Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Lyric Jorgenson
Acting Associate Director for Science Policy
Office of Science Policy
National Institutes of Health
6705 Rockledge Dr Ste 750
Bethesda, MD 20892

February 26, 2022

Re: Request for Information on Proposed Updates and Long-Term Considerations for the NIH Genomic Data Sharing Policy (NOT-OD-22-029)

Dear Dr. Jorgenson,

Thank you for your continued development of the NIH Genomic Data Sharing Policy ("GDS Policy"). The GDS Policy and its predecessors have provided an essential foundation for maximizing the benefits of and ensuring the robustness of NIH-funded genomic research. The GDS policy also provides global leadership in genomic data sharing and establishes norms that influence the sharing of genomic data and other biomedical data worldwide. A revised policy will enable improved data sharing and the acceleration of scientific progress, building on NIH's strong record in open data.

We submit this response to your proposed updates of the policy on behalf of the Global Alliance for Genomics and Health (GA4GH). GA4GH is an international, nonprofit standards organization formed in 2013 to accelerate the potential of genomics research and medicine to advance human health. Bringing together 660+ leading organizations across 90+ countries working in healthcare, research, patient advocacy, life science, and information technology, the GA4GH community collaboratively develops policy guidelines and technical standards to enable the responsible, voluntary, and secure sharing of genomic and related health data.

GA4GH products provide a standard, interoperable, community-driven framework for achieving a global ecosystem of responsible genomic data sharing. A secure, privacy-ensuring framework should be established that balances risks with the benefits of scientific advances from data linkage. This ecosystem should support data provenance and ensure appropriate attribution for data donors and stewards.

Many successful examples already demonstrate this vision in practice within the NIH ecosystem. For example, the NIH Cloud Platforms Interoperability (NCPI) effort has shown the utility in making ~11 PB of high quality research data accessible across AnVIL, BioData Catalyst, Cancer Research Data Commons (CRDC), and Kids First. NCPI employed GA4GH products—including GA4GH Passports via the Researcher Auth Service (RAS), and the GA4GH Data Repository Service (DRS) standard—to provide access to authorized researchers. Our suggestions below build on these and other data sharing successes around the globe, NIH's continuing use of GA4GH products in facilitating this success, and reflect the consensus-driven recommendations of our international community. We called upon our broad international network to provide feedback on the GDS Policy, and refined our collective feedback through a series of virtual and in-person discussions. The recommendations reflect the overall work of GA4GH and our ongoing mission to employ international genomic data sharing in order to activate the right of all humans to benefit from science, as outlined in Article 27 of the UN's 1948 *Declaration of Human Rights*.

Below, please find responses to selected individual requests for input.

I. Maximizing Data Sharing while Preserving Participant Privacy and Preferences

3. Data linkage

The ability to link data across datasets, sources, and institutions is critically important for research and healthcare. Specifically, the ability to create a "union" record of one participant, merging records from multiple independent datasets, is essential to garner a complete picture of any participant in health research, as most participants receive healthcare in multiple places. The de-duplication inherent in this process is also essential to minimize analytical biases, as unknown duplicate and unmerged records will reduce the accuracy of population estimates of prevalence of genotype or phenotype, or the rate of association between the two.

Privacy-Preserving Record Linkage (PPRL) enables "putting the patient back together again," joining different data modalities for a given participant (https://doi.org/10.1109/TCBB.2018.2855125). For example, a researcher could use a PPRL to unify genomics, imaging, and clinical phenotypic data where those data live in different repositories. PPRL allows for more robust cohort discovery and analytics, since the data about a participant and a set of participants is more complete.

Large-scale data of different modalities will continue to reside in different repositories, which will drive an ongoing technical and ethical need to join data across sources. The GDS Policy should permit data linkage of biological and health data whenever data has participant consent for research and is de-identified.

When one or more data sources do not meet all GDS Policy expectations for de-identification, this provides a greater challenge, as identifiable information in one data source increases the risk of re-identification of other linked datasets. When linking with data sources that do not meet all GDS Policy expectations for de-identification or consent, researchers and IRBs need to take extra care in considering risks to study participants, and disclose these risks to participants, when possible.

PPRLs generated from sensitive participant data should be regarded as sensitive data themselves, and protected accordingly. PPRLs should be derived with a cryptographic method, with security and privacy impact carefully assessed. Biometric data, which is subject to numerous legal restrictions and prohibitions, should not be used as input to the PPRL hash. Data storage systems that implement PPRLs should support the right of participants to withdraw data, without the risk of the data re-emerging. Importantly, PPRLs should not emerge as "universal identifiers", as this would allow linkage across institutions without oversight, and would present the risk that a breach in one system extends to vulnerabilities in multiple systems.

It seems ethically acceptable to link two consented research datasets, with appropriate risk assessment. Such risk assessment can be estimated from controlled-access data or synthetic data via assorted metrics (https://cloud.google.com/dlp/docs/compute-risk-analysis), or through adversarial analysis.

Linking a research dataset post-deposit with unconsented clinical data should only be allowed when an IRB deems it of minimal risk and when consent cannot be obtained. Where possible, however, it is ethically preferable to obtain consent from participants to data linkage at the time of the initial study.

4. Consent for data linkage

To maximize the utility of collected data, ideally investigators will inform study participants of the possibility of data linkage, even if investigators do not plan any data linkage at the time of consent. We recommend using the language from the GA4GH Regulatory & Ethics Toolkit (https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/) Consent Clauses for Genomic Research (https://ga4gh.org/consent-clauses-for-genomic-research-docx/):

> Will my information be linked with any other data?
> [Name of database] will link the [information] and [information] you have contributed as part of this study with your sequencing and other data.

Explicit consent for linkage may not be required when the risks associated with data linkage are low, especially when the individual datasets involved all have research consent. It should be clear that lack of explicit consent for linkage does not prevent data linkage, but only enhances the need for researchers and IRBs to consider risks to study participants, and disclose these risks to participants, when possible. When doing so, we recommend that researchers also inform the participants of the goals and benefits of such linkage.

II. Expectations for Alternative NIH-Supported Genomic Data Management and Sharing Resources that Store Human Genomic Data

5. Data management and sharing principles for NIH-supported resources

b. Any aspect of the principles described for Data Access.

The principles proposed by NIH focus on legal and technical aspects of data access agreements that follow an approval in principle to share controlled-access data. Throughout the existence of dbGaP and other controlled-access repositories, some researchers seeking to access controlled-data have encountered frustration with and delays in research due to unclear criteria to get that approval in principle in the first place.

NIH should require that grantees make clear the requirements for data access and the procedures by which data access committees (DACs) evaluate data access requests. We recommend the GA4GH Regulatory & Ethics Toolkit (https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/) Data Access Committee Review Standards (DACReS) Policy (https://www.ga4gh.org/wp-content/uploads/GA4GH-Data-Access-Committee-Guiding-Principles-and-Procedural-Standards-Policy-Final-version.pdf) to guide DACs and those establishing them. NIH should require that grantees establish DAC policies that include the elements described in the DACReS policy, including Terms of Reference formalizing the membership and authority of the DAC, Standard Operating Procedures, and Criteria for Assessing Access Applications. Grantees should be required to deposit these policies transparently and publicly in a third-party repository that mints data object identifiers (DOIs), such as Zenodo.

As custodians of data collected with public funds, DACs must focus their criteria for assessing access applications on ensuring that the applicants meet ethical and legal standards, and promoting the interests of study participants and science generally, rather than the narrow interests of individual researchers or institutions. To ensure this, NIH should require that DACs operate at arm's length from study investigators and that those involved in a study must not take part in decisions on access to the study data. An IRB or Institutional Animal Care and Use Committee (IACUC) member would not be allowed to take part in ethical review of their own study due to the obvious conflict of interest. The same principle applies to DACs.

The DOIs of relevant DAC policies must be available in the Data Management and Sharing Plan and in manuscripts that refer to the data. The Data Management and Sharing Plan and manuscripts should also include a description of permitted purposes for the data using the GA4GH Data Use Ontology (DUO; https://github.com/EBISPOT/DUO).

III. Policy Harmonization

6. Harmonizing GDS and DMS Policies

It is desirable to have some harmonization of administrative processes for review of data management and sharing plans. For example, avoiding the need to submit two redundant plans to satisfy both the GDS and Data Management and Sharing (DMS) policies is an improvement.

Nonetheless, it is essential not to reduce substantive requirements of the GDS Policy to the lowest common denominator in the DMS Policy. The GDS has served as an exemplar for data sharing policies for 7 years, and its requirements are well-accepted by researchers who produce genomic data and relied upon by researchers who independently analyze genomic data. Reducing substantive requirements for genomic data sharing after many years of established practice will disrupt genomic research.

NIH specifically asks about changing the current policy which has non-human data as a subject and restricting the GDS Policy to human data only. We oppose this change. Research and data on non-human organisms are no less important to biomedical research than research and data on humans. There is no justification for this change, which will add needless friction to research on non-human genomics. In fact, the lack of privacy and controlled-access data considerations means it should be easier for those working with non-human organisms to achieve compliance with the GDS Policy.

To achieve increased harmonization between the GDS and DMS policies, we recommend that NIH strengthen the next revision of the DMS policy to match the more stringent requirements of the GDS policy. This would bring the advantages of harmonization without reducing to a lower standard an effective policy that has worked well for years.

7. GDS and DMS data sharing timelines

GA4GH strongly opposes relaxing the current timelines for prepublication sharing of large-scale genomic data, such as submission of cleaned data within three months of data generation. The timelines in the GDS policy derive from international agreements in the landmark Bermuda Principles of 1996 and subsequent Fort Lauderdale Agreement of 2003. Reducing these expectations after a quarter-century of a synergistic ecosystem between data producers, methods developers, and other researchers that has brought great benefits would be regrettable. The COVID-19 pandemic has made it clearer than ever before that not only is rapid prepublication data sharing and preprint sharing achievable, but it has improved scientific collaboration, expedited the dissemination of scientific results, and has saved countless lives.

NIH justifies relaxing the establishing timelines because they "have posed challenges for compliance". Clearly, the challenges for achieving compliance at the end of the performance period will be much, much greater. At the end of the performance period, direct funding for sharing activities are over and the NIH loses the ability to enforce compliance through suspension of an ongoing award. The ability of NIH to achieve widespread compliance with the new DMS policy at the end of the performance period is unproven.

IV. Long-Term Consideration of the Scope of GDS Policy

8. Types of research covered by the GDS Policy

a. Whether there are other types of research and/or data beyond the current scope of the GDS Policy that should be considered sensitive or warrant the type of protections afforded by the GDS Policy

As noted above, we believe the GDS Policy serves as an exemplar policy for many kinds of data sharing. We agree that human proteomics data, and other data types where re-identification risk is similar to genomics data, should be considered sensitive and warrant the protections afforded by the GDS Policy.

b. Whether small scale studies (e.g., studies of fewer than 100 participants) and those involving other data types (e.g., microbiomic, proteomic) should be covered under the GDS Policy, and if training and development awards (e.g., F, K, and T awards) should be covered by the GDS Policy

We support adding non-genomic data types commonly referred to as "omics" data to the GDS Policy. A non-exhaustive list of these data types should include proteomic, metabolomic, microbiomic, lipidomic, and radiomic data. The GDS Policy's definition of genomic data already includes other sorts of omics data, such as transcriptomic or epigenomic data. The GDS policy should also require including accompanying metadata, including clinical phenotypes critical to the use of genomic and other omics data. When collected at scale, the GDS policy should also include clinical data as a primary data source.

Some non-genomic data types discussed above, such as proteomic data, include information about human genetic variants that could be used for re-identification, and so should be considered sensitive. Other data types, such as metabolomic data, carry much less risk of re-identification, and therefore should not be considered sensitive in the same way. NIH should state in its guidelines that these data types are much less likely to require controlled access.

We support having the GDS policy cover small-scale studies, and studies funded by training and development awards. It would be acceptable to use the timelines in the general DMS Policy instead of the accelerated GDS Policy timelines for small-scale studies only. This should only depend on whether the study is large-scale or small-scale, and not on the funding mechanism. Otherwise, the standard for the extent and nature of expected data sharing for omics data should be the same for small-scale studies.

c. Whether NIH-funded research that generates large-scale genomic data but where NIH's funding does not directly support the sequencing itself should be covered by the GDS Policy.

All NIH-funded research that generates large-scale omics data should be covered by the GDS Policy. To exempt research based on the source of funds for sequencing itself in a project otherwise paid for by NIH would be a huge loophole that would prevent full utilization of the data. This should not be allowed.

9. Data sharing expectations under the GDS Policy

NIH should absolutely apply the GDS Policy beyond genomics to other large-scale data. Other data types referred to as “omics” and related clinical data and metadata, are prime candidates for application of the GDS Policy. Beyond omics data, NIH should consider applying similar expectations to other awards where generating large-scale data or a data resource is a major focus. One can identify such awards either through language in funding opportunity announcements or through specific aims focused on the generation of large-scale data or data resources. The Office of Extramural Research should review submissions of funding operating announcements and Notices of Special Interest (NOSIs) for the NIH Guide for Grants and Contracts to identify those where enhanced data sharing requirements should apply.

Conclusion

Thank you again for your continued development of the GDS Policy. The GDS Policy has a long track record of success. We support efforts to build on this success by strengthening the policy and increasing the number of areas to which it applies.

Sincerely yours,
the GA4GH Executive Committee

Heidi Rehm Kathryn N. North

Ewan Birney Heidi Rehm Kathryn North Peter Goodhand

on behalf of the GA4GH Steering Committee.

Response drafting group:

- Michael Hoffman (University Health Network; University of Toronto)
- Michael Baudis (University of Zurich; Swiss Institute of Bioinformatics)
- Ewan Birney (European Molecular Biology Laboratory—European Bioinformatics Institute)
- Anthony Brookes (University of Leicester)
- Melissa Cline (University of California, Santa Cruz)
- Manuel Corpas (Cambridge Precision Medicine)
- James Eddy (Sage Bionetworks)
- Robert R. Freimuth (Mayo Clinic)
- David Glazer (Verily Life Sciences)
- Melissa Haendel (University of Colorado Anschutz Medical Campus)
- Brian O’Connor (Broad Institute of Harvard and MIT)
- Angela Page (Broad Institute of Harvard and MIT; GA4GH)
- Heidi Rehm (Broad Institute of Harvard and MIT; Massachusetts General Hospital)
- Augusto Rendon (Genomics England)
- Adrian Thorogood (McGill University)