharing emerge. In addition, specific guidance is needed for investigators who have little experience and infrastructure to support open science data. For this reason, we believe guidelines should be referenced, established and frequently updated as standard practices evolve. NIH should also support research to improve discovery of effective practices, and should support the effective dissemination of these best practices to investigators, institutions, and institutional review boards (IRBs). Special consideration for dissemination should be given for IRBs, who would have a greater burden with implementing step-wise recommendations.

The most important consideration for timing is that these requirements should apply only to new funding awards, and to awards where the requirements for data sharing can be specified in funding guidelines. First, data sharing considerations should be understood before the consent

process is defined, as participants have a right to know in how their data may be shared. Second, while it is reasonable for the sharing plans to be acceptable as a condition for funding, the negotiation of what constitutes an effective plan and terms should not be done individually as a condition of an award.

In terms of phased adoption for different types of research or funding mechanisms, we recommend the following considerations. First, NIH has for various projects applied open data sharing, and these have typically been large multi-institutional programs. Extending implementation first from these groups seems reasonable. Second, clinical and translational research domains may be appropriate for early implementation of the provisions, due to the proximal impact the findings may have on direct application to health. Third, training programs and small grants programs may be more appropriately considered for later implementation, with larger programs prioritized to mitigate additional reporting and management burdens. Fourth (and perhaps most important), the policy should be implemented as quickly as is reasonable, as there are clear benefits to open research data.

Finally, NIH should advance the infrastructure for storage of shared data. Current recommendations are for considering repositories available at no cost for extended periods of time. Without NIH support, such repositories may become either compromising or unsustainable. If investigators and institutions agree to share data, the actual storage of that data to enable sharing should not be a primary concern of the data contributors.

Additional considerations

These are additional considerations for the provisions that were noted by our members.

More clarity is needed regarding “Compliance and Enforcement.” The concept of making data available to the scientific community “as long as it is useful” lacks definition, is highly subjective, and is problematic. Investigators cannot require unlimited consent from research subjects, and this as written is similar to requesting unrestricted access to data. Terms of the agreement for sharing should be clear and unambiguous.

We support the goals of open research data and these efforts to advance them by the NIH. We also recognize that there will be situations where due to licensing restrictions or privacy risks, some data may not yet be appropriate for sharing. However, we hope that exclusions will be seen as exceptional and that the burden of proof should be on the investigators or institutions to define the exclusions, rather than an equal burden to exclude or include data. Data sharing currently carries sufficient burden in implementation that it should be facilitated, and the goals of open research data for NIH-supported research should be promoted.

In Provision #6, “Data Sharing Agreements, Licensing, and Intellectual Property” under “Requirements for Data Management and Sharing Plans,” the term “intellectual property” was noted as unclear, difficult to define, and problematic. It will be difficult for researchers and the NIH to always have a working agreement about the meaning of the term, which will create issues. Instead, the NIH should define precisely what it intends in regard to the various legal

rights for research data. Explicitly defining a term like “proprietary interest” may be a better approach.

As alluded to prior, it is important to increase consideration of the research participant or subject with regard to sharing data. Studies have shown that patients’ perspectives on appropriate use of data can be very different than that of investigators. These provisions are focused on the researcher perspective, which is appropriate given the context for implementation, but additional consideration is needed of the subject perspective. How subjects and patients understand data sharing in the context of data privacy, ownership, and consent will be critical for successful implementation. Such perspectives will be important for defining governance for appropriate use of shared data, which also seemed lacking in the provisions. Researchers may be able to advocate in the data sharing plan elements for protection of proprietary interest, but the provisions are not clear in defining protections for appropriate use from patient perspectives and ensuring protection of privacy.

Conclusion

AcademyHealth appreciates the opportunity given by NIH to provide input regarding these proposed provisions for a new data management and sharing policy. Our organization has a shared interest in the goals of the policy and its successful definition and implementation. We believe our organization and members can be helpful in disseminating guidelines and best practices for data sharing plans as they are discovered and as they evolve. We believe a successful policy can achieve many benefits, but one that is not well defined or implemented can actually impede the use of various data for generating evidence that can improve health and the performance of the health system.