Blue Ribbon Panel Enhanced Data Sharing Policy Companion Document

V1.0, 17 January 2017

In 2016, the White House Cancer Moonshot Task Force established an advisory panel of leading experts in a range of fields relating to cancer prevention and treatment. The panel developed recommendations through working groups and consultation with the cancer community. The final Cancer Moonshot Blue Ribbon Report 2016 is available online.

This document identifies pressing policy issues and recommendations developed by the Enhanced Data Sharing Working Group. These recommendations were recognized to be beyond the scientific scope of the Blue Ribbon Panel. We have summarized them below so that they are captured for the US National Cancer Institute (NCI) and international research communities, as they were viewed to be essential to the long-term impact of the Moonshot Program.

We have tagged each of the recommendations with one of *three tiers*:

Tier 1 – Immediately actionable items (by NCI)
Tier 2 – Policies that the Moonshot task force can address (NIH wide policy changes)
Tier 3 – Changes that need legislative action (i.e., items outside NIH control/external to NIH, which need external engagement and initiatives)

I. MOTIVATIONAL APPROACHES FOR DATA SHARING

Critical to any discussion of data sharing is how the data will be de-identified to remove or mask personally identifiable information from individuals to minimize the risk of unintended disclosure of their identity and information. The primary standard for de-identifying personally identifiable information (PII) and personal health information (PHI) is the HIPAA Privacy Rule (45 CFR 164.514). The Privacy Rule outlines appropriate uses and disclosures of PHI and provides mechanisms for using and disclosing health data responsibly without the need for explicit patient consent (via HIPAA Safe Harbor that focuses on 18 data elements and the Statistical or Expert Determination methods). While this provides guidance, there is a need to develop new approaches to protect data, based on both the proposed usage and level of trust for those that use that data. Focus on the Privacy rule alone ignores the fact that data de-identification is a continuum. This includes (1) identifiable data (which have the highest potential of re-identification), (2) masked data with or without identifiable attributes, (3) actively managed data (in which the degree of re-identification is monitored and assessed) and (4) aggregated data (in which individuals are unable to be identified). We must consider the entire continuum in our considerations of potential

policy changes.

We must consider data sharing in the context of the entire research life-cycle. Rather than just focusing on data after it has been generated, we must consider how to motivate researchers to consider data sharing as early as possible, optimally before the data is generated. This can ensure best practices with regard to data management and data standards, maximizing the potential for reusability and interoperability. In addition, we need to consider the potential value of the data (i.e., for use in secondary analyses to confirm findings, to explore different research questions, or to develop or refine analytic methodologies) and assess the costs associated with data sharing, including deposition, management, and access.

Motivate Patients to Engage Throughout Research Life Cycle

In order to motivate patients to engage in research and participate in studies, we need to address concerns by patient groups about lack of transparency and understanding of the benefits of collaboration and participation, in addition to the potential risks.

Recommendations:

1) Require researchers to provide patients with results of their research in understandable, lay terms. To facilitate this, provide guidelines and templates on best practices for communicating results. Support research into approaches for how to better educate the public on the benefits of data sharing and re-contact. (Tier 2, AACR)
2) Consider strengthening the Genetic Information Nondiscrimination Act (GINA) to all forms of insurance, including life insurance. (Tier 3, AACR)
3) Provide standard consents (or consent clauses) across multiple contexts that can be used by a broad range of studies and make it easy for patients to understand how their data will be shared. (E.g. "*You agree and understand that data you contribute to this study will be made available to researchers according to the following guidelines...*".) (Tier 3, AACR)

Motivate Researchers to Provide Early Wide Data Availability

Attribution: The current tenure and funding processes are a fundamental roadblock in data sharing. While we recognize the policies around this vary by institution, there are common themes, particularly around publication and attribution.

Recommendation:

4) Credit should be given for any type of contribution such as data generation, data curation, data analysis, and providing reusable code and workflows (available via code and tool repositories, executable in cloud environments, and referenced

using Digital Object Identifiers (DOIs)). For example, contributors could be awarded points that are noted during evaluation of grants. Points can be awarded for tools according to the frequency with which they are used and the ability to use them to easily reproduce results of publications. One suggestion (based on ideas from NCI EGRP and NSF work, e.g., Ioannidis and Khoury [PMID:24911291]) is to develop a data/tool-sharing index (S-index) for a researcher, following the commonly used H-index for evaluating citations of publications. As another example, journals could be encouraged to allow separate author lists (with first/last authors) for data curation, data generation, data analysis, etc. This is being done in a limited way by, e.g., *Science Magazine*, and supported by new standards efforts. (Tier 2 but requires collaboration outside of NIH with journals etc., AACR, NSF)

NIH Funding Structure: Current NIH data sharing plans don't result in meaningful sharing of data for a number of reasons. Reviewers of NIH grants don't necessarily have the technical expertise to review and evaluate data sharing plans. There are also no formal mechanisms for follow-up on the execution of and adherence to data sharing plans. Similarly, there is no mechanism to ensure sharing of analytical tools that together with the data can reproduce the results. The current structure of NIH grants could be changed in order to promote data and tool sharing in a meaningful way.

Recommendations:

5) Implement an NIH data sharing plan approval process. Similar to human subjects review, a failure to have an adequate data sharing plan in a grant proposal will result in rejection of the application. (Tier 2)
6) Implement a "no sharing, no payment" policy. This could be implemented as staged rewards throughout a grant period (e.g., researchers get paid for phase k only after they have shared the data, tools, and results of phase k-1) following explicit deliverables in the data sharing plan. To allow for rapid progress through each phase, data could be deposited with a timer for public release after a fixed period of time (e.g., 6 months), allowing the researcher exclusive access for that time. (See examples from astronomy.) A metric for data sharing compliance could be created and required on biosketches. In addition, researchers need incentives to work from the beginning in a collaborative cloud or commons platform environment, which would facilitate release of data, tools, and workflows. For large-scale RFAs, this could be part of the terms and conditions of funding. We should also leverage non-traditional data science and information stewards in the review of data sharing plans. (Tier 2, AACR, NSF)

Motivate Insurers and Clinical Labs to Make Test Results Widely Available

Recommendation:

7) Develop policies with insurers, clinical lab accreditation agencies, and regulators to **require data sharing of all variants (both benign and pathogenic)** as a requirement for reimbursement, lab accreditation, or streamlined review of lab tests approval (recent draft guidelines from FDA). As a component of these policies to facilitate transparency, labs should provide confidence metrics with respect to accuracy and consistency of these results. Provide guidance for CMS and Palmetto on a data sharing policy that includes resources to work with policy makers on how to articulate the importance of data sharing. (Tier 3, AACR)

Facilitate IRBs to Support Enhanced Data Sharing

IRBs currently unwittingly introduce constraints that make it difficult for participants to share their data to the extent they want to share them.

Recommendation:

8) Make sure policies assist IRBs in preventing addition of constraints on data sharing that do not reflect the wishes of participants. Identify existing practices that introduce these constraints and educate IRBs to avoid these practices. (Tier 3, AACR)

II. MECHANISMS FOR DATA SHARING

Licensing

Aggregating data can be significantly impeded by the heterogeneity and diversity of licenses or terms-of-use associated with the source data and the legal hurdles they present. This is not just an impediment in the current model of independent data providers but also would serve as a major barrier for a common platform given intellectual property and other related concerns. This applies not only to the original data but also to derivatives of that data, impacting the potential for computational approaches and tools focused on knowledge discovery, meta-analyses, and confidence metrics or evidence. The NIH should support research and help in harmonizing agreements in a way that dependencies can be managed (e.g., Model Data Use Agreements, etc.).

Recommendation:

9) Adopt (at a minimum) a license following the Uniform Biological Material Transfer Agreement (UMBTA) with multiple opt-in signatories for data generated by NIH funding. While it is recognized that UMBTA does not handle revenue sharing terms or auditing / enforcement, a task force should be created to examine alternative models such as those used in government-convened standards consortia such as [F]RAND licensing or in Industry-led standards consortia (RF or RAND licensing) that can provide path for licensing revenue and interoperability. Similar considerations should be applied to tools and code, although developers are more familiar with code licenses and they are more commonly used compared to data licenses. (Tier 2)

Privacy

Better articulation and guidelines around what can be shared are needed. For example, HIPAA/privacy issues are used as reasons to not share data, even when it may not be applicable (e.g., is genetic information PHI regardless of source?). This is further complicated by the fact that patients should be able to control their own data. The Enhanced Data Sharing Working Group participants have emphasized the need for patients to be able to access all of their own data and easily consent for its use in research studies, perhaps by using digital signatures. There are model Initiatives such as BlueButton and Sync for Science that are working on ways for patients to donate their data to research.

Recommendations:

10) Develop explicit guidelines regarding privacy and data sharing. Create policy in which researchers are required to provide patients with a copy of their data (e.g., test results, medical records, and image data). Create policy that not only allows patients to freely contribute their data to research but also focuses on human-data interaction (agency, negotiability, transparency). (Tier 3, AACR)
11) Develop and implement electronic trackable consents with clear and simple individual-level preferences for sharing data that can be dynamically modified or revoked. This may require research on how to verify that patients understand the risks and benefits of contributing their data. Require that all entities holding data accept a digital signature.

Common Data Schema/Ontologies

One key barrier to effective analysis of data is heterogeneous data schemas/ontologies. A uniform schema/ontology would make analyses much more efficient and informative. Multiple schemas exist today, in part, due to a lack of central coordination. However,

there are also fundamental differences in these ontologies because different researchers have framed diseases and findings in different ways to suit their theories of disease pathology.

Recommendation:

12) Incentivize the community to adopt a common model at some level of abstraction (e.g., caDSR roadmap and EVS). Focus is not on biological categorization, but rather on data sharing and data integration. Data sharing tooling and measures of 'meaningful data sharing' through interoperable data representations will inform the scientific recommendations. (Tier 1 and Tier 3)

During the drafting of these recommendations, we consulted and borrowed from these other recommendations on policy issues:

(i) AACR Workshop July 23-24, 2016

(ii) National Science Foundation (NSF) Cyberinfrastructure Task Force Report (March 2011)

(iii) National Institutes of Health (NIH) Plan for Increasing Access to Scientific Publications and Digital Scientific Data (Feb 2015)

(iv) Wellcome Trust Expert Advisory Group on Data Access Governance of Data Access (June 2015)

(v) National Human Genome Research Institute (NHGRI) Policy, legal and ethical issues in genetic research

(vi) Research Data Alliance, BioSharing Registry: connecting data policies, standards & databases in life sciences Working Group

(vii) Lawler M et al. "All the World's a Stage: Facilitating Discovery Science and Improved Cancer Care through the Global Alliance for Genomics and Health (GA4GH)", Cancer Discov (2015).

(viii) Biotechnology and Biological Sciences Research Council (BBSRC) Data Sharing Policy (March 2016)