Compiled Public Comments on Request for Information: Proposed Updates and Long-Term Considerations for the NIH Genomic Data Sharing Policy

Guide Notice Number: NOT-OD-22-029

November 30, 2021 – February 28, 2022

Table of Contents

41. Michael Saito, Epic
42. Mary Lee Kennedy, Association of Research Libraries
43. Brian Scarpelli, Connected Health Initiative
44. Regulatory Intelligence, Regeneron Pharmaceuticals, Inc.
45. William B. Coleman, PhD, American Society for Investigative Pathology
46. Glenn Martin
47. Elizabeth A. McGlynn, PhD., Kaiser Permanente
48. Gretchen Purcell Jackson, MD, PhD, American Medical Informatics Association
49. Denise Dillard
50. Mette Peters, Sage Bionetworks
51. Joseph M Yracheta, Native Bio-Data Consortium
52. Pam Dixon, World Privacy Forum
53. Anonymous
54. Steven E. Brenner
55. Emily Bonkowski, CGC
56. Janis Geary
57. Kevin Mcghee, New York Genome Center

ID: 1793

Submit date: 12/9/2021

I am responding to this RFI: On behalf of myself

Name: Brent Wilkerson

Name of Organization: Medical University of South Carolina

Type of Organization: Nonprofit Research Organization

Role: Scientific researcher

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

Single cell RNA-seq datasets in GEO are frequently sliced or transformed in ways that make them useless for integrative analysis. Some investigators upload datasets with cells removed and with counts normalized or transformed in unspecified ways, possibly to prevent competitors from using their data. In some cases, BAM files can be downloaded and realigned to recover missing cells and raw counts, but this should be unnecessary. Realignment is very computationally intensive and wastes time. Please update standards for genomics data sharing to require upload of raw counts tables without any filtering.

ID: 1794

Submit date: 1/2/2022

I am responding to this RFI: On behalf of myself

Type of Organization: University

Role: Scientific researcher

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Studies involving other data types (macrobiotic and proteomic) should not be covered by the GDS policy. These are not human identifiable data.

ID: 1795

Submit date: 1/3/2022

I am responding to this RFI: On behalf of myself

Name: Erik Imel

Name of Organization: Indiana University

Type of Organization: University

Role: Scientific researcher

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

It should be acknowledged that in the setting of rare diseases, some information may become more readily identifiable, inadvertently. There need to be adequate safeguards for example such that if there are only a handful of people with a disorder that it may be easier for third parties to determine who that person is, which may infringe on participant privacy. We have had persons refuse to participate in studies because of concern that their information would end up in a government database, which is to the detriment of research in the rare disease, as every participant is very valuable in that scenario.

ID: 1796

Submit date: 1/13/2022

I am responding to this RFI: On behalf of myself

Type of Organization: University

Role: Scientific Researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

The crux of this issues is the assumption that genomic data may in the long run be more "identifiable" than currently expected. To address this concern, the controlled access database should be revised such that investigators are obligated to maintain privacy. The hurdles to data use should not be administrative and/or review of research plans. Instead, each applicant from a validated research institute should agree to the terms of data sharing, with the goal of protections of privacy.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Data linkage is critical for broad assessments across populations. These linkages should be available to researchers. Again, if the investigator/institute are bound to follow guidelines through contractual agreement, the concerns with privacy of large cohorts may be mitigated. Such agreements should be designed for rapid and easy enactment so research is not blocked/delayed significantly.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

dbGaP is unwieldy and takes many months of significant back and forth to post data; in contrast other biorepositories are much more streamlined and easy to use. This issue needs to be addressed. For other data management resources, and data sharing, it is critical to avoid additional steps that prevent reaching the goal of more data sharing. For example, requiring that the repository obtain a data submission agreement seems unnecessary, as we already have IRBs, etc that monitor data sharing. In general, the process should be minimal and streamlined. Requiring that outside repositories follow dbGaP's, for example, may result in less data sharing.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

It needs to be clarified what "data generation" means, and what type of data NIH would like submitted. As an example, for scRNAseq data, three months after sequencing of the full cohort, one would really just have the raw data. QC, validation, and endstage annotation takes time. Cancer cell annotation/identification in scRNAseq data is hard—non-cancer tumor associated normal cell annotation is even harder. So does linking clinical and experimental data—and often full clinical outcome data can be delayed compared to the genomics. Three months is not sufficient time to perform the QC and annotation needed for posting the data. As the annotation and accurate classification of the cells is critical for downstream analyses, there is value in releasing the data with well-vetted and validated information. Indeed, it may take significantly more than three months to get through data compilation and dbGaP paperwork and approvals (the dbGaP process itself averages over six months for our cohorts). Further, as it takes significant effort to post the data through dbGaP, an iterative process where we have to update the data multiple times is not practical--what will likely happen if there is a three month mandate is that data is posted in an immature form, and not updated with relevant critical information.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Absolutely small scale studies and other data types are also important to share with the community and should be included in the GDS policy. There is no rational basis to only include a subset of data types. Projects linked to NIH funded research should also submit their data, or alternatively, ensure that the entire project is independent of NIH funding.

ID: 1804

Submit date: 1/20/2022

I am responding to this RFI: On behalf of myself

Type of Organization: University

Role: Scientific Researcher

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

Yes, please harmonize with DMS policy. The 3 month policy is unrealistic regarding allowing the initial researchers time to process/analyze/interpret/publish initial findings.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Please don't make broad sharing mandatory for small-scale projects. Folks with rare genetic disease could easily be identifiable by the WGS or WES, and don't always want to consent to broad sharing of their genomes- so such requirements impair research into rare disorders. Please don't make broad sharing required for sequence that is not generated with NIH funds. Using nonNIH funds is sometimes the only way to enable study of rare disorders and patients/families who do not want their DNA sequence posted on the web for %anyone to peruse. In this regard, I'll also mention the policies of NHGRI. We are no longer able to participate in the current incarnation of the Centers for Mendelian Genomics, as that initiative now requires 'broad' data sharing, rather than sharing to databases with controlled access- and we're no longer allowed to specify the range of conditions for which the sequence can be accessed. Given our consent forms specify the disorders, and our IRB has only approved sharing for study of specific ranges of disorders, samples from these consented participants can no longer be used in CMG-related studies. Constantly re-consenting to keep up with every change isn't feasible, some participants don't want their sequence available to anyone for every possible purpose, and precious legacy samples can no longer be studied with such funding.

ID: 1813

Submit date: 1/28/22

I am responding to this RFI: On behalf of myself

Name: Wayne Shreffler

Name of Organization: Massachusetts General Hospital

Type of Organization: Health Care Delivery Organization

Role: Scientific researcher

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

I have had trouble finding guidance on genomic data sharing that appears to distinguish data sets with respect to their inherent risk for de-identification. It does not seem that whole genome sequencing, consisting of long sequence reads for the purpose of identifying germline variations, should be treated in the same manner as very short read transcriptomic studies (e.g., RNA-seq), which will generate very little data on individual germline variation and therefore carry low risk for de-identification.

ID: 1814

Submit date: 1/28/22

I am responding to this RFI: On behalf of myself

Name: Charles Warden

Name of Organization: City of Hope National Medical Center

Type of Organization: Nonprofit Research Organization

Role: Scientific researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

I believe there is room for improvement in making clear to the scientific community that i) consent is important for sharing genomic data from cell lines and ii) the raw data other than Whole Genome Sequencing and SNP chip genotypes can in fact identify the subject (relative to other data made directly available to the subject, such as uploaded 23andMe data). There are some situations where I thought the data for public data deposits was being removed out of an abundance of caution, which I agree with. However, to give examples where I could successfully self-identify myself with genomic data for myself: **Example 1)** Low-Coverage and or Limited Genotype Self-Identification - I needed to sign a HIPAA release to obtain what might be considered a relatively small number of genotypes (from Color) - If selecting the sufficiently informative genotypes, I could get clear separation of relatedness estimates with less than 2,000 genotypes. - If using low-coverage Whole Genome Sequencing data, I would say I had a good chance of self-identification with at least 1 million paired-end 100 bp reads. - If family-level identification counts, perhaps even 0.1 million reads would be enough, and I think caution for future developments and/or other analysts is worth taking into consideration. Full summary available here: https://cdwscience.blogspot.com/2020/03/testing-limits-of-self-identification.html **Example 2)** Exome Self-Identification with 23andMe Genotype Data Some decrease in performance relative to comparing 23andMe genotypes to Whole Genome Sequencing data, if directly using on-target GATK Exome variant calls. However, kinship estimate already not horrible, and other methods / analysts may be able to find a way to get an even better estimate with only on-target reads. Also, self-identification relatively clear with the right continental-level ancestry guess and using GLIMPSE for imputation of off-target reads that were within 2,000 bp of coding exons. So, I would consider FASTQ files from Exome and RNA-Seq data to be genetic identifying information. Full summary available here: https://cdwscience.blogspot.com/2020/07/broad-ancestry-predictions-for-my-exome.html All of this analysis performed on a personal computer with 32 GB of RAM and 4 cores. If imputation is not used, the computational requirements are even less. **Example 3)** RNA-Seq Self-Identification While not currently published / public, I have found that samples compared so far appear to match with high relatedness estimates from genotypes when they are from the same subject. I believe that independently matches what is described in this other publication: https://pubmed.ncbi.nlm.nih.gov/31501877/ In general, you can access the raw data from the following locations: https://github.com/cwarden45/DTC_Scripts (various links) https://my.pgp-

hms.org/profile/hu832966 (not completely up to date due to PGP site maintenance, but includes what may be most important datasets paired for myself)

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

b) I currently have concerns that all of the large datasets are not fully enforced to be shared to meet the policies. However, I think the patient consent matters for even small datasets. For example, even if data sharing is not required by the NIH, I would consider samples without consent or with unknown consent to not be acceptable for deposit of raw data. So, if the policy helps ensure the data can be published to meet journal requirements and still respect patient privacy, then I think that is helpful. That said, I have encountered issues with getting local Institutional Certification for cell lines. I am not sure if this is something that the NIH can assist with (to be like HeLa cell lines). While not an example of a small cell line dataset, I frequently assist with data deposit for smaller studies and I have listed some of what I encountered here: https://www.nature.com/articles/s41586-019-1186-3#article-comments **c)** There are some studies (such as with commercial cell lines) that can be deposited into GEO/SRA (even if not meeting GDS requirements), but they can't get local Institutional Certification (at least when I attempted to do so). I don't believe any Non-NIH dbGaP applications that I assisted with were successful, and I believe that process can be somewhat time consuming. I think this means that genetic identifying information without explicit consent is being made public, and I am not sure if one or both of these components can be changed to make that less likely.

ID: 1830

Submit date: 2/2/22

I am responding to this RFI: On behalf of myself

Name: Philip O Alderson, MD

Name of Organization: Saint Louis University School of Medicine

Type of Organization: University

Role: Medical provider

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

To answer questions posed in the NIH release: 1) Yes, permission for current or future data linkage to other datasets that meet GDS specifications should be obtained when patients provide consent to use their data. 2) Yes, it is important to include permission to link to other types of data ,e.g., proteomics , micrbiomics, in the consent form. This is of particular interest and potential with respect to artificial intelligence (AI)-enabled approaches to rapid and accurate assessment of 3D protein structure such as demonstrated by the Alpha Fold algorithm. Such AI approaches offer great future potential to explore 3D protein configuration epigenetics in health and disease. 3) Yes , certainly consent should be obtained for linking to GDS certified and compatible datasets. Linkage should not be requested to data obtained without patient consent or to datasets that might allow patients to be identified. Although access to large amounts of such data might be scientifically rewarding, the potential damage and liability are too great. In addition , such uncertified datasets are likely to prove in part incompatible with current and future data review algorithms and that will increase the likelihood of data capture errors. 4) Consideration should be given to making the revised GDS Policy more explicit regarding data formats for linkage. Since 2014 many new approaches to improved natural language processing (NLP) have been developed. Transformer technology like BERT and its many "offspring" have the potential to allow more rapid and accurate NLP data retrieval and joining from both large and smaller datasets. In addition, newer data compression formats are being applied to genomic datasets. Accordingly, explicit GDS policy directions about acceptable/preferred approaches to data compression, storage and retrieval may be warranted. Thank you for the opportunity to comment.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

I included comments about consent for data linkage in my response to the Data Linkage section above. In short, yes, consent for data linkage should be obtained in the initial consent form . In patient datasets the links should be to data for which patient consent has been obtained and patient identities are

completely concealed. Links to proteomic datasets and related scientific data should be included. Please see above.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

See Data Linkage sections above.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

Datasets that are bioinformatically and computationally compatible will perform better together. The new GDS Policy should be more explicit about data formats and accessibility. See Data Linkage above for relevant comments.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

The DMS Policy approach seems a bit more liberal and should allow greater data linkage. However, genomic expression is a dynamic variable so tight comparison intervals may need to be established for specific comparisons that rely on more biologically dynamic processes.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Yes , what you term small scale studies in this section -- microbiome , proteomic -- definitely should be included. 3D protein folding related to gene expression will be an important scientific and clinical frontier and the GDS Policy and consent approach must be designed to facilitate inclusion of such data. See Data Linkage section above.

Email: philip.alderson@health.slu.edu

ID: 1833

Submit date: 2/2/22

I am responding to this RFI: On behalf of myself

Name: Qunfeng Dong

Name of Organization: Loyola University Chicago

Type of Organization: University

Role: Scientific researcher

De-identification. **The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.**

I support the expanding de-identification options.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

I think that it is acceptable to use potentially identifiable information

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Yes, I strongly support data linkage between datasets

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

When patients were asked to sign the original consent, provide an option for them to consent whether this data can be linked to other consented data .

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

All those NIH-supported resources should provide a digital copy of their database to NIH as part of the backup to ensure long-term lifetime of the data.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

a. Yes, any published study using NIH fund should deposit their data at NIH, even if they are just ugly excels or raw plain text. b. Yes c. Yes

Data sharing expectations under the GDS Policy. **Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).**

Raw data (at least raw sequence data) should be shared, not just processed data, since the bioinformatics processing pipelines consistently get updated.

Email: qdong@luc.edu

ID: 1847

Submit date: 2/3/22

I am responding to this RFI: On behalf of an organization

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/hymwwnOJPG.pdf

Email: tbadmington@primr.org

ID: 1877

Submit date: 2/8/22

I am responding to this RFI: On behalf of myself

Name: Leslie M Thompson

Name of Organization: UC Irvine

Type of Organization: University

Role: Scientific researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

I agree there needs to be expansion of de-identification options. But there needs to be clear language and transparency in hospital and research consents that data may be generated and utilized in this way

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Circumstances could be laid out in the context of consents or when this information might be used. Build into pilots initially to generate trust and understanding. Could align around public good problems. Have option to opt out although that gets difficult after data disseminated in any way

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Data linkage should be a component of GDS policy as that is what will really transform the use of genomic and other potentially identifiable data even with de-identification coding. This becomes tricky if data sets do not meet all GDS policy expectations.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Yes this should be part of transparent consent processes. If NIH had some standard language for broad and ethical consent process, that would be helpful

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

Transparency is key for data that is deposited into NIH supported resources for any potentially identifiable information

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Small studies could also be covered by GDS policy although many of these have project specific consents.

ID: 1879

Submit date: 2/16/2022

I am responding to this RFI: On behalf of myself

Name: Michael Gorin

Name of Organization: David Geffen School of Medicine - UCLA

Type of Organization: University

Role: Scientific researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

There are means of de-identifying data in such a fashion as to preserve relative dates that would not jeopardize the ability to look at sequential events, information and/or disease progression. Such methods should be encouraged and facilitated rather than simply eliminating birth dates and visit dates (for example). Realistically this strategy for de-indentification of time stamps and relevant dates will be imperfect for complete anonymization since one can infer a great deal from relative intervals between dates. For example if an individual receives a rare treatment at a young age. However the benefits outweigh the compromises. Other potential data that has to be de-identified such as addresses and locations (e.g. zip codes) could be managed in alternative manner if the investigators know what elements are related to that information and then have the identifiers recoded with the corresponding element values, thus preserving the ability to make associations. An example would be the socio-economic status of a patient's residence. The address would be de-identified in the data and the socio-economic status value has been linked to the dataset. Other values addressing level of education, insurance status, occupation would all be amenable to pre-hoc recoding.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

see my comments above for potentially identifiable information. The investigators should be able to specify the surrogate values for that information that would be relevant for analysis and these would be encoded at the time of de-identification.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

The challenge with data linkage is being able to combine multiple data sources (and images) for a given individual and also be able to deal with the possibility that a single individual may be represented in multiple datasets that have been created for alternative studies (and in some cases, related studies). One needs to know if the same person is contributing to multiple analyses and also to ensure that there is high fidelity with the data linkage. On a personal note, many years ago, I sent a request to a large number of specialists in my field looking for cases of a specific toxic retinopathy. A number of them responded with cases, making it appear that the condition was fairly prevalent. When we obtained the information about those cases, we discovered that most of them were of the same individual who had been presented at several conferences and the images had been distributed to the attendees. Though the patient's identity remained anonymous throughout the entire process, the fundus images allowed us to readily identify that they are from a common source.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

The challenge in getting meaningful consent is that one can't always predict which data may be appropriate for linkage in the future. This can be medical records, laboratory results, dietary information, imaging studies, etc. The key to this is to not ensure which data is linked together but how the data will be used. A person should not have to provide a consent that prevents self-incrimination. Thus it should be understood that these datasets can be used by law enforcement, employers or by companies with a potential financial interest in identification of the individual (such as misuse/fraud of health insurance). The boundaries of data use need to be clearly defined so as to be informative of the potential "risk" to the participant. In other words, individuals or groups who use the data for disallowed purposes are potentially subject to litigation and the expense of such litigation should be borne by the agency that approved and funded the research (NIH). If a person becomes aware that their data (regardless of how it has been linked) for biomedical research has been used for other purposes, they should be able to file a claim with the NIH for investigation and potential prosecution of the perpetrators. This would include institutions that are responsible for "holding" the information and who either provide that information to an unauthorized user or has the information hacked and exploited.

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

The key with genomic data is to be sure that the dataset for an individual includes the platform that is used, the version and when the data was last analyzed. Assignments of the pathogenicity of variants are constantly changing and new methods can detect variants that are not identified from older techniques. So this meta data is key to quality control and should have a string attached to it, much like they use for bitcoins and NFTs

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

There should be a standardization of the metadata used to attach to genomic data (much like a DICOM standard)

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

No concerns regarding the initial data sharing timelines. The only concern is really how one handles the updating of the data analyses and attributes and how that will get incorporated into the ongoing shared datasets

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

From the standpoint of practicality, it is reasonable to exclude small scale studies from being required to comply with the GDS policy but they should be encouraged to participate if they wish to do so. This will work best if the data sharing process is simple and straightforward. I would not make this policy contingent on the type of award (that leads to a host of issues). Rather it would be if the data is considered pilot and/or exploratory. The policy should apply if NIH funds a portion of the research (e.g. funding that is enabling) even if it doesn't directly support the sequencing itself.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

One needs to cover imaging data that is also linked to genetic/genomic data and this includes AI analyses of those imaging datasets.

Email: gorin@jsei.ucla.edu

ID: 1880

Submit date: 2/17/2022

Name: Gaia Cantelli

Name of Organization: EMBL-EBI

Type of Organization: Nonprofit Research Organization

Role: Institutional official

I am responding to this RFI: On behalf of an organization

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

With regards to other omics data types, a recent white paper entitled "Data management of Sensitive Human Proteomics Data: Current Practises, Recommendations, and Perspectives for the Future" (PMID: 33711481) was published last year, authored by the managers of the main proteomics data repositories in Europe and USA (part of the ProteomeXchange Consortium, including EBI's PRIDE database, the world-leading resource), and by the leaders of proteomics activities in ELIXIR, the European infrastructure for life sciences. In this white paper, the current state-of-the-art with regards to the potential identifiability of proteomics data was summarised. Identifiability can potentially arise from different proteomics data types, including protein sequence variation information (e.g. Single Amino Acid Variants) and/or protein expression levels, among others. The proposed way forward to handle ethics issues could be summarised with the principle "as open as possible, as closed as necessary". This way, two main goals would be achieved: ensure that open policies in proteomics are not hindered (and the corresponding benefits for the field remain) while, at the same time, ensuring that potential risks for individuals are appropriately considered and managed where required. Several concrete recommendations were included in the white paper, e.g.: 1) Research is needed. Tailored studies must be performed to learn more about identifiability risks for the different proteomics data types and approaches, in order to be able to take better informed policy decisions in the future. 2) Policy making. Specialists on policy about biological data need to understand the different data types included in proteomics experiments and the inherent differences with DNA/RNA sequencing data. Members of the Data Access Committees/IRBs must also have adequate training with the same overall objective in mind. In any case, in our view data access restrictions should not be higher than the existing ones for

transcriptomics data. 3) Bioinformatics Infrastructure. There must be infrastructure (data repositories, submission pipelines, interfaces to access the data, potentially tailored data formats as well) that can support appropriately controlled-access proteomics data. Funders must realise that this whole ecosystem will be required. Additionally, a commentary entitled "The growing need for controlled data access models in clinical proteomics and metabolomics" (PMID: 34599180) was also published last year in Nature Communications, authored by the managers of the EMBL-EBI resources EGA (European Genotype-Phenome Archive), MetaboLights (for metabolomics data) and PRIDE (for proteomics data). This article highlighted the need for controlled accessed bioinformatics infrastructure for other omics data types, focusing on proteomics and metabolomics approaches. There, it is stated that "existing controlled access resources such as EGA, dbGaP and JGA are not ideal for proteomics and metabolomics datasets. The data model of these resources is based on the Sequence Read Archive data model, which is tailored for sequencing-based assays and cannot appropriately represent proteomics and metabolomics datasets". It is also stated there that, in addition to identifiability, "controlled access to proteomics and metabolomics data may become necessary because of requirements related to patient consent or due to personal data regulations like GDPR (General Data Protection Regulations) or any other relevant legislation".

ID: 1881

Submit date: 2/22/2022

Name: Barbara E. Bierer, MD

Name of Organization: MRCT Center

Type of Organization: Other

Type of Organization – Other: Academic Medical Center

Role: Scientific researcher

I am responding to this RFI: On behalf of an organization

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/bbXIzNqADO.pdf

Description: MRCT Center response to NOT-OD-22-029

Email: bbierer@bwh.harvard.edu

ID: 1882

Submit date: 2/24/2022

I am responding to this RFI: On behalf of myself

Name: Lynn Brielmaier

Type of Organization: Patient Advocacy Organization

Role: Member of the Public

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

I am against this. It is well established that by aggregation of this data with other databases would result in the ability to identify people personally. Then this could be employed for workplace discrimination in hiring and advancement, insurance acceptance and rate hikes. It will also be used against people looking to serve in public office.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

I am against this. It is well established that by aggregation of this data with other databases would result in the ability to identify people personally. Then this could be employed for workplace discrimination in hiring and advancement, insurance acceptance and rate hikes. It will also be used against people looking to serve in public office.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

I am against this. It is well established that by aggregation of this data with other databases would result in the ability to identify people personally. Then this could be employed for workplace discrimination in hiring and advancement, insurance acceptance and rate hikes. It will also be used against people looking to serve in public office.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Most people don't understand the implications of consent, and will end up being discriminated against without their knowledge.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

Most individuals will not dig into these subtopics, and do not understand the implications of consent.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

policies should be harmonized, but there is danger that States could subvert the rules.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

the more databases involved, the easier it will be to suffer a CyberSecurity breach, and leak the data. It is well established that any database is subject to Nation State attack. It is easy to envision the same technologies will trickle down to criminal and commercial exploit. Political exploit is also available.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

the types should be considered and harmonized.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

the types should be considered and harmonized.

ID: 1884

Submit date: 2/24/2022

I am responding to this RFI: On behalf of an organization

Name: Anurupa Dev

Name of Organization: Association of American Medical Colleges

Type of Organization: Nonprofit Research Organization

Role: Member of the Public

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Protecting the interests of research participants should continue to be a central principle of the NIH GDS Policy and will require the agency to carefully examine its existing criteria around de-identification and policies for informed consent. The AAMC appreciates the NIH's acknowledgement that "the concept of 'identifiability' is a matter of ongoing deliberation within the scientific and bioethics communities." Particularly when working with certain human genomes, such as those from a rare disease or underrepresented population, it can become significantly easier to link genomic data to an individual. Furthermore, as our understanding of genomics and analytic capabilities constantly improve, data that is marked as de-identified now would possibly be re-identifiable in the future. We have concerns about the requirements in the current GDS Policy to de-identify human genomic data by applying standards from both the HHS Regulations for Protection of Human Subjects (The Common Rule) [45 CFR 46.102(e)] and the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [45 CFR 164.514(b)(2)]. This unclear combination of two completely separate regulations is a suboptimal approach to de-identification. The removal of the 18 identifiers in the HIPAA privacy rule do not ensure de-identification, and as noted in the RFI may actually impede certain types of research due to the removal of data elements, such as granular location data, which may be needed for scientific inquiries. In future updates, we recommend that the GDS Policy adhere to the standards set forth in the Common Rule to ensure that the identities of research subjects cannot be readily ascertained by the investigator or associated with the data. We note that the revised Common Rule includes a mechanism for periodically revisiting the concept of identifiability in light of currently existing technologies and urge that the GDS policy follow those standards as they develop. In 2016 comments in response to proposed rulemaking for the Common Rule , the AAMC recommended against including specific safeguards for identifiability in regulation and instead suggested the use of "examples of reasonable safeguards presented within a tiered risk-based framework" as guidance. Both the scientific community and research participants would be served by the GDS Policy not requiring certain actions for data de-identification but rather focusing on developing optimal and study-specific methods for minimizing identifiability and communicating the expectation to research participants that users of data will not seek to identify individuals, but the potential for re-identification may exist.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Another element that will likely increase patient privacy concerns is the linkage of genomic sequences with data that do not necessarily meet the GDS Policy, like disease outcomes collected in clinical settings. With the rise of personalized medicine, being able to link genomic data to phenotypes, social data, and other information, can greatly increase the value of data and our ability to look for and develop diagnostics, prognostics, therapeutics, and risk assessment—however, these additional capabilities also increase the amount of information linked to each individual research participant.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Permitting data linkage also emphasizes the critical role of the consent and widespread educational processes to adequately inform research participants that their genomic and phenotypic data may be used for future research purposes and shared broadly, and what the associated risks might be. The continual emergence of new analytics and techniques may present research opportunities that cannot be specifically noted in a consent form. Obtaining sufficiently expansive consent is also critical given that locating a research participant months or years after the initial research has concluded is very difficult and itself requires re-identification. AAMC recommends that the NIH develop sample language and/or consent form templates to clearly explain the rationale for genomic research and data linkage, as well as any associated privacy and confidentiality risks. Ideally, this guidance would assist individual Institutional Review Boards (IRBs) in understanding what should be required of investigators and communicated to research participants. Importantly, consequences of these risks should not fall solely on the patient—as NIH moves into a new era of increased genomic data sharing and data linkage, with potential changes in risks to research participants, we believe it is important for the agency to formulate and implement appropriate enforcement actions for misuse or inappropriate sharing or transfer of data, including monetary penalties or suspension of current or future funding. Finally, effective data linkage will require oversight and standards to ensure that data are robust and contain the appropriate metadata. Drawing connections between linked data may require not only processed data, but raw data and the appropriate metadata regarding software and scripts to be able to conduct or corroborate an analysis. While preparing high quality data is expensive and time-consuming, it is also necessary for data integration and meta-analysis.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

It is clear that the amount of large-scale genomic data being generated will only continue to increase and will accordingly require an expansion of platforms and repositories for data sharing, storage, and analysis. We agree that it is critical for all these NIH-supported resources to maintain appropriate standards and protections for data, and to adhere to existing principles for data access and security. Despite the fact that some of these databases exist outside of the NIH, it is key to maintain controlled access models where necessary, have systems for user authentication, and procedures for managing inappropriate or unauthorized use or access. Although the need to develop and use new data platforms is clear, it could also result in or exacerbate inequities across the research enterprise. Institutions with more limited resources may not be set up to support investigators who want to share or access data. Typically, institutions require strong existing infrastructure and data scientists to support investigations in the genomic sciences. As NIH invests in awardee institutions to develop resources, it should consider the differences in institutions and their capabilities and ensure that the ability to conduct scientific investigation or the populations able to participate in research are not limited by disparate resources.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

In its comments to NIH on the DMS Policy, AAMC emphasized that "It is critical to have as much as harmonization and standardization as possible across the NIH in both the policy requirements and implementation. This includes all grantees as well as consistency in evaluation of compliance and in institute-specific requirements." The AAMC appreciates NIH's intent to harmonize the GDS Policy and GDS Plan elements, submission, and review with the DMS Policy and strongly encourages this harmonization. We support inclusion of genomic data sharing within the DMS Plan, as outlined in the RFI, to avoid the duplicative submission of plans by the researcher. We additionally agree that these plans, including the budget for genomic data management and sharing, should be assessed during the grant review process.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

While the DMS Policy does set a more flexible timeline for data sharing, any decision to change GDS Policy data submission timelines should involve careful consideration of potential impacts on policy compliance, data use and re-use, and what is most effective for the genomic research community.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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**

("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).

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.**

The ongoing expansion of studies in "omics" to include not just genomics, but also transcriptomics, proteomics, microbiomics, or metabolomics, raise new questions on the ideal scope of the GDS Policy. It is clear that, like genomic sequences, these data can be a valuable source of information especially when linked with other phenotypic data. The AAMC recommends that questions on whether and how these data types are integrated into the GDS Policy should involve further targeted outreach to experts, particularly as standards in these fields are still under development. We also note the potential benefits of including smaller-scale studies under the GDS Policy. Advancements in artificial intelligence and machine learning make it easier to aggregate and analyze data from disparate sources. Additionally, meta-analysis across multiple smaller data sets can assist researchers in obtaining a broader population sample and including more underrepresented individuals in research studies.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

While the DMS Policy does set a more flexible timeline for data sharing, any decision to change GDS Policy data submission timelines should involve careful consideration of potential impacts on policy compliance, data use and re-use, and what is most effective for the genomic research community.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/nRUdYNYFbA.pdf

Email: adev@aamc.org

ID: 1885

Submit date: 2/25/2022

I am responding to this RFI: On behalf of an organization

Name: Anonymous

Name of Organization: Anonymous

Type of Organization: Other

Type of Organization-Other: Genomics lab

Role: Medical provider

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/YZssLwNRUF.pdf

Email: madeline.gitomer@hoganlovells.com

ID: 1886

Submit date: 2/25/2022

I am responding to this RFI: On behalf of myself

Type of Organization: University

Role: Scientific researcher

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Current regulations about the sharing of potentially identifiable information are overly onerous and substantially impede scientific progress. As a medical practitioner, when I consent patients to provide tissue for scientific research, they have little concern about its potential identifiability. Conversely, they very much want their tissue to be used to advance new therapeutic and preventative measures for disease (in this case, cancer). However, as a scientist, I find that I am unable to access important datasets due to regulations around privacy concerns and their interpretation by various institutions--although I regularly analyze identifiable data in secure environments. For example, GENIE data are minimally analyzable without access to raw data, which have been withheld due to privacy concerns. Similarly, essentially all genomic data from Europe are no longer accessible to me or many other US-based analysts. These issues are often most problematic with non-NIH-funded data. For example, the European issue is due to European and not NIH regulations. But the scope of data that are then inaccessible to NIH-funded researchers seems large enough for the NIH to make every effort to make those data accessible. For example, GENIE data obtained in clinical settings might not fall under GDS policy, but the institutions obtaining those data are heavily NIH-funded, and the assays used resulted directly from NIH-funded advances. The fact that data generated by NIH funds can be used in Europe but European data cannot be used by NIH-funded researchers seems counter to our national interest.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

b. Yes, results of any size study should be shared upon publication or by the award end date (the earliest of the two).

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

If data defining genomic/epigenomic/transcriptomic are produced by an NIH-supported study, these derived and processed data should be submitted into an appropriate public repository (e.g., dbGaP, GEO, SRA) along with raw data for that study (Level 1/2 as appropriate). This shall include, but not be limited to the following: genetic variation (SNV, INDEL, SV, CNV, LOH), transcript/gene abundance, differentially methylated regions or inferred fraction of DNA methylation, accessible chromatin regions, regions enriched for a particular histone modification or transcription factor, topologically associating domains (TADs) or other chromatin interactions, and regions associated with enhancer activity. Such data shall be provided in standard formats, e.g., VCF/MAF, bed/bedGraph/bigwig, or tab-delimited text such that they are viewable on commonly used genome browser software or importable to common statistical software environments and shall be appropriately normalized to control for variation in sequencing depth. Further, as a means of assessing data reproducibility and integrity of derived results, read coverage shall be provided in addition to any data vectors described above on a per-sample basis. Derived and processed data describing genomic changes are used by the Environmental Health Sciences community as well as by a broad community of genetic, genomic, and clinical investigators. The current GDS policy indicates the requirement of sharing derived and processed genomic data (Level 3 Data) only in general and describes the types of Level 3 Data in Supplement (Table 1) without sufficient detail. The common practice for most current studies is to submit only raw data, even if genomic/epigenomic/transcriptomic features were analytically determined and used to support or produce conclusions in reports and/or publications. Sometimes, the derived data are included in the publication as partial tables or supplementary materials in computationally incompatible formats greatly reducing data findability, reusability, and computability. Listing main types of derived and processed genomic data (Level 3) in the main part of the Policy document and providing more detailed list in the Supplementary sections would highlight this requirement to researchers at the beginning of the project, rather than pointing to a gap in data sharing post factum. Repeating analyses to independently derive data describing genomic change may consume significant computational resources and may diminish experimental reproducibility. In addition, failure to share derived data precludes researchers that lack computational resources from fully using genomic data.

ID: 1888

Submit date: 2/25/2022

I am responding to this RFI: On behalf of an organization

Name: Elizabeth Sun

Name of Organization: International Society for Biological and Environmental Repositories (ISBER)

Type of Organization: Professional Org/Association

Role: Member of the Public

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

See attachment.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

See attachment.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

See attachment.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

See attachment.

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

See attachment.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

See attachment.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

See attachment.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.

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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).

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.

See attachment.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

See attachment.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/KEJoNPwEys.pdf

Email: elizabeth.sun@malachite-mgmt.com

ID: 1890

Submit date: 2/25/2022

I am responding to this RFI: On behalf of myself

Type of Organization: Nonprofit Research Organization

Role: Scientific researcher

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

If participants have the ability to opt-in with the applicable consent form(s) and researchers are required to obtain IRB approval for work with such data, then we would be supportive of this. Most rare disease parents/patients are supportive of research but it would be best if clarified up front and they are allowed to opt-in (or out) of such sharing.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Same concept as above in #2. If participants have the ability to opt-in to link data with the applicable consent form(s) and researchers are required to obtain IRB approval for work with such data, then we would be supportive of this. Most rare disease parents/patients are supportive of research but it would be best if clarified up front and they are allowed to opt-in (or out) of such data linkage.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Data linkage should be addressed when obtaining consent to the researchers' best knowledge. We have kept our consent form broad to allow sharing to anyone studying the specific rare disease but did not specify depositing in larger repositories such as dbGap etc. We strongly support that individuals participating in research have the choice to elect to share their data at the time of consent but we also understand that data collected prior to institution of this policy may need to be exempt.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

Genomic data may not meet the requirement to be sufficient to replicate findings especially if completed under "research" category vs. in a clinical setting, therefore it could be exempt from falling under definition of “scientific data” (and thus, no requirement for sharing). Most patients are willing and open to sharing data and willing to accept some risks of de-identification to study their rare disease and typically want to make sure it is used but we do need to address/consider the limits on researchers if all genomic data falls under this policy.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

We feel strongly that timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first) to eliminate unnecessary burden on research teams. Three months goes by far too quickly, especially in a laboratory setting to require such a quick turnaround.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 (“Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards,” NOT-OD-14-111).**
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.**

In regards to (b), it is currently unclear how much data is available to ascertain how potentially identifiable such other data types could be. It would be difficult to apply the policy without some knowledge on this. We agree that 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.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

By requiring that NIH-funded researchers deposit data into a larger repository for their research to be funded, this policy removes the participant’s ability to participate in research in a meaningful way while maintaining control of how their data is used. They should be able to elect to share such data as they wish to the best of our ability (perhaps special exemptions for historical samples or data-more on this

below). Another concern is the barrier this policy may place on collecting data from a diverse group of participants. The additional requirements to deposit data may act as a barrier to certain groups of people to participate in research. There are already barriers that include access to study sites or inherent fear/negative perceptions of research within some communities; we should actively work to avoid generating additional barriers. If data has been historically collected prior to the establishment of these requirements and is completely de-identified (e.g., not able to re-identify), researchers should be allowed to utilize such samples for analyses. Otherwise, the research community loses access to study a substantial number of available samples and/or data currently sitting at various academic institutions unused. While we openly support and advocate for open science, we also feel strongly that depositing genetic data in such cases where it is not possible to obtain dbGap consent or other appropriate consent(s) should not be required. This policy could be adapted for all future prospective data and/or sample collections but some leniency is needed within this policy for previously stored data/samples. Specifically, data and biological samples should be permitted for use in NIH-funded studies without the requirement for data depositing when obtaining consent is not a possibility. Perhaps a specific committee can be established to ascertain the validity of such use cases?

ID: 1891

Submit date: 2/25/2022

I am responding to this RFI: On behalf of an organization

Name: Mary Beth Terry

Name of Organization: Breast Cancer Family Registry

Type of Organization: Other

Type of Organization-Other: Research cohort consortium

Role: Scientific researcher

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/CLWjIIMIvy.pdf

Email: mt146@cumc.columbia.edu

ID: 1892

Submit date: 2/26/2022

I am responding to this RFI: On behalf of an organization

Name: Angela Page

Name of Organization: Global Alliance for Genomics and Health

Type of Organization: Other

Type of Organization-Other: Non-profit Standards Development Organization

Role: Institutional official

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

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.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

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.

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

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).

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

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.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

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.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

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.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS 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.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/utklWGPcTq.pdf

Description: PDF document of full response

Email: angela.page@ga4gh.org

ID: 1896

Submit date: 2/27/2022

I am responding to this RFI: On behalf of myself

Name: Steven A. Roberts

Name of Organization: Washington State University

Type of Organization: University

Role: Scientific researcher

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

Currently NIH-supported studies that generate genomic data are only required to deposit raw sequencing reads to an appropriate public repository. However, the utilization of these data by other researchers, either for investigating new questions or repeating the originally published analysis is often limited by the ability of groups to conduct computationally intensive calling of derived data. My group's research has personally been slowed by the lack of depositing calls for single nucleotide variants and normalized RNA-seq expression data. I support the recommendations provided by the NIEHS on the updated GDS policy that "If data defining genomic/epigenomic/transcriptomic are produced by an NIH-supported study, these derived and processed data should be submitted into an appropriate public repository (e.g., dbGaP, GEO, SRA) along with raw data for that study (Level 1/2 as appropriate). This shall include, but not be limited to the following: genetic variation (SNV, INDEL, SV, CNV, LOH), transcript/gene abundance, differentially methylated regions or inferred fraction of DNA methylation, accessible chromatin regions, regions enriched for a particular histone modification or transcription factor, topologically associating domains (TADs) or other chromatin interactions, and regions associated with enhancer activity. Such data shall be provided in standard formats, e.g., VCF/MAF, bed/bedGraph/bigwig, or tab-delimited text such that they are viewable on commonly used genome browser software or importable to common statistical software environments and shall be appropriately normalized to control for variation in sequencing depth. Further, as a means of assessing data reproducibility and integrity of derived results, read coverage shall be provided in addition to any data vectors described above on a per-sample basis. Justification: Derived and processed data describing genomic changes are used by the Environmental Health Sciences community as well as by a broad community of genetic, genomic, and clinical investigators. The current GDS policy indicates the requirement of sharing derived and processed genomic data (Level 3 Data) only in general and describes the types of Level 3 Data in Supplement (Table 1) without sufficient detail. The common practice for most current studies is to submit only raw data, even if genomic/epigenomic/transcriptomic features were analytically determined and used to support or produce conclusions in reports and/or publications. Sometimes, the derived data are included in the publication as partial tables or supplementary materials in computationally incompatible formats greatly reducing data findability, reusability, and computability. Listing main types of derived and processed genomic data (Level 3) in the main part of the Policy document and providing more detailed list in the Supplementary sections would highlight this

requirement to researchers at the beginning of the project, rather than pointing to a gap in data sharing post factum. Repeating analyses to independently derive data describing genomic change may consume significant computational resources and may diminish experimental reproducibility. In addition, failure to share derived data precludes researchers that lack computational resources from fully using genomic data."

Email: steven.roberts2@wsu.edu

ID: 1897

Submit date: 2/27/2022

I am responding to this RFI: On behalf of an organization

Name: Leigh Burchell

Name of Organization: Allscripts

Type of Organization: Other

Type of Organization-Other: Technology developer

Role: Member of the Public

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Allscripts supports the use of expert determination as an acceptable strategy for de-identification and currently relies on this approach. The likelihood of patient re-identification is largely dependent on the anticipated recipient (with this also a focus of HIPAA), and an anticipated recipient with access to large genomic data sources (e.g., 23andMe) could re-identify a patient with some effort without the need for a geneticist. It is unlikely that most organizations would have access to such large genomic data sources that could allow them to be able to re-identify. Allscripts believes it is necessary to harmonize the definition of "de-identification", considering definitions and guidance from HIPAA and Safe Harbor rules, to improve the software development community's ability to respond to these issues in a consistent way. We also recommend that the NIH work with the Department of Health and Human Services Office for Civil Rights (OCR) to provide greater clarity regarding the types of genomic data or scenarios in which genomic data would qualify as a biometric identifier, setting clearer expectations for entities engaging in research activities that use genomic data.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Allscripts requests clarification regarding the purpose of this proposed change since adding potentially identifiable data to a HIPAA de-identified data set would potentially render it as PHI. We do want to be able to store and share identifiable information where appropriate, but what, exactly, is the proposal here? Is the NIH planning to create a repository for public consumption? Would such a repository fall under the HIPAA framework for limited data sets, if so, how would this work?

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with

consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

There are many different concepts being combined here so responses from industry will need parsing; we would recommend separating them out to avoid confusion. A narrow reading of this could lead to a policy that indicates certain linkages are not allowed even where it is possible to craft a linked data product that includes elements of both underlying datasets that can meet the GDS Policy requirements even where a full linkage would not. Allscripts would support a GDS Policy permitting data linkage under these criteria but would not support a framework that would require different safeguards than are already in place. It is our recommendation that there be no language prohibiting linkages, because our experience tells us that the combined data product may pose a potential problem more so than the act of linking itself. Rather, it is our recommendation to differentiate the contents of the initial linked database as separate from the linked data product itself and then to carry through with Expert Determination for the linked sets. We note that Allscripts and our Veradigm data business are already used to successfully and safely linking data, and it would be problematic if there were to be an issue because of the new policy if we were to continue using partial information. Please clarify if the purpose of this proposal is to ensure that consent is carried all the way through to link to other data that was not consented. The final product needs to remain de-identified via Safe Harbor or Expert Determination. It is worth reinforcing here that it is the responsibility of the provider to obtain patient consent, so where we serve as the data resource to people conducting research, we cannot always know how that consent was obtained, nor would it be reasonable to expect us to.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

It is Allscripts' recommendation that these three issues be split into separate items to avoid confusion; for clarity, we will attempt to address them separately here. We agree that consent should address potential data linkage but would not want to see this become a blanket policy covering all scenarios specific to data linkage language in consent documents. We agree that an IRB should weigh risks of linkage, current or future, but also be willing to revise policies as future use cases expand. The industry is still in the infancy of genomic data collection and sharing, so we caution against blanket policy approaches. Ensuring that consent is meaningful is a separate issue that is much larger than just this request for comment. We recommend a separate request for comment around this one topic. Allscripts does not believe there is any reason to treat genetic data differently than other protected health data. Although there may is currently clearly elevated concern about genetic data, we believe that in practicality, it should be treated consistently with other sensitive health information. Genetics is not addressed by HIPAA at all, and to this point, no one has identified any part of the genome as protected. Conversely, treating all sensitive data consistently would certainly ease the work of technical resources, including the software development community, and could do a great deal to address confusion among the provider community that currently has to remember different requirements and limitations associated with different types of data. It would be simplest for all involved if all true PHI is treated identically as PHI. The fact that it has already been agreed that the whole genetic sequence is PHI, but some fragments are considered to be PHI while others are not, is another reason that we should rely on Expert Determination. Further, given that what is not identifiable today may become so tomorrow as

technology advances, the need for Expert Determination is expected to remain for many years. In all cases, it is our recommendation to try to avoid building walls between or around any data in ways that are different from others. We do need guidance but would not want to create something new for this purpose. Our genetic experts believe there is not actually more risk associated with this data than other PHI, but that the industry simply needs to agree how better to deal with it in a consistent manner.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

Allscripts does not have a strong opinion about these data management and sharing principles but is eager to learn where the revised GDS Policy lands on this so we can abide by any and all changes. We recommend that the NIH provide clear parameters so industry can make informed business decisions and stay compliant.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

Based on our experience, Allscripts believes there would be a huge potential impact on time and resources required by these two endpoints, and we suggest that effort should be focused on avoiding making it burdensome for whoever has to submit the data. It seems the later endpoint would be less burdensome. We simply recommend a harmonization of polices and timelines. Allscripts supports and appreciates the approach proposed to the topic of timelines.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

It is not clear why sample size would be a measure used to identify the allowability of an exception, and the figure of 100 participants seems arbitrary. A better definition of 'small scale' would be helpful, but we remind the NIH that that the smaller the scale, the higher the possibility of re-identification.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/xYjlkPVfet.pdf

ID: 1898

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Name: Anna Malkova

Name of Organization: University of Iowa

Type of Organization: University

Role: Scientific researcher

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

If data defining genomic/epigenomic/transcriptomic are produced by an NIH-supported study, these derived and processed data should be submitted into an appropriate public repository (e.g., dbGaP, GEO, SRA) along with raw data for that study (Level 1/2 as appropriate). This shall include, but not be limited to the following: genetic variation (SNV, INDEL, SV, CNV, LOH), transcript/gene abundance, differentially methylated regions or inferred fraction of DNA methylation, accessible chromatin regions, regions enriched for a particular histone modification or transcription factor, topologically associating domains (TADs) or other chromatin interactions, and regions associated with enhancer activity. Such data shall be provided in standard formats, e.g., VCF/MAF, bed/bedGraph/bigwig, or tab-delimited text such that they are viewable on commonly used genome browser software or importable to common statistical software environments and shall be appropriately normalized to control for variation in sequencing depth. Further, as a means of assessing data reproducibility and integrity of derived results, read coverage shall be provided in addition to any data vectors described above on a per-sample basis. Justification: Derived and processed data describing genomic changes are used by the Environmental Health Sciences community as well as by a broad community of genetic, genomic, and clinical investigators. The current GDS policy indicates the requirement of sharing derived and processed genomic data (Level 3 Data) only in general and describes the types of Level 3 Data in Supplement (Table 1) without sufficient detail. The common practice for most current studies is to submit only raw data, even if genomic/epigenomic/transcriptomic features were analytically determined and used to support or produce conclusions in reports and/or publications. Sometimes, the derived data are included in the publication as partial tables or supplementary materials in computationally incompatible formats greatly reducing data findability, reusability, and computability. Listing main types of derived and processed genomic data (Level 3) in the main part of the Policy document and providing more detailed list in the Supplementary sections would highlight this requirement to researchers at the beginning of the project, rather than pointing to a gap in data sharing post factum. Repeating analyses to independently derive data describing genomic change may consume significant computational resources and may diminish experimental reproducibility. In addition, failure to share derived data precludes researchers that lack computational resources from fully using genomic data.

ID: 1899

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Name: David Kwiatkowski

Name of Organization: Dana Farber Harvard Cancer Cener

Type of Organization: Nonprofit Research Organization

Role: Scientific researcher

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

Suggestion: If data defining genomic/epigenomic/transcriptomic are produced by an NIH-supported study, these derived and processed data should be submitted into an appropriate public repository (e.g., dbGaP, GEO, SRA) along with raw data for that study (Level 1/2 as appropriate). This shall include, but not be limited to the following: genetic variation (SNV, INDEL, SV, CNV, LOH), transcript/gene abundance, differentially methylated regions or inferred fraction of DNA methylation, accessible chromatin regions, regions enriched for a particular histone modification or transcription factor, topologically associating domains (TADs) or other chromatin interactions, and regions associated with enhancer activity. Such data shall be provided in standard formats, e.g., VCF/MAF, bed/bedGraph/bigwig, or tab-delimited text such that they are viewable on commonly used genome browser software or importable to common statistical software environments and shall be appropriately normalized to control for variation in sequencing depth. Further, as a means of assessing data reproducibility and integrity of derived results, read coverage shall be provided in addition to any data vectors described above on a per-sample basis. Justification: Derived and processed data describing genomic changes are used by the Environmental Health Sciences community as well as by a broad community of genetic, genomic, and clinical investigators. The current GDS policy indicates the requirement of sharing derived and processed genomic data (Level 3 Data) only in general and describes the types of Level 3 Data in Supplement (Table 1) without sufficient detail. The common practice for most current studies is to submit only raw data, even if genomic/epigenomic/transcriptomic features were analytically determined and used to support or produce conclusions in reports and/or publications. Sometimes, the derived data are included in the publication as partial tables or supplementary materials in computationally incompatible formats greatly reducing data findability, reusability, and computability. Listing main types of derived and processed genomic data (Level 3) in the main part of the Policy document and providing more detailed list in the Supplementary sections would highlight this requirement to researchers at the beginning of the project, rather than pointing to a gap in data sharing post factum. Repeating analyses to independently derive data describing genomic change may consume significant computational resources and may diminish experimental reproducibility. In addition, failure to share derived data precludes researchers that lack computational resources from fully using genomic data.

ID: 1901

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: AIMEE GOLBITZ

Name of Organization: Mass General Brigham

Type of Organization: Health Care Delivery Organization

Role: Institutional official

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/FzHOkTLAhl.pdf

Description: Mass General Brigham GDS comments

ID: 1902

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Type of Organization: University

Role: Bioethicist

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/PrnQYWeTFq.pdf

Description: General comments and suggestions for GDS policy

ID: 1903

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Name: Natalie Saini

Name of Organization: Medical University of South Carolina

Type of Organization: University

Role: Scientific researcher

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

If data defining genomic/epigenomic/transcriptomic are produced by an NIH-supported study, these derived and processed data should be submitted into an appropriate public repository (e.g., dbGaP, GEO, SRA) along with raw data for that study (Level 1/2 as appropriate). This shall include, but not be limited to the following: genetic variation (SNV, INDEL, SV, CNV, LOH), transcript/gene abundance, differentially methylated regions or inferred fraction of DNA methylation, accessible chromatin regions, regions enriched for a particular histone modification or transcription factor, topologically associating domains (TADs) or other chromatin interactions, and regions associated with enhancer activity. Such data shall be provided in standard formats, e.g., VCF/MAF, bed/bedGraph/bigwig, or tab-delimited text such that they are viewable on commonly used genome browser software or importable to common statistical software environments and shall be appropriately normalized to control for variation in sequencing depth. Further, as a means of assessing data reproducibility and integrity of derived results, read coverage shall be provided in addition to any data vectors described above on a per-sample basis. Justification: Derived and processed data describing genomic changes are used by the Environmental Health Sciences community as well as by a broad community of genetic, genomic, and clinical investigators. The current GDS policy indicates the requirement of sharing derived and processed genomic data (Level 3 Data) only in general and describes the types of Level 3 Data in Supplement (Table 1) without sufficient detail. The common practice for most current studies is to submit only raw data, even if genomic/epigenomic/transcriptomic features were analytically determined and used to support or produce conclusions in reports and/or publications. Sometimes, the derived data are included in the publication as partial tables or supplementary materials in computationally incompatible formats greatly reducing data findability, reusability, and computability. Listing main types of derived and processed genomic data (Level 3) in the main part of the Policy document and providing more detailed list in the Supplementary sections would highlight this requirement to researchers at the beginning of the project, rather than pointing to a gap in data sharing post factum. Repeating analyses to independently derive data describing genomic change may consume significant computational resources and may diminish experimental reproducibility. In addition, failure to share derived data precludes researchers that lack computational resources from fully using genomic data.

Email: sainina@musc.edu

ID: 1904

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Doug Fridsma

Name of Organization: Datavant Group

Type of Organization: Other

Type of Organization-Other: Health IT company

Role: Institutional official

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

The challenge with genomic information is that full redaction of genomic information–either by removing information or abstracting it to less specific information–can often render the data useless for analysis. While genomic information is highly specific and often unique to an individual, in isolation, the risk of re-identification of these highly specific and often unique sequences is low. Care must be taken when pooling and linking data that contains genetic information. Given the unique characteristics of genomic data, and the complexity and rapidly evolving nature of genomic data sharing, expert determination may be the only option for de-identification of genomic information. The Omnibus Rule in 2013 did not include genetic information as one of the 18 direct identifiers as defined by HIPAA, but redaction of the 18 identifiers as part of Safe Harbor is not sufficient to classify datasets containing genetic information as de-identified. The application of expert determination as a method for assessing the disclosure risk of genomic data has the benefit of being tailored to individual datasets, with a bespoke handling of the trade-offs between utility and privacy. To be effective, expert determination requires robust quantitative estimates of the risk contained within a dataset. Such robust assessment of risk is based on the distributions of potentially identifying values within the data and the intersectionality of those distributions as compared to reference data. To support expert determination, the NIH should support additional research on mutant allele frequencies, non-coding DNA mutations, sequence variability in certain regions of the genome, frequency of silent mutations, chromosome phenotypes, single SNP and combinations of polymorphisms and other genetic and genomic characteristics. Without the underpinning of these baseline assessments, expert determination will be either too conservative or too permissive in the evaluation of disclosure risk. Expert determination provides the only nuanced approach to managing genomic information, and can take into consideration key attributes of the genomic data to be shared: whether the sequence is from a tumor or somatic cell line with tumor sequences at a lower risk for re-identification than inheritable or somatic cell lines the frequency of specific mutations with rare or low frequency mutation at a higher risk for re-identification the length of the sequence with shorter sequences at a lower risk for re-identification the comprehensiveness of the dataset with more comprehensive information about an individual at higher risk for re-identification. Although HIPAA has determined that heritability does not limit the ability of an

individual to share their data, with more genomic data being available for analysis, it will be important to continue to monitor the re-identification risk and potential harm to groups and families. This and the factors listed above should be considered when determining the risks for re-identification in sharing genomic information. For example, expert determination may be the only way to assess the risk of re-identification of somatic cell mutations (such as the BRCA-1 or CFTR genes). Common sequences such as the CFTR FΔ508 may not create a risk of re-identification (putting a cystic fibrosis patient in a group of about 30,000), while n-terminal missense mutations such as c.14C>T are very rare, and potentially identifiable. A more nuanced approach to redaction and de-identification is only possible with expert determination. In addition, we should not assume that two de-identified data sets, when combined, will remain de-identified. Expert determination should be used to evaluate the risk of linked and combined datasets, allowing for more thoughtful approaches to protecting patient information. Remediation could include abstraction (substituting "the existence of a mutation" for the actual mutation), redaction (removing the sequence) or requiring additional security and access controls (for example, not allowing the data to be downloaded or removed from a data enclave). This would create more flexible ways of managing and sharing genomic information for research purposes, while always considering the risks of re-identification. Finally, every effort should be taken to use new approaches to linking and preserving patient privacy. New technologies such as privacy preserving linkages (PPRL) and honest broker governance structures are already being used by the NIH to support research into Covid-19. These approaches provide a way to link patient data without exposing PII, and do so within a governance framework that provides a trusted intermediary that can manage the data and prevent potentially identifiable data from being shared. These approaches could also be used for genomic information that when combined with expert determination of the original and linked datasets, can assure that the re-identification risk is low.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Genomic data can and should be shared within repositories so that other researchers can build on that research. As described above, expert determination (with assessment of mutation frequency, tumor vs somatic cells, sequence length, etc) can determine the risk of data submitted to a repository. Data deemed potentially identifiable should be treated as full PII data, and maintained in repositories and enclaves that control the use of that data. Restrictions on downloads and working within highly secure enclaves can reduce the chance that data would be removed from the repository and potentially re-identified. Even data that is deemed low risk should be treated cautiously and every effort should be made to reduce the likelihood that genomic data is linked to other datasets that make re-identification possible. It is possible to both link data into enclaves (without allowing identifiable data to be removed from the enclave) and PPRL tokens (or other privacy-preserving linkage technology) can be used to link data without requiring identifiable data to be shared. Our experience with the National Covid Cohort Collaborative (N3C) suggests that using honest broker intermediaries can support linking data in ways that protect patient privacy. For example, tokens (without identifiable information) may be allowed to leave a repository enclave, and if linkable data is found, that data can be moved into the enclave to offer higher security and access controls. Honest brokers, coupled with enhanced IRB review, and appropriate security and access control processes, can minimize the risk of re-identification, while providing more

value to the research community. Finally, for investigators that are using genomic data, data sharing plans submitted with grant applications should make these plans a scorable element. This would increase the attention paid to these data sharing plans, it would incentivize novel ways to share data, and would make the importance of safely sharing genomic information front and center for a research investigator. Over time, researchers will develop new and better ways to share and link data, and have the success of those approaches evaluated through peer review. The NIH will send an important message to the research community, and will only invest in research that makes data protection and data sharing a priority.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

We have divided our comments into three sections: a) Data collected, submitted to repositories (or linked) in which all the data has been consented (clinical trials or IRB-reviewed, consented research) b) Data that is collected without consent (data in electronic health records and claims data), and c) data in which some of the data has been consented, but other linked data has not (for example, linking clinical trials genomic information with real world data found in EHRs) Data collected with consent For patients who are participating in research and can consent to the use of that data, permission to share the data within protected enclaves and to link that data to other data that the patient has consented to for use (from a registry or other IRB-reviewed research), should be permitted. It is still important that the data be protected as PII, and used within access controlled repositories or de-identified before it is used for other purposes. With meaningful informed consent, patients can be fully aware of the risks and benefits of consenting to sharing data within repositories and linking that data to data that a patient may have. Data collected in which some (or all) of the data is collected without consent Data collected without consent is pervasive within EHRs, claims and consumer data. As genetic testing becomes more wide-spread, genomic data will be present in these other data sources, for which no consent for sharing or linking has been provided. In this setting in which data has been collected without consent, before this data is shared or linked, it should be de-identified and reviewed so that the risk of re-identification remains low. There are examples in which this kind of data can be useful for clinical research, and support research that can (and should) provide patients with informed consent. For example, it should be possible to use privacy preserving record linkage methods to identify patients who have data in two different data sets (without exposing PII) or identifying cohorts in datasets. If these datasets are linked, then expert determination should be used to assure that the new datasets are still appropriately de-identified. In linking to data that has been collected without consent, it is important to not expose an individual's PII as part of the linking process. Linkages should use privacy preserving mechanisms to link data, and the resulting linked data set should be evaluated to ensure that the risk of re-identification of the resulting dataset remains negligible. Data links in which some data is consented and some data is not An increasingly common scenario is when a patient consents to participate in an IRB-reviewed clinical study, and consents to the use and linkage of that data. In this case, the clinical study data has been properly consented, but it is combined and linked with other data sources that may not have been consented for use. This could include individual level claims or EHR data, or aggregate data related to

social determinants of health. In either of these cases, linkages between consented data and non-consented data can increase the risk of re-identification. This scenario is similar to the scenario above, and should be treated as if all of the data has not been consented. The non-consented data should be de-identified and no PII shared for either linking or for analysis. PPRL methods can still permit privacy preserving linkages. The resulting data set should be reviewed to assure that the new, linked dataset still conforms to the uses and restrictions of the original consented data. For example, if that data is to be held in access restricted enclaves, the new data set should be held in that same environment. If the consented data has been de-identified before it was shared, then the resulting dataset should be reviewed to be sure that it remains de-identified, and if not, remediation should be applied to the data set to assure it falls under the consented uses. The NIH should make sure that meaningful informed consent obtained for both sharing and linking considers the scenarios in which that data may be used, and follows best practices to protect the privacy and confidentiality of the patient's data.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

As described above, when patients participate in IRB-reviewed, consented research, future studies should consent participants for both sharing data and linking data to other datasets. In this setting, the informed consent should include scenarios for sharing and under what conditions that data will be shared (de-identified, identified within a protected enclave, restricted access with IRB-review and controls, etc). It should also include scenarios outlining how that data might be linked (no linkages allowed, de-identified links allowed, fully identifiable links allowed) and how the investigators plan to protect the data. However, data that is obtained in other settings (from EHRs, claims data, or is repurposed for secondary-uses in research), there should be no expectation to get meaningful informed consent for these kinds of data. In these circumstances, data should be held in secure enclaves to minimize risk of unauthorized access, every effort should be made to remove identifiable information, modern, privacy-preserving techniques for linking data sets should be used, and expert determination should be used to assess the risk of re-identification in the linked datasets. As suggested above, the importance of good data sharing and linking plans cannot be overstated. If the NIH believes that good data sharing and linking is essential to scientific advancement as well as protecting patients who participate in clinical studies, it is imperative that the NIH consider this a scorable element on grant applications, and that studies with inadequate plans for protecting patient data should be removed from consideration. Without a clear incentive to ensure the safety of genomic data, data sharing plans will remain an afterthought, and not carry the importance that these important data resources require.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/PmARMxHwbh.pdf

Description: Summary of Datavant comments to NIH Genomci data sharing RFI

Email: doug@datavant.com

ID: 1906

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Type of Organization: Nonprofit Research Organization

Role: Scientific researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Regarding the statement "Furthermore, data from multiple sources may not have been obtained under the same consent and de-identification expectations as the GDS Policy", it may be helpful to provide an additional question in consent forms about consent to data linkage with external data.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Regarding the phrase "there is certain potentially identifiable information that would not be acceptable to submit" more clarification is needed. Each data submission may include two components: one is without identifiable information and one with identifiable information. Users who want to access the one with identifiable information needs to submit a detailed plan on safeguarding identifiable information.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

note that if consent has not been collected, such data cannot be used.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

Consider that GD may include both raw and processed data, and a detailed description on how raw data were processed.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

Regarding “time of an associated publication”, clarification is needed. Is this a reference to a journal publication or a presentation at a professional meeting?

ID: 1907

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Deven McGraw

Name of Organization: Invitae Corporation

Type of Organization: Biotech/Pharmaceutical Company

Role: Member of the Public

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Please see attached document.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Please see attached document.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Please see attached document.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Please see attached document.

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

Please see attached document.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

Please see attached document.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

Please see attached document.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Please see attached document.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

Please see attached document.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/LBaRmcVIqv.pdf

Description: Invitae's Response to the RFI

Email: megan.boyd@invitae.com

ID: 1908

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Nichole Holm

Name of Organization: American Society of Human Genetics

Type of Organization: Nonprofit Research Organization

Role: Institutional official

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

The American Society of Human Genetics (ASHG) is a society of more than 8,000 genetics professional members, with the mission to advance human genetics and genomics in science, health and society through excellence in research, education, and advocacy. Many of our members conduct NIH-funded human genome research and are therefore subject to the Genomic Data Sharing Policy (GDSP). Broad data sharing is fundamental to the advancement of genomic sciences. At the same time, given that human genome information is personal and sensitive, it is important that data is shared in a way that preserves the privacy of research participants. ASHG welcomes the NIH exploring how the GDSP might be improved to achieve this. We believe that the requirement to remove the HIPAA identifiers from genomic data submitted to the NIH does impose limitations on the potential to draw correlations between genotype, phenotype, and environmental information. Therefore, we see the significant benefit to the GDSP allowing for flexibility in how data should be de-identified if individuals' privacy can be preserved.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

In general, ASHG believes removal of HIPAA identifiers is important to protect research participant privacy. However, the inclusion of select HIPAA identifiers may enable further scientific insight without significantly increasing the risk of re-identification. We therefore think there is value in providing alternatives to de-identification policies while recognizing the need to consider the greater risks of re-identification for individuals within particular populations. Some HIPAA identifiers, such as zip codes, age, and other dates related to an individual, could provide valuable data for examining biological and environmental correlations. For example, access to ages and dates of symptom onset, diagnoses, or treatments would be highly beneficial for research on age-related diseases/phenotypes or longitudinal studies of disease etiology. Additionally, access to environmental identifiers would allow researchers to compare potential sociological or environmental factors in association with genotypes and diagnoses to better elucidate gene-environment interactions. To protect the privacy of research participants, we urge additional caution regarding inclusion of HIPAA identifiers in datasets in certain circumstances: (1)

Where individuals are at a greater risk of re-identification due to their geographic location, individuals should not have their zip code retained in de-identified GDS databases. Geographic properties that pose a higher risk to re-identification include zip codes within lower population-dense regions or within or near tribal reservations/land. We recommend implementing a minimum population threshold for the inclusion of an individual's zip code. We also recommend exclusion of zip codes that overlap with tribal geographic jurisdictions. (2) Where the datasets include individuals with rare diseases, we advise additional caution and consideration of the re-identification risks associated with harboring rare genetic variants and unusual clinical records. (3) Where ages and dates of diagnosis/treatment are included in datasets, Date of Birth (DOB) should not be retained as this could significantly increase the risk of re-identification.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

ASHG supports broader data linkage and sharing, as data linkage has proven highly beneficial for human genetics and genomics. However, implementation of data linkage should not pose any additional risks to participant privacy or re-identification, and methods of data linkage should ensure privacy of the individual.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

ASHG agrees informed consent must include the potential for future data sharing under the GDS Policy, as well as disclosure of the risks associated with inter-repository sharing. To ensure this disclaimer is meaningful and understood by the individual, it should be clearly outlined in the consent along with the methods of protection in place to prevent re-identification.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

ASHG supports the harmonization of GDS and DMS policies, thereby alleviating the administrative burden for researchers having to comply with two separate policies. We recommend that data sharing plans be submitted in concordance with the DMS Policy requirement, requiring submission as a component of the initial funding application. Such a modification in the policy would highlight the importance of management and sharing of genomic data with all other aspects of study design.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**

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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).

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.

ASHG affirms that the sharing of non-genomic –omics data (e.g. transcriptome, proteome, microbiome, etc.) does not pose the same kind of risk of re-identification risk as the sharing of genomic information. Given the state of current science, we do not believe that they warrant the same level of protection at this time. However, we would welcome sharing of additional data types that present a more comprehensive profile of phenotypic involvement, and do not see other types of –omics data as posing additional identifiability risks. This should be revisited if the technology advances and re-identifiability becomes a risk.

Email: nholm@ashg.org

ID: 1909

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Michelle McClure

Name of Organization: American College of Medical Genetics and Genomics

Type of Organization: Professional Org/Association

Role: Institutional official

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/wzSwrZsMkn.pdf

Description: ACMG Comments

Email: mmcclure@acmg.net

ID: 1910

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Sarah Thibault-Sennett

Name of Organization: Association for Molecular Pathology

Type of Organization: Professional Org/Association

Role: Medical provider

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/oQpwZiMHQA.pdf

Description: AMP comments for RFI Proposed Updates and Long-Term Considerations for the NIH Genomic Data Sharing Policy

Email: sthibaultsennett@amp.org

ID: 1911

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Kasey Nicholoff

Name of Organization: Electronic Health Record Association

Type of Organization: Professional Org/Association

Role: Member of the Public

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Permit Expert De-Identification The EHRA supports adding the Expert Determination method as an acceptable option for de-identification under the GDS Policy. When using this method to de-identify data sets, the person responsible for determining that there is only a "very small risk" of re-identification should be made aware of the intention to submit the data set to an NIH Repository, as well as that repository's policies for access and re-disclosure of the data set. Those details will influence the determination of the degree of risk of re-identification, though we note that it is unlikely that many organizations would have access to large enough genomic data sources to allow them to be able to re-identify patients. Clarify De-Identification Expectations with OCR There is ambiguity regarding the extent to which genomic data is considered a biometric identifier for the purposes of the 18 identifiers required for de-identification according to the HIPAA Privacy Rule. This has created challenges for entities engaging in research activities to know whether they have satisfied de-identification expectations when submitting data sets to meet obligations under the NIH's current Genomic Data Sharing Policy. We also note that it would be helpful if the definition of "de-identification" across various regulatory bodies was harmonized to improve the software development community's ability to respond to these issues in a consistent way. We recommend that NIH work with the Department of Health and Human Services Office for Civil Rights (OCR) to provide greater clarity regarding the types of genomic data or scenarios in which genomic data would qualify as a biometric identifier, setting clearer expectations for entities engaging in research activities that use genomic data. We also recommend adopting a policy that considers the degree to which the genomic data could be used to identify a unique individual. If the genomic data that is part of the data set being submitted could not itself be used to identify a unique individual, it should not be considered a biometric identifier.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Robust privacy and security measures must be implemented by NIH Repositories before it would be appropriate for potentially identifiable information to be submitted under the GDS Policy. When considering protections warranted in submitting data sets containing potentially identifiable data, the

EHRA recommends employing expectations analogous to HIPAA's privacy and security rules for the stewardship of protected health information, requiring the implementation of physical, administrative, and technical safeguards to prevent inappropriate access, use or disclosure of identifiable information. Further, repositories should be required to hold a Certificate of Confidentiality to prevent them from being compelled to disclose identifiable information. They should also require strict adherence to data use agreements for any individual or entity accessing potentially identifiable information, with commensurate penalties for unauthorized use or inappropriate disclosure. Entities with permission to access potentially identifiable data should be prohibited from attempting to re-identify individuals in the data set. We also request clarification to be sure that we understand what the NIH is proposing here. For entities who do want to be able to store and share identifiable information where appropriate, what exactly is being proposed? Is the NIH planning to create its own repository for public consumption? Would such a repository fall under the HIPAA framework for limited data sets, and if so, how would that work?

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

A narrow reading of this could lead to a policy that indicates certain linkages are not allowed even where it is possible to craft a linked data product that includes elements of both underlying datasets that does meet the GDS Policy requirements, where, in some instances, a full linkage would not meet the same requirements. It is important not to assume risk is inherently increased as a result of linking data sets. Recognizing the benefit of allowing combinations of existing data sets to enable more robust medical research, the EHRA supports permitting data linkage between datasets that meet GDS Policy expectations and potentially identifiable information. However, if linking two data sets compromises the de-identified nature of the resulting data set, in order to maintain patient privacy protections, we recommend that researchers combining or linking datasets be accountable for verifying that the resulting data set continues to be de-identified or take remedial action to de-identify the data set. If that is infeasible, the data set should not be re-disclosed without protections that would prohibit recipients from attempting to identify individuals who are a subject of the information and that would prevent the use or disclosure of the information for unauthorized purposes.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

At a high level, the EHRA is unsure whether there is a legitimate reason beyond public perception to treat genetic data differently than other protected health data. We believe that in practicality, it should be treated consistently with other sensitive health information, which all deserve robust and intentional protection. Making a decision to treat all sensitive data consistently would ease the work of technical resources, including the software development community, and could also do a great deal to address confusion among the provider community that currently has to remember different requirements and

limitations associated with different types of data. The EHRA does recognize the challenges inherent in prospectively informing participants about potential data linkages and appreciates the NIH objective to respect patient autonomy. Understanding that current rigorous GDS patient consent expectations include consent for the use of information in secondary research studies, we suggest that it is unnecessary to collect additional specific consent for linking data sets, particularly if due diligence is undertaken to validate that those linked data sets continue to meet de-identification expectations. We also point out that ensuring that consent is meaningful is an issue that is much larger than just this NIH request for comment. We recommend a separate request for comment around this one topic. Lastly, we agree that an Institutional Review Board (IRB) should weigh risks of linkage, current or future, but also be willing to revise policies as future use cases expand. The industry is still in the infancy of genomic data collection and sharing, so we caution against blanket policy approaches that may impede opportunities to maximize all patient data in the future.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/dNhaEEQWMW.pdf

Email: knicholoff@ehra.org

ID: 1912

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Ashley Delosh

Name of Organization: HIMSS

Type of Organization: Professional Org/Association

Role: Patient advocate

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/EcdlOfMIOl.pdf

Description: Official HIMSS comments in response to NIH Genomic Data Sharing Policy RFI

Email: ashley.delosh@himss.org

ID: 1914

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: COGR

Name of Organization: COGR

Type of Organization: Professional Org/Association

Role: Institutional official

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/okihCEhUgo.pdf

Description: COGR letter to NIH GDS RFI

Email: jbendall@cogr.edu

ID: 1915

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Deborah Motton

Name of Organization: University of California System

Type of Organization: University

Role: Institutional official

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

The UC system appreciates NIH's recognition that genomic data requires unique considerations, as the standard of de-identification for genomic data is a moving target. As understanding of genomics and analytic capabilities constantly improve, data that was once marked as de-identified now could be re-identifiable in the future. While UC supports expanding de-identification options to include the expert determination described at 45 CFR 164.514 (b)(1) as an acceptable method under the GDS Policy, we would like to note that Department of Health and Human Services Office for Civil Rights does not specifically address de-identification of genomic data, which can often be complex. In addition, the current GDS Policy requires that investigators follow both the Common Rule "readily identifiable standard " and HIPAA Safe Harbor method for de-identification. However, the two approaches are not always compatible. Therefore, we emphasize the need for guidance specific to the de-identification standard for genomics data. Such guidance, rather than a regulatory mandate that may add additional barriers to an evolving area, can offer a principle-based approach for the use and sharing of genomic data, and more importantly considerations around downstream use of such data. Such an approach can provide examples of safeguards and a tiered risk-based framework to help researchers, their institutions, and IRBs navigate issues in an environment where definitions, standards, and legal requirements are continuously changing. Additionally, we note the need for identifications tools in this space. For example, for projects focused on sequencing microbes that may unintentionally include a subset of identifiable human data, the National Center for Biotechnology Information (NCBI) has been helpful in providing tools for de-identifying such data. It would be important that such tools across NIH continue to be developed, funded, and made available to the community.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

We recommend that NIH allow submissions of Limited Data Sets as described at 45 CFR 164.514(e) to potentially include in repositories under the GDS Policy. However, we emphasize that if identifiable data are to be shared, consent should be required. In many cases, research participants provide their data in order to facilitate biomedical discovery. The risk of undermining that goal should be considered

alongside other risks. Likewise, the magnitude of risks to individual research participants should be considered in the context of other routinely experienced privacy risks (e.g., location tracking by apps and phones) rather than as isolated risk to be mitigated at all costs. We encourage NIH to be attentive to the independent submissions from UC researchers, such as from Steven Brenner on latent privacy risks in functional genomics data.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

UC supports the ability to link participant data from diverse datasets, such as electronic health records, with genomic information. For example, to meet the goals of precision medicine, data linkage can serve as a powerful tool to understand how participants'genetic and molecular characteristics may produce different outcomes throughout their lifetime. Simply securing genomic information in a repository has limited utility compared to linking genetic variants and mutations to clinical outcomes across the lifespan. In that regard, some amount of linkage will always be necessary to advance scientific understanding, but the scientific community needs guidance from NIH on the breadth of data linkages it would allow balanced with tradeoffs for protecting research participants' privacy. Specifically, the UC system recommends that NIH clarify the exact types of data linkage it would allow and any restrictions on combining certain datasets. Along with this guidance, we ask that NIH provide examples of what is allowable. To more concretely address the questions posed in section of the RFI, UC favors permitting data linkage. For datasets that may not meet the Policy expectations, research participants, upon being adequately informed about what will happen with their information and how it will be used, should be able to make the choice of whether or not to provide their data. This approach is discussed further below in Consent for Data Linkage. Importantly, consequences related to breaches to privacy from data linkages should not fall solely on research participants nor investigators demonstrating best efforts to abide by shifting privacy standards and requirements. As NIH moves into a new era of increased genomic data sharing, with potential changes in risks over time, we believe it is important for the agency to develop ongoing training opportunities but at the same time consider how to implement appropriate enforcement actions for misuse or inappropriate sharing or transfer of data, including monetary penalties.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

UC values the importance of properly seeking consent from those participating in research, and disclosing risks associated with data sharing. We agree that data linkages and risk considerations related to data submission to NIH repositories should be disclosed when obtaining consent for protocol under the GDS Policy. In cases where datasets are collected without providing disclosure on data linkages, we recommend that NIH note this to other investigators before providing information from the repository. It is also important to acknowledge that there is uncertainty in long-term privacy protections, and in

some cases, privacy protections are not guaranteed for years to come. The benefits of utilizing consent for data linkages include recognizing the autonomy of participants by affording them the opportunity to optimize the data they are contributing to research and maximizing societal benefits of their participation. Risks related to data linkage will be dependent on the types of data used. They may include an increased risk to confidentiality, possible damage to reputation, employability, criminal suits, among others, depending on the nature of the data. We ask that NIH develop guidance and sample consent language that address what data linkage means, how it could affect research participants, and whether subjects can still participate in a study without agreeing to data linkage. It is important that NIH be clear on its expectations on the use multiple data sources, including cases where data obtained utilized different consents or lacked consent. Such guidance and sample language would help researchers and IRBs who may not have the expertise in the intricacies of genomic data identifiability.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

The UC system agrees that the use and sharing of large-scale genomic data will continue to grow, and as such, the appropriate tools, resources, and platforms will need to be developed to keep pace with the needs. These NIH-supported resources, whether external or internal to NIH, should maintain appropriate standards and protections for data, and to adhere to existing principles for data access and security. For example, it would be important to have alignment in controlled access models, user authentication, and procedures for managing inappropriate or unauthorized use or access across the systems.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

We support inclusion of genomic data sharing within the DMS Plan, as outlined in the RFI, to avoid the duplicative submission of plans by the researcher. We also agree that these plans should allow for the budgeting for long-term genomic data management and sharing. We recommend that NIH assess these plans as part of Just-in-Time (JIT) documentation for extramural awards. Requiring submission of the plan at the JIT phase rather than at the proposal stage minimizes administrative burden for both the applicant and peer reviewers. While the DMS Policy sets a more flexible timeline for data sharing, any decision to change GDS Policy data submission timelines should involve careful consideration of potential impacts on policy compliance, data use and re-use, and what is most effective for the genomic research community.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).

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.**

Training and development awards (e.g., F, K, and T awards) should be excluded from this Policy. Small-scale studies, which may bare added risks to privacy, should be excluded as well.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

The UC system recommends that questions on whether and how to integrate data types beyond genetic information, such as transcriptomics, proteomics, microbiomics, or metabolomics, into the GDS Policy should involve further targeted outreach to experts, particularly as standards in these fields are still under development.

Description: University of California Comment Letter on NIH RFI on Genomic Data Sharing Policy

Email: agnes.balla@ucop.edu

ID: 1916

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Dana B. Hancock, PhD

Name of Organization: RTI International

Type of Organization: Nonprofit Research Organization

Role: Scientific researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

RTI supports the expansion of de-identified options for data sharing as a means to extend data sharing resources. The risks to participant privacy must be addressed through appropriate de-identification and data security protections; consent for research studies that permits data sharing; and using expert determination regarding the appropriateness of sharing and linkage of data obtained without consent (e.g., clinical data). Further, RTI recommends differentiating level 1 and 2 data from levels 3 (aggregated data) and 4 (summary statistics/analysis results). As described below, it is also important that the NIH attend to the potential for researcher and institution-based inequity, particularly to non-R1 institutions, that could inadvertently arise through the implementation of these efforts.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

The submission of some types of potentially identifiable information, such as zip code or dates associated with medical care, are viewed by RTI as acceptable for inclusion in repositories under the GDS policy. Participants are not likely to perceive such data as private and personal information. Expert consensus should be sought as to what potentially identifying information is permitted, and in what contexts. For future data collection, RTI recommends exploring the feasibility of allowing participants to elect what personally identifying information is made available, using defined categories of data that would allow a reasonable person to determine the extent to which the information is viewed as private and personal. RTI strongly endorses the use of expert determination, as described in the HIPAA Privacy Rule, as an acceptable approach to defining appropriate de-identification methods. HIPAA guidance on satisfying the expert determination method highlights expertise in statistical and scientific principles for de-identification, which is critical; additional expertise, such as in bioethics and/or community perspectives, may be useful in situations where the determination of 'very small' risk is particularly challenging.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and

whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

The linkage of data, and particularly clinical and research data, provides a rich research resource but raises additional ELSI and practical challenges. RTI recommends the use of standardized, global unique identifiers to connect data in different datasets without compromise to de-identification. In 2021, Congress reversed course to uphold funding for programs that implement a unique identifier, after previously prohibiting use of unique identifiers. The GDS Policy should permit data linkage between datasets that meet GDS Policy specifications. In addition, with the rise of new tools and technologies, including artificial intelligence and machine learning, the GDS policy should be expanded to cover a broader definition of data.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

In the case of new research data collection, RTI recommends that consent forms for NIH studies collecting genomic data clearly state that it is not possible to eliminate all risk of re-identification. RTI endorses obtaining consent for data linkage at study entry, even when no initial plans to link data exist (consistent with the NHI GDS Policy). Options for data linkage should be presented at the time of first consent, to include linkage with other research datasets as well as clinical (non-consented) datasets. Given the practical challenges that are associated with re-consent, informed consent forms should be developed to reduce the necessity of reconsent. Effective evidence-based practices for obtaining consent should be prioritized (e.g., Kraft et al. 2017. Clin. Trials 14(1):94–102; Doerr et al. AJOB Empir Bioeth. 2021 Apr-Jun;12(2):72-83). Such practices may not, however, be consistent with standard informed consent procedures upheld by individual IRBs. More specific guidance that is relevant to investigators and IRBs may be placed on the existing NIH webpage, “Informed Consent for Data Sharing” (https://www.genome.gov/about-nhgri/Policies-Guidance/Genomic-Data-Sharing/informed-consent-for-GDS). The NIH should consider prioritizing research funds to assess meaningful consent for data sharing; resulting outcomes could inform recommendations directed to investigators and regulatory agencies. Similarly, assessing attitudes toward data sharing among different populations could provide evidence of potential barriers to data sharing to be addressed. For historical research data, investigators should honor the existing consent to the extent possible; if it is feasible to reconsent participants, they should do so. Where it is not feasible to obtain reconsent for linkage with non-consented datasets, expert determination should be sought for whether linkage is acceptable, with the expert determination group to include relevant methodological expertise, a bioethics expert, and a community or patient advocate if warranted. The determination should include consideration of the potential for benefits and harms to the individuals in the datasets and the communities to which they belong. People from different cultures, racial backgrounds and generations may have different norms and expectations related to data sharing, which will need to be addressed in establishing policies. Large initiatives such as All of Us are making progress in this area, and policies should capitalize on existing research in working with diverse populations to ensure that consenting processes and data sharing practices build on evidence. RTI supports the need to determine whether specific data sharing requirements are necessary

for pediatric data. This remains a challenge to be addressed (e.g., Rahimzadeh V et al, AJOB Empir Bioeth. 2020 Oct-Dec;11(4):233-245).

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

Overall, RTI agrees with the suggestions around data submission, data access, and data security. Regarding Data Submission, there should be a formal agreement in place before data is submitted to NIH-supported resources. However, a full review and signature from a certified institutional signing official (SO) for each data submission seems overly restrictive for deidentified level 3 (aggregated data) and level 4 (summary statistics/analysis results) data that has been consented for sharing, especially given that the data sharing plan would have been reviewed by the SO in the original institution certification. This additional step could delay scientific discoveries, further strain project budgets, and would be particularly difficult in combination with tighter data sharing deadlines (e.g., the proposed timeline of sharing within 3 months after data generation) if enacted. Further, for investigators at non-R1 institutions, a lack of resources may constitute an unfair burden. We feel that clearly defining and expanding the definition of various levels of data in line with data levels from the NHGRI GDS data standards (https://www.genome.gov/about-nhgri/Policies-Guidance/Genomic-Data-Sharing/data-standards) and outlining specific requirements by level would aid in the interpretation of these policies and could be used consistently throughout data sharing agreements. These data standards could be expanded to provide examples from additional omic data types for improved clarity. In addition, providing a standard data submission agreement template for NIH funded repositories would be helpful for both the repository managers and investigators. Regarding Data Access, we agree that repositories and platforms should expect users to comply with the "NIH Security Best Practices for Controlled-Access Data Subject to the NIH GDS Policy." Of note, the GDS requirements could result in inequity for non-R1 intuitions, which do not have the information technology (IT) support network to maintain security recommendations that are necessary with this policy. Federated data systems such as NHLBI BioData Catalyst and NHGRI AnVIL, which provide the necessary security, storage, and computing environment, along with the IT support to minimize these disparities. The availability of such systems should be made aware to the researcher when applying for access and downloading the data. Lastly, the best practices mention that data should be housed on a system without internet access and behind an institution firewall. We agree with this, but as mentioned above, deidentified level 3 (aggregated data) and level 4 (summary statistics/analysis results) are data types where such tight security may not be needed. Regarding Data Security, repositories and platforms should employ an authentication system, such as eRA commons, with two-factor authentication. Implementation of these authentication systems should be accompanied with standards. There should be clear expectations to the user and repository maintainers as to what user data should be collected and maintained. This user registration data should be considered sensitive, personally identifiable information (PII), and protected in all necessary means. Further where multiple systems are integrated into one system the complete system should align with researcher auth services (RAS), therefore having the user to only sign on once for all services within the system. All genomic repositories should comply with the Federal Information Security Management Act (FISMA, for ensuring data protection), Federal Risk and Authorization Management Program (FedRAMP, for data protection on cloud computing), and authorities to operate (ATO) when dealing with level 0

(raw data) to level 2 data (raw data after initial cleaning) or data containing PII. Clear guidelines and best practices should be outlined for information security, administration, and scientific personnel about which levels of genomic data fall into the impact levels of FISMA and FedRAMP. All security measures should start with an ATO to operate within the data security framework outlined in the data use agreement. RTI further supports the suggestion that repositories and platforms should have procedures in place for handling data management incidents, a communication plan to notify appropriate NIH staff of incidents, and report data use statistics. Overall, providing standardized templates and outlining specific data use statistic reporting mechanisms would help maintain consistency across platforms.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

RTI supports the proposed changes to harmonize the GDS and DMS policies. Requiring a single data sharing policy, while still retaining genomic data sharing requirements for genomic data, would reduce burden on researchers and simplify the process while fulfilling the original intent of data sharing plans. RTI also supports harmonizing the designation of data as "sensitive" and budget review process for genomic data with the current DMS policy. RTI supports the sharing of non-human data consistent with the DMS and does not foresee a negative impact to extending the timeframe to "by time of publication or closeout of project". RTI supports a longer timeline for requiring data sharing in general. The "3 months after data generation" sharing requirement is burdensome and often not realistic or met by researchers. Implementing quality control measures is a critical part of data analysis that requires a significant investment of time. Longer timeframes allow researchers to fully analyze and quality control data prior to sharing, ensuring that researchers are satisfied and comfortable with the quality of the data they share. It also reduces the risk of researchers having to revise mistakes that are discovered after data release which requires informing any research groups already using the data of the corrections and could be time-consuming. Additionally, time spent preparing data for sharing/submission and navigating the submission process takes away valuable time from data analysis. Allowing a longer timeframe for data sharing allows researchers time to share only high-quality data that is ready for release and avoids premature release of data that has not been through sufficient quality control.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

RTI supports the proposed policy changes, however there are some points we encourage NIH to take into consideration. Although sharing data at the "time of publication" is sufficient in most cases, there are cases in which rapid release of data prior to publication (e.g. 3 months after data generation) benefits the broader scientific community. These cases include studies creating reference datasets, cataloging data, large-scale studies, studies of timely public health impact (e.g., COVID), or studies intended for use by the broader scientific community. The current policies set the data sharing timeframe for human and non-human data differently, with the human data having the shorter timeframe. However, we suggest having a timeframe based on study type. Certain types of human data do not need to be shared on a rapid-release timeline, and similarly there are significant benefits to the

sharing of animal data prepublication. For example, early access to the mouse genomic datasets in Geneweaver (https://www.geneweaver.org/) could be valuable to scientists using mouse models. Additionally, as the division between human and non-human data becomes increasingly ambiguous with recent advances in technology, it becomes more difficult to disentangle human and non-human data for data sharing purposes. For example, there have been advances in machine learning that increase analytical power through the combining of human and non-human datasets. This suggestion of delineating data sharing timelines by study type as opposed to data source (human vs non-human) is based in part on a 2009 workshop which brought together scientists, ethicists, lawyers, and editors to discuss the topic of prepublication data sharing (rapid release of data) [Toronto International Data Release Workshop Authors. Prepublication data sharing. Nature. 2009 Sep 10;461(7261):168-70]. The group developed a set of policy recommendations and guidelines, dubbed the Toronto Agreement, governing which publications merit a "rapid release" requirement to benefit the broader scientific community. The Toronto Agreement authors endorsed the rapid release of large reference datasets with broad utility spanning a diversity of data types including chemical structure and data from tissue banks. They categorized projects as either "prepublication data release recommended" or "prepublication data release optional". Their recommendations were based on study type and the nature of the data and did not take the data source (human vs non-human) into account. For example, "Genomewide association analysis of thousands of samples" was recommended for rapid data release whereas the recommendation for "Genotyping of selected gene candidates" was to make rapid release optional. In the Toronto Agreement, studies were separated into 10 categories, from Genome sequencing to 3D-structure elucidation studies, and examples of studies where prepublication release would be recommended versus optional were provided for each category. If NIH were to adopt a study-type based data sharing policy, it wouldn't be necessary for NIH to be as granular as the Toronto Agreement in the breakdown of study types. For example, projects could be split into two categories for data sharing purposes, "reference studies" and "non-reference studies". The "reference studies" category would encompass the study types described above as having broad utility, such as studies creating reference datasets or cataloging data. Projects which fit into this "reference studies" category would operate on a rapid release timeline requiring release of data within 3 months of generation. Projects considered to be "non-reference studies" would require data sharing "at time of publication or closeout of project". Although human data has unique requirements for data security, this does not necessarily hold true for data sharing. Thus, data sharing requirements would need to be separated from security requirements, which are necessarily based on the human or non-human origin of the data. Project budgets would also have to take into account any extra costs incurred by "reference studies" due to the rapid data release requirement.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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**

("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).

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.

RTI supports a broad definition of data types to be covered under the GDS policy to maximize the potential of NIH-funded research. With the rise of Artificial Intelligence and machine learning (AI/ML) tools, the field is moving beyond single layer data analyses (for example, genome-wide association studies) to concurrent integration of multiple data types (e.g., omics, environmental exposures, clinical and phenotypic data) that are linkable. Linkable data enables multiple data types from the same individual to be aligned and analyzed together. In its simplest form, the reuse and reanalysis of even single data layers are limited when few outcomes, exposures, and demographics are provided (for example, sharing genomic data in dbGaP but providing limited phenotypic data beyond the original study's primary outcome). Multi-faceted data will enable new discoveries, and the clinical utility of these discoveries will rely on understanding the underlying AI/ML algorithms and training models and having the data interoperate and reproducible. In the current state, interpretation of AI/ML findings is often complicated by not fully understanding how the algorithms and training models work, resulting in what is often referred to as the "black box". To advance from this current state, more open sharing of multiple data types is needed, recognizing that careful curation of the data and understanding of the original study design are essential to mitigate risks of participant identification but is feasible. With regard to sample size, there isn't a gold standard threshold for declaring large vs. small, and RTI recommends that even studies with fewer than 100 participants and studies with other omics be covered by the GDS policy. Smaller datasets can bring great value in varied ways, including: serving as the basis for methodological comparisons, being used as the source data for advancing researcher/scientist informatics skills and expertise, and enabling more statistical power via combination with other relatively small studies. Datasets with more deep phenotyping (i.e., comprehensive collection of individual-level traits to characterize disease manifestations and bodily systems) and studies collected to assess outcomes of rare diseases tend to be smaller in sample size, and when analyzed together (for example, when combined through meta-analysis), the smaller datasets collectively offer much greater statistical power for making discoveries. To enable this type of meta-analysis and have them done most efficiently, RTI encourages NIH to make smaller datasets publicly available. Similarly, data sets with multiple omics layers tend to be smaller in sample size, and the broader data may be more powerful than larger data sets with single omics, and thus also need to be made publicly available. The promise of AI/ML and concurrent integration is that the sum of multiple omics is greater than their parts. RTI recognizes that funding of genomic sequencing and related technologies by NIH and other sources will require careful consideration. Sequencing costs are one aspect of a full research plan, and if NIH funds are used to support other aspects of the research leading to the sequencing (for example, recruitment of the cohort and samples and/or computing costs for analyses), RTI recommends that these data should be covered under the GDS policy, in acting on what is best for the science.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

RTI supports providing both raw and processed data, with complete experimental information, is best for achieving high quality and reproducible science. Further, one of the greatest challenges in public sector science is interoperability, so a key feature of sharing multiple data version and data types will involve making these data linkable. While difficult, making data open, linkable, yet protected often yields the greatest amount of scientific progress. However, complying with additional data sharing requirements and/or preparing data in a shorter timeline for sharing must be balanced with the associated personnel and other costs. For most study types, RTI supports a timeline that enables a sufficient embargo period, such as time of publication or the end of the award period (i.e., at close out, following no-cost extension if needed) whichever comes first. RTI requests that NIH refrain from mandating a shorter timeline, in general, as study investigators often have not yet thoroughly performed their quality control procedures, and providing multiple versions of the data will be needed as quality control and statistical analyses are completed and cleaner data become available. RTI recognizes that there are special cases where NIH may need to insist on more rapid sharing of data that would benefit the scientific community at-large, such as creating reference datasets, cataloging data, or conducting large-scale studies on timely public health issues (e.g., COVID-19 research). In these special cases and overall, RTI points out that project budgets need to account for added personnel costs for the time spent preparing multiple data versions for sharing and navigating the submission process, in addition to their other project responsibilities.

Email: dhancock@rti.org

ID: 1917

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Michael Saito

Name of Organization: Epic

Type of Organization: Other

Type of Organization-Other: Health IT Developer

Role: Member of the Public

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/etqQMJJAql.pdf

Description: Epic's Comments on the Genomic Data Sharing Policy

Email: msaito@epic.com

ID: 1918

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Mary Lee Kennedy

Name of Organization: Association of Research Libraries

Type of Organization: Professional Org/Association

Role: Institutional official

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/VJPQJpKwwD.pdf

Email: mkennedy@arl.org

ID: 1919

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Brian Scarpelli

Name of Organization: Connected Health Initiative

Type of Organization: Professional Org/Association

Role: Medical provider

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

See attached letter.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

See attached letter.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

See attached letter.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

See attached letter.

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

See attached letter.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

See attached letter.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

See attached letter.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

See attached letter.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

See attached letter.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/gEuLxOnRKg.pdf

Email: bscarpelli@actonline.org

ID: 1920

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Regulatory Intelligence

Name of Organization: Regeneron Pharmaceuticals, Inc.

Type of Organization: Biotech/Pharmaceutical Company

Role: Institutional official

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Please refer to Regeneron comments provided in the attached document.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Please refer to Regeneron comments provided in the attached document.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Please refer to Regeneron comments provided in the attached document.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Please refer to Regeneron comments provided in the attached document.

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

Please refer to Regeneron comments provided in the attached document.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

Please refer to Regeneron comments provided in the attached document.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

Please refer to Regeneron comments provided in the attached document.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Please refer to Regeneron comments provided in the attached document.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

Please refer to Regeneron comments provided in the attached document.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/enKWJxisKp.pdf

Description: Regeneron comments letter - NOT-OD-22-029

Email: regulatoryintelligence@regeneron.com

ID: 1921

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: William B. Coleman, PhD

Name of Organization: American Society for Investigative Pathology

Type of Organization: Nonprofit Research Organization

Role: Institutional official

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

ASIP supports expanding de-identification options under the GDS Policy to include expert determination as described in 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule). However, it should be noted that HIPAA neither defines the qualifications of the expert nor sanctions any set of methods that would yield such a determination. Moreover, there is no uniform set or standard that defines very small risks or accounts for the fact that those risks may change over time. Notably, institutions may not have the expertise to confidently make such determinations, particularly for complex datasets. Therefore, we encourage the NIH to continue to explore the use of alternate de-identification strategies and support the development of such strategies so that they may be made more accessible to the research community at large. We also encourage NIH to develop an approach to ensure the routine assessment of changes in risk that may develop over time as a result of changes in the information environment, security vulnerabilities, or computing capability. We urge NIH to develop policies that will promptly remediate new vulnerabilities as they emerge or are identified. Finally, we urge that NIH promptly disclose to research participants the inappropriate disclosure of their research data, whether intentional, unintentional, or the result of malicious activity.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

ASIP supports the use of potentially identifiable information under the following circumstances: • IRB reviews the research to determine the risk associated with the research use and sharing of potentially identifiable information to ensure appropriate risk-benefit ratio; and only approves research with an appropriate risk-benefit ratio. • IRB review incorporates appropriate expertise in data science and cybersecurity, as evidenced by specialized training and/or work experience. • The research participant has been informed of the risks and benefits of the research and has provided consent for the research use and sharing of potentially identifiable information. • All data are stored in a secure database with controlled access. • A certificate of confidentiality is issued. • That every user requesting access to data must sign a Data Usage Agreement that prohibits purposeful re-identification. We take this opportunity to encourage the NIH to review their data security and access policies to ensure they are adequate for

genomic datasets that include potentially identifiable information and to implement any necessary changes prior to permitting the submission of potentially identifiable information.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Data linkages will increase the potential for re-identification and loss of privacy. ASIP supports data linkage between datasets that meet GDS Policy expectations under the following circumstances: • Expert determination determines that the risk for re-identification of the linked datasets is low. o The risks associated with data linkage will vary such that each data linkage event should be subject to expert review and determination of risk. To this point, additional framework and guidance is needed for IRBs, data access committees (DACs), etc. who are responsible for reviewing research using linked datasets. • Linked datasets are stored in a secure database with controlled access. • A certificate of confidentiality is issued. • That every user requesting access to data must sign a Data Usage Agreement that prohibits purposeful re-identification. • In the future, research participants should be informed of the risks associated with data linkages and provide informed consent (see below). o We acknowledge that this is not possible for previously collected datasets and feel that expert review (e.g., IRB, DACs, etc.) of the potential risks of data linkages is a required and appropriate safeguard that facilitates the maximal utility of existing datasets compliant with the GDS Policy. Data linkages to datasets that do not comply with GDS Policy such as data derived from archival diagnostic tissue blocks found in pathology departments is a common example of datasets that often do not meet GDS Policy expectations because often the research is exempt under 45 CFR 46.104 or is performed with a waiver of informed consent under 45 CFR 46.116. As a society of investigative pathologists, we acknowledge the significant scientific impact of data sources such as these. However, we propose that before expanding the GDS Policy to include datasets that do not meet current GDS Policy (such as those derived from clinical settings without informed consent or datasets derived from direct-to-consumer genetic testing), the NIH first implement processes for data linkages for those that do meet GDS policy and demonstration to both the public and the research community that appropriate safeguards may be applied to datasets.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

The benefits and potential risks of genomic data sharing, data linkage, as well as the sharing of potentially identifiable information should be disclosed to research participants at the time of informed consent. We recognize that the risks of future data linkages cannot be reasonably ascertained at the time of consent, nor can they be quantified; however, we feel strongly that research participants should be made aware of the increased risk of loss of privacy and confidentiality. As such, we encourage the NIH to add language pertaining to data linkage to the informed consent language for future use of data and biospecimens (NOT-OD-21-131). This should include suggested language and/or methods for

mitigating risks associated with data linkages. Establishing recommended consent language and corresponding guidance will provide necessary guidance to the research community, IRBs, DACs, etc.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

All the established principles for Data Submission, Data Access, and Data Security should apply to NIH-supported resources to ensure that genomic data protections are consistent with the terms of the GDS Policy. Additional considerations may be required for storage of potentially identifiable information and data linkages. If the GDS Policy is expanded to permit these activities, the NIH should reconsider whether the current Data Security requirements of FISMA and FedRAMP Moderate Authority to Operate (ATO) are sufficient. Furthermore, any data management incidents (DMI) related to potentially identifiable information and/or data linkages should be closely monitored by the NIH to identify any gaps in Data Submission, Data Access, and/or Data Security requirements.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

ASIP agrees that GDS and DMS policy harmonization is appropriate and will ultimately decrease the burden to the scientific community.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

ASIP agrees that the timeline for genomic data submission under the GDS should be harmonized with the DMS Policy. Allowing more time for data submission will alleviate current challenges associated with the current GDS Policy submission timeline and allow investigators more time to ensure the quality of data submitted. For data that will become part of an academic publication, waiting until the data pass peer review would add another layer of certainty that the data are meaningful and correct.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**

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.

a. ASIP recognizes that other types of "omic" data may warrant the types of protections afforded to genomic data by the GDS Policy. However, prior to expanding the GDS Policy to include these alternate data types, the NIH should consider whether the soon to be implemented DMS Policy affords the necessary protections. If found lacking, future expansion of the GDS Policy to such research/data may be warranted. b. ASIP believes that small scale studies, including training and development awards, should not be covered under the current GDS Policy. The amount of funding awarded is likely inadequate to meet the requirements of the GDS Policy. Compliance with GDS Policy and data submission for studies deriving genomic data should be encouraged but not required and should not be prohibitive of next award. c. Yes, any NIH-funded research that generates large-scale genomic data, even when NIH funding does not directly support the sequencing, should be covered by the GDS Policy. The scientific impact of including these data in NIH/NIH-supported resources is critical to the mission of the NIH.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/LVzRsJNoeq.pdf

Description: Summary of Responses to NOT-OD-22-029

Email: wbcoleman@asip.org

ID: 1923

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Name: Glenn Martin

Name of Organization: ISMMS

Type of Organization: University

Role: Institutional official

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

I believe the premise of this question is incorrect. The genome is the one true identifier and to continue policies that do not acknowledge that fact is a major flaw. Since HIPAA came into effect and the GDS policy was initiated the sheer number of genomes available in some form other, privately, commercially, publicly, restricted or not has exploded. Cross identification of individuals and family is an established fact. To continue without this recognition is not helpful.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

With truly informed consent there should be very few limitations. Adding elements should not be allowed after the fact without obtaining the subjects' specific additional permission. I would be very mindful of increased use of location tracking software, 9 digit zips and certain geocoding. Individual houses and less than a handful of apartments may be included within a single zipcode.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

NIMH has been linking data for years with prospective informed consent and consideration. I believe it is limited to linking research data. This makes sense. This cannot be "forced" The OHRP admonition not force data sharing for unspecified future uses in the context of research interventions that hold out the prospect for benefit must be honored. This is not always the case, especially around NCI studies.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**

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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Please consider the fact that all clinical information is potentially sensitive. We tend to focus on psychiatric issues, STD's etc, but even something that may at first glance appear not to be sensitive, e.g. a history of a successful intervention for early cancer, can still negatively impact ones ability to find a spouse of a job. HIPAA de-identification should never be used as anything other than a "good start"

Email: glenn.martin@mssm.edu

ID: 1924

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Elizabeth A. McGlynn, PhD.

Name of Organization: Kaiser Permanente

Type of Organization: Health Care Delivery Organization

Role: Institutional official

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

See attached comments

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

See attached comments

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/xpQtRrmoxj.pdf

Description: Comment Letter

Email: lori.potter@kp.org

ID: 1925

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Gretchen Purcell Jackson, MD, PhD

Name of Organization: American Medical Informatics Association

Type of Organization: Professional Org/Association

Role: Member of the Public

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/IzEVfwJPzG.pdf

Email: carrie@korrisgroup.com

ID: 1926

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Name: Denise Dillard

Type of Organization: Health Care Delivery Organization

Role: Scientific researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Tribal affiliation is considered identifiable data by many tribes and many tribes would require tribal approval for the collection of and additional use of this data element. Further expansion of de-identification options may erode the protections that currently exist and introduces risk that will further diminish already low participation of American Indian and Alaska Native communities and community members in genomic research given relatively recent harms of Havasupai tribal members. Expert determination as described in the HIPAA Privacy Rule is not an acceptable method for de-identification of American Indian and Alaska Native data unless the expert is cognizant of the sensitivities of using tribal affiliation as well as the small sizes of many American Indian and Alaska Native tribal groups. Tribes should not be identified unless use of the data is authorized by the tribe or potentially a tribe has given its authority to an expert to make a de-identification determination on its behalf. It seems that a Tribal Data Access Committees should be established for any repositories that plan to include American Indian and Alaska Native data who could provide another level of protection prior to the release of "de-identified" data. American Indian and Alaska Native communities are inadequately protected under the proposed expansion of de-identification options as described.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Secondary research with genomic data has violated individual AI/AN and tribal consent (e.g, Havasupai case), and had consequences for other communities as well (e.g., Henrietta Lacks). Specific informed consent should be required for all studies in which genomic data are used, and broad consent should not be allowed. Tribal consent should be required for all uses of genomic data from American Indian or Alaska Native people. Even when data point does not identify the individual participant, the tribe may be named. If specimens and data are then used in ways not authorized by the tribe, there is the potential for group harm and stigmatization of the tribe in resulting publications and reports. Submission of data elements to repositories under the current GDS Policy already presents unacceptable risk for American Indian and Alaska Native peoples. Contrary to the focus of the current GDS Policy, as stated in this RFI, even the suggestion of submitting potentially identifiable information to repositories under the GDS Policy without tribal consultation and tribally approved plans for the protection of these

data does not strike "an appropriate balance between accelerating scientific research through rapid genomic data sharing and minimizing risk." Establishing institutionalized Tribal Data Access Committees in NIH institutes and agencies and outside entities with NIH-funded data repositories would be a step in the right direction for establishing appropriate protections. However, a Tribal Data Access Committee should not supplant the tribal authority of an individual tribe to govern use of potentially identifiable information from its members.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

The GDS Policy should not permit data linkage between datasets that include American Indian and Alaska Native data without informed consent of individuals and Tribal approval from all involved communities. The risk to American Indian and Alaska Native communities and individuals from the use of data without American Indian and Alaska Native individual and community consent would include further eroding the trust and willingness of American Indian and Alaska Native communities and community members to participate in genomic data gathering efforts. Without more specific provisions in the proposed revisions that specify the essential role of American Indian and Alaska Native tribes in providing research oversight of the use of genomic data in federally supported research, use of genomic data from American Indian and Alaska Native people should not be included in any datasets, linked or otherwise.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

If specimens and data are used for secondary analysis in ways not authorized by the tribe, there is the potential for group harm and stigmatization of the tribe in resulting publications and reports. Rigorous protections should be applied to genetic information and specimens containing DNA as they by definition contain individually identifiable information. Specific informed consent should be required for all studies in which an individual's DNA or data are used and general informed consent should not be allowed. Future research use of data should require informed consent for secondary analysis as people's preferences and willingness to participate in research may change over time. Further, any definition of a biospecimen should include information on the ethical protocols and policies involving biological samples collected from humans who have since passed away (or who are now deceased). Data linkage should absolutely be addressed when obtaining consent for data sharing and future use of data under the GDS Policy; however, it is unclear how such consent would be obtained and exactly what individuals and communities would be consenting to in such circumstances. Consent can only be meaningfully obtained from American Indian and Alaska Native community members if it emerges from research initiated by communities themselves for purposes that communities identify and prioritize as important to the community and where infrastructure to support community ownership and oversight of research data and dissemination efforts is established.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

Tribal Data Access Committees should be enacted for any data sharing that involves American Indian and Alaska Native community members' data. Tribal Data Access Committees should not supplant the tribal authority of individual tribes unless a tribe has officially ceded their authority to this Committee. The data management and sharing policies and principles related to American Indian and Alaska Native data should be initiated within American Indian and Alaska Native communities and organizations and then adopted by NIH, rather than extended from NIH to communities, in accordance with community standards and values. Broad data management and sharing is already problematic for American Indian and Alaska Native communities and further erosion of federal policies governing protection of participant data would only weaken the capacity of NIH funded researchers to work with American Indian and Alaska Native communities, leading to further underrepresentation of these communities in clinical research. The authority and role of American Indian and Alaska Native tribes in overseeing research on their lands and with their community members needs to be addressed in proposed changes. There is also a need to specify what consequences researchers and research institutions/groups/organizations face when harm resulting from research practice has been documented and reported.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

If this process strengthened the protection of American Indian and Alaska Native data and communities, for example through the establishment of a Tribal Data Access Committee, it could be a positive development. Any weakening of protections to American Indian and Alaska Native data and communities would be detrimental.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

GDS data should not be treated differently than other data and no data sharing or harmonization of American Indian and Alaska Native data should occur without approval by the appropriate Tribal authorities. Data sharing agreements that meet Tribal standards and requirements are necessary and critical for doing any research with American Indian and Alaska Native individuals and communities, including genomic research.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**

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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

GDS Policy should be revised to increase protections for data collected from American Indian and Alaska Native individuals and communities, not weaken them. These proposals appear to weaken rather than increase these protections, and in so doing are unacceptable without substantial revision.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

The proposed changes weaken rather than strengthen protections for research participants. Risks to communities and groups such as American Indian and Alaska Native people are not adequately considered. Policy language needs to be added that explicitly addresses the role of tribal review within research regulation to ensure that research occurring with American Indian and Alaska Native peoples truly meets ethical standards. Any tribal exceptions to the policy should be repeatedly stated.

Email: dadillard@scf.cc

ID: 1927

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Mette Peters

Name of Organization: Sage Bionetworks

Type of Organization: Nonprofit Research Organization

Role: Member of the Public

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

The degree and type of de-identification are not equivalent. Per HIPAA, data may be de-identified using a Safe Harbor (rule-based) or an Expert (risk-based) approach. Other country's regulations like the European GDPR favor using pseudonyms or codes to mask identifiers, which is also a risk-based method of de-identification. Neither rule nor risk based de-identification is ideal in all circumstances, since neither acknowledges disclosure of risk. The Safe Harbor method is currently the only de-identification option under the GDS policy. It requires the removal of 18 identifiers that are considered to enable direct re-identification of an individual. While the Safe harbor rule is easy to follow, this "all or nothing" requirement restricts the utility of data sets, and prevents important research that relies on some of these 18 identifiers. For example, research about the spread of diseases requires access to precise timelines and localization information. Longevity research, and other research on conditions that affect the elderly necessitates knowledge of specific age. Currently, full dates, full zip codes, and ages over 89, are some of the direct identifiers that must be removed to comply with the GDS policy as currently written. A more nuanced approach to de-identification is preferable. The risk-based approaches require that someone knowledgeable in statistics and scientific methods determines -and documents- that the risk of using the data to re-identify a data subject is low. These are contextual and should consider any possible data linkage and any likely method to perform re-identification. These imply that the risk of re-identification must be evaluated for each proposed data use at the time of processing (not at the time of data collection). A concern with this approach is whether assessing the risk of re-identification and the adherence to the data minimization requirement at the time of data request would delay access. As algorithms and AI get more sophisticated, the risk of re-identification is greater. Adding tiers of data protection could resolve some concerns and minimize compliance burdens. GDS could request disclosure of risk-assessment methods and provide identifiability risk metrics to datasets. Furthermore, in the absence of unambiguous subject consent, requiring ethical approval of requests to datasets that include identifiable information but are deemed de-identified using a risk-based approach.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with

consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

While data linkage is necessary for effective biomedical research, there are privacy risks involved every time one set of data is linked to another, enhancing the possibility of algorithmic identification of individuals or expanding the information to the point of easy identification of an individual. Data linkage also may involve technologies that are not secure or retain data that can be found and identified. These technical solutions must be grounded in a privacy framework, and if NIH allows data linkage, guidelines for technologies used for this must be spelled out in the policy or in an addendum. (https://doi.org/10.1093/jlb/lsaa010). Other technologies are available that could be used for privacy protection, such as using patient identifiers in seeded hash code combinations using a Health Insurance Portability and Accountability Act compliant SHA-512 algorithm that minimizes re-identification risk (http://doi.org/10.1093/jamia/ocv038) In the NIH Workshop on the Policy and Ethics of Record Linkage, the chair, Dr. Pilar Ossorio "noted that data linkage can raise policy and ethical issues, especially in contexts in which people never anticipated their information captured throughout clinical care would be used in research. Additionally, she noted the need to understand when, how, and what policy implications of how linked data should be used in analyses when considering federating data choices" (https://datascience.nih.gov/nih-policy-and-ethics-of-record-linkage-workshop-summary). For example, while linking genomic data to electronic medical records may create a rich, useful dataset for scientific discovery, the individuals must be made aware that their genomic data and their electronic medical data is being combined to create richer information about their personal health and genetic profile. Proper informed consent for data linkage must be robust, explaining what data linkage is and the risks it involves for potential re-identification.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

There are inherent dangers of personal information exposure when data linking occurs between datasets, especially those that do not meet GSD policy expectations. First of all, as noted in previous sections, it is difficult to obtain reasonable, meaningful informed consent that covers all applicable scenarios in which data linkages may impact the use and release of personal information, and the consequences that may result from it. The potential risks are complicated and detailing risks not only the individual study participant, but also their family and community is difficult to scale but yet, the risk of re-identification must be disclosed, and the potential for group harm from data linkages must be addressed. Granted these complexities, we propose that consent for data linking should be considered on a case by case basis with an ethics board that evaluates the impact on the individual, the next-of-kin, and the community.

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

a. We agree that repositories storing and sharing data under the GDS policy should obtain a data submission agreement on par with the Institutional Certification described under section IV.C.5 of the current GDS Policy and the additional Certification of Confidentiality as required by dbGAP. An additional consideration related to data linkages should be taken into account, such as IRB approvals for GDS and non-GDS policy data linkages b. We agree that the research community benefits from clear guidelines on how data must be managed and shared through the application of a data submission agreement at the time of ingress and a data access agreement at the time of egress. The sensitive nature of the data being shared may warrant degrees of controlled-access on this data. For instance, genomic data use may necessitate the execution of agreements with submitting or requesting institutions due to the risk of re-identification while other data, such as expertly determined de-identified clinical data, may benefit from less stringent requirements. Selecting from a menu of controls including user authentication, agreements executed by the institution, and/or acceptance of terms and conditions for use by individual users - these may be applied flexibly or leveraged across datasets based on specific use cases. A multi-tiered approach that is responsive to any given data's conditions for use enables greater and more efficient processing of requests, in support of federal data sharing and open science efforts, that has been a long-time challenge for researchers and contributes to a delay in their access to dbGaP-registered data. Moreover application of a stable and shared governance ontology, such as the Data Use Ontology (DUO, https://github.com/EBISPOT/DUO), enables researchers to query datasets of interest whilst also surfacing structured, governance attributes that inform them of user requirements and/or limitations to data use with which they must comply. Application of a governance ontology also provides greater transparency to the request and review process undergone by data access committees (DACs) and provides users with information regarding how requests are evaluated. As data is shared through federated or hybrid sharing models, i.e., federated data and centralized governance or vice versa, this governance ontology provides even greater context for how data may be utilized across platforms and repositories, including studies. A governance ontology can also help define "who" is able to access data - from research investigators to engineers to community members and the general public. We do have concerns about limiting data access to individuals with eRA Commons ID as that excludes independent analysts that are not eligible to receive federal funding. We propose instead a Validated User approach as exemplified by access to mHealth data on the Synapse platform (https://help.synapse.org/docs/User-Account-Tiers.2007072795.html#UserAccountTiers-ValidatedUsers). Finally, we support reporting requirements, e.g., data use statistics and data management incidents, that build trust and accountability among the research community and platform or repository administrators. We contributed to the GA4GH's Data Access Committee Review Standards Policy (DACReS, https://www.ga4gh.org/wp-content/uploads/GA4GH-Data-Access-Committee-Guiding-Principles-and-Procedural-Standards-Policy-Final-version.pdf) outlining reportable metrics on platform or repository use, such as collection size by data type, number of approved and denied requests, average length of time for making decisions on requests, and information about users. Reporting efforts provide transparency in the review process, including criteria that underpin access decisions, and engage research communities, including research participants, in the ongoing scientific process. Reporting on incidents that can compromise the integrity or security of data, and actions taken to correct and prevent these incidents, is also a critical component of maintaining public trust in data sharing efforts. c. The bar for achieving FISMA and FedRAMP Moderate Authority to Operate (ATO) is costly and resource intensive. Platforms and repositories that currently store data in accordance with GDS Policy would benefit from NIH support in achieving this threshold for data security. Without agency

sponsorship and adequate financial support, data sharing in future may be limited to private groups that can pay for this certification and limit services provided by not-for-profit organizations and/or exclude not-for-profit organizations from participating in the FedRAMP Marketplace as Authorized.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

The DMS policy requires submission of a full Data Management and Sharing Plan at application, while the current GDS policy requires only a minimal sharing plan at application followed by a more detailed plan and the Institutional Certificate as just-in-time information. We agree that harmonization with the DMS policy in this area would be beneficial. Submission of a data sharing plan at application has the advantage that it enables the investigator to fully assess budget needs for data management and sharing, and emphasize the message that data sharing should be an integral part of a research project. Submitting a data sharing plan at application allows for peer review comments and for the NIH over time to better learn what reasonable data sharing costs are. These considerations are if anything more pertinent to genomic data than other research data types due to size, data sensitivity, and research community needs. We also believe that there is a need for the Institutional Certificate to be part of the at application requirements in order to identify limitations on data use prior to funding decisions. To accommodate genomic data in the DMS plan we believe the DMS plan could be strengthened to be more explicit about what summarization level data will be shared, with an emphasis that raw as well as processed data should be provided, unless limited by consent. We believe that non-human genomic data should be treated on par with human genomic data in terms of DMS harmonization, as these are often valuable research resources that are costly to replicate. One specific concern is animal research ethics and reducing the need for wasteful research on animal models. We appreciate the amount of work required for data sharing, and that many non-human studies do not meet the current GDS sample threshold, but recommend that a simple sample size cutoff is replaced with an assessment of data value.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

We do not agree with the proposal to relax the expectations for when human genomic data should be shared. Reverting to a more permissive policy will not benefit the research community at large. The change in the policy would make it difficult for Data Coordination Centers as the funding is often on the same cycle as the data generating grants. If a large portion of data is only shared at the end of the funding period there is a risk that resources to support data sharing may be limited. We are also concerned that this may result in less compliance to data sharing as this will limit the use of the yearly progress report as a tool to monitor compliance to data sharing milestones. Instead, we recommend strengthening the timeline requirements through the implementation of data sharing updates as part of a grants yearly progress report, where deviations from the initial data sharing timeline and scope must be justified. We believe that non-human genomic data considered to be of sufficient scale data should be under the same data sharing timeline as human genomic data.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

a. We recommend including proteomics and metabolomics data to be included under the GDS policy due to emerging privacy concerns. This is particularly true of Mass Spec based data where recent studies have indicated a need for controlled access models (https://doi.org/10.1038/s41467-021-26110-4). In addition, many large studies generate data at multiple omic levels. Requiring data sharing under one policy allows for data linkages under the GDS policy. b. Limiting the GDS policy to studies with more than 100 participants has the risk of reducing data sharing for rare diseases and understudied populations, where sharing may be even more important than for studies of larger cohorts. We recommend removing a specific threshold and instead evaluate if small scale studies should adhere to the GDS policy based criteria such as uniqueness and research needs. We agree with the current GDS policy of not requiring F, K, and T awards unless the research involves the generation of large scale, or rare disease, genomics data. c. If the NIH funding is secured based on the premise of data being generated from non-NIH sources we believe the data should be treated according to the GDS policy

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

The GDS policy should apply to all omics and related metadata, and be considered for other data types that are of a sufficient scale, sensitivity, and research community value to benefit from from the GDS policy rules. An example has been the need to facilitate rapid and safe sharing of COVID research data, ranging from human omics, to clinical information, and viral sequencing. The growing need for establishing data linkages calls for a more uniform policy in terms of consent, IRB approvals and data sharing and use requirements.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/KuuWnXlhDO.pdf

Description: Cover Letter

Email: mette.peters@sagebase.org

ID: 1928

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Joseph M Yracheta

Name of Organization: Native Bio-Data Consortium

Type of Organization: Nonprofit Research Organization

Role: Scientific researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

While we appreciate the need for some data that is conventionally excluded in de-identified datasets by HIPAA, the alternative processes are underspecified and may not adequately protect members of small, identifiable communities. American Indians, Alaska Natives, federally unrecognized Indigenous and Latin Americans of high indigenous ancestry are easily identified by zip code, constellation of comorbidities, environmental and pharmaceutical exposures, type of health insurance, ICD10 codes, and genomic architecture (long runs of homozygosity). Local repositories in the emerging federated approach would require additional (costly and substantial) resources to use these approaches.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

The inclusion of additional identifiable information in a data resource that already includes sequence data (which is, of course, the most identifiable data of all) raises serious concerns. It is not clear how that could be centrally managed. As stated in the response to I.1, Indigenous peoples of the Americas are highly identifiable and customarily subject to overt and systemic racism that cannot be solved by the NIH and is not likely to be legislated against via Congressional action. As an alternative, we recommend continued investments in the emerging federated approaches that permit local control, including investments in the required systems for data security and personnel for monitoring and specific sanctions against commercialization from corporate entities until such a time that American Indians and Alaska Natives can capitalize on such data for themselves.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

The inclusion of additional data without consent and requisite tribal approvals is problematic & a violation of the Reserved Rights of Tribes stated in Treaty and International Law.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Consent is crucial for data linkage, but it is not at all clear what individuals would be consenting to. The growth of a federated system permits ongoing participant engagement and community oversight that might enable a more iterative approach. Continuing down the path of broad open and unrestricted consent in centralized resources raises serious concerns about the erosion of patient, family, and community interests in data. Such practices are evidence of the continuing breakdown of cultural, spiritual, legal, and community cohesion of Native communities that is antithetical to self-governance and self-determination. Such actions can also be considered by some as meeting the definition of genocide. They are genocidal in the usurpation of intellectual, economic, political and health resources.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

Because Tribes are Domestic Dependent Nations with Sovereignty recognized internationally and because such resources are communally owned and managed, we believe that local controls are required to be extended into NIH systems to govern shared resources. In most cases, these local controls are in much closer connection to participant, family, and community perspectives and should be preserved. Extending a weakened system of participant protection to local resources makes the situation worse from the perspective of patients, families, and communities.

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

Insofar as this process would weaken other protections, it seems problematic. Insofar as it would strengthen them, it seems good. This would require additional resources for stewardship in local repositories.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

Because of the sophistication of computing and machine learning, there is no reason to continue to treat genomic data as unique. All sources of data about individuals, families, and communities raise risks and deserve attention in similar ways. The problem is that the proposed policy weakens rather than strengthens these protections (unless one counts a waiver of one's future interests ala Henrietta Lacks,

in what happens to tissues and data as a protection). An extended timeline, however, does provide additional time to work out data-sharing agreements and would probably be a good idea.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Many data types raise related concerns, especially in small populations, but extending a weakened rather than strengthened system of protections would make the problem worse. It has already been demonstrated that proxy data in machine learning systems are at risk if not more than proximal data and this too should belong to Indigenous peoples. Further, the ethics of secondary data and use of sequence data was not conceived and constructed with Indigenous and vulnerable populations in mind. This is particularly risky in the new technological and commercializable environments. The landscape has changed significantly and so should the NIH and its ethical policies and estimation of social harm. This should be seen, instead, as an opportunity to tighten protections.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

Any policy that strengthens protections for participants, families, and communities would be welcome. Weakening protections by requiring broader and less restricted sharing of additional data types moves in a direction diametrically opposed to the goals and needs of Indigenous people.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/RscdfoarVJ.pdf

Description: Amerindigenous Datasharing Concerns

Email: joseph@nativebio.org

ID: 1929

Submit date: 2/28/2022

I am responding to this RFI: On behalf of an organization

Name: Pam Dixon

Name of Organization: World Privacy Forum

Type of Organization: Nonprofit Research Organization

Role: Institutional official

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Please see attached file

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Please see attached file

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Please see attached file

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Please see attached file

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

Please see attached file

Harmonizing GDS and DMS Policies. Any aspect of the approach to harmonize GDS and DMS Policies and Plans described in the Notice, including for non-human genomic data.

Please see attached file

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

Please see attached file

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

Please see attached file

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

Please see attached file

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/jYBfVarnQd.pdf

Description: Comments of World Privacy Forum

Email: pdixon@worldprivacyforum.org

ID: 1931

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Type of Organization: University

Role: Scientific researcher

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

Precision medicine approaches to the treatment of disease, in which genomics data is generated from patient specimens, is a state-of-the-art technology of great interest for guiding treatment decisions for genetic diseases and for cancer. In the clinical setting, precision medicine has already shown great promise for determining appropriate treatments for patients with solid tumors. For example, at the University of Washington, the UW-OncoPlex™ Cancer Gene Panel (a multiplexed mutation assay for tumor tissue) assesses mutations >350 genes related to cancer treatment, prognosis, or diagnosis. As per this website, https://testguide.labmed.uw.edu/public/view/OPX, the UW-OncoPlex™ uses next-generation "deep" sequencing to detect most classes of mutations, including single nucleotide variants, small insertions and deletions (indels), gene amplifications, and selected gene fusions. The test can also detect microsatellite instability (MSI) status and total mutation burden (TMB) for relevant cancer cases. Currently the test is designed to detect somatic mutations in cancer, and is not designed to detect germline (heritable) mutations. In order to expand the use of precision medicine to a wide variety of diseases, including those caused by germline mutations (whether monogenic or the result of multiple genetic variations, as in the case of most heart disease), it is necessary to study a vast array of patient specimens and companion medical data. By mining diagnostic and treatment data longitudinally from the medical records, and combining this health data with genomics studies conducted on specimens (often residual, archived from clinically-indicated biopsies), correlations can start to be made between genetic variation and treatment responses. While it is desirable to obtain patient consent for the conduct of these genomic studies, it may not be feasible, particularly when the number of patients is large and/or the patients are unavailable for consenting (for many logistical reasons) and/or the genomics studies are retrospective. The major question exists as to whether linking various data sets, including clinically obtained data, and conducting genomics/transcriptomics and other “omics” studies on clinically obtained specimens, and sharing that data on NIH databases, WITHOUT CONSENT, violates patient rights in a meaningful way. There is no question that the linking of these clinical data sets, including medical records, and retrospective studies on previously obtained, archived clinical specimens, would greatly advance our understanding of disease and treatment response. Currently, the IRB grants Genomic Data Sharing Certification to an investigator when the investigator meets the consenting requirements for sharing that data to an NIH database. In the event that consent has not been obtained (e.g., the patient is deceased or otherwise not accessible, or the consent form did not contain the specifically required GDS language), the IRB may grant GDS Certification if the investigator provides a

compelling enough reason for Certification (e.g., the patient has/had a rare genetic disease and they and their family "would have wanted" the data to be made widely available). However, it is not feasible to request an exception to the GDS policy for every specimen, not to mention every disease state, that is being studied. It is important for the NIH to understand that publications use the GDS policy as a standard, even for unfunded research. In other words, publications expect investigators to make data available, which investigators may achieve by uploading data from unfunded studies to NIH databases. Hence, NIH's GDS Policy has wide-ranging implications for what research can be done in general. Currently, doing any genomics studies on clinical specimens, without consent, and sharing that data, is not easily justified, including when the research is not Federally-funded. Please also note that a source of confusion for investigators is the difference between the various NIH databases, some of which are restricted and some of which are not. GDS Certification from the IRB indicates to which kind of database an investigator may upload patient data, based on the consent form, but investigators are not always clear on the differences between them. Investigators are keen to protect patient privacy, but in the case of genomics data, the likelihood that a patient may be re-identified from their data increases as time passes, as more individuals are sequenced and the hacking approaches becomes more sophisticated. The risk of re-identification is the same, whether or not a patient has given their consent to have their data uploaded to an NIH database. We look to NIH to determine robust methods to protect patient privacy. It is also important to note that the currently required consent language that describes the NIH databases, and the risks of data sharing, are difficult to explain to a research participant, and ostensibly difficult for a research participant to assess. This is not a justification for removing the requirement for informed consent, but it needs to be acknowledged. Given the existence of the Genetic Information Nondiscrimination Act, it is not clear what the real risks of re-identification are. It must be assessed whether a patient re-identified by their genomics data that was uploaded to an NIH database would lose their job or life insurance, or whether re-identification would harm others in their family. In summary, it would be hugely beneficial to the field of precision medicine, and the improvement in understanding of disease and treatment options, if clinical datasets and clinical specimens could be utilized for genomics studies, and the results of these studies could be shared on NIH databases, without specific consent from the patients.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/ECfNzTxLSZ.pdf

Description: Response to issue I.3. regarding Data Linkage

Email: ejonlin@uw.edu

ID: 1933

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Name: Steven E. Brenner

Name of Organization: University of California, Berkeley

Type of Organization: University

Role: Scientific researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Functional genomics summary data, such as gene expression values and DNA methylation sites, are widely believed to be safe and widely shared without restriction. In our recent work (unpublished draft manuscript attached), we found it is possible to infer distinctive genotypes from functional genomics summary data. Searches of genealogy databases using the inferred genotypes can directly identify a person's or their relatives' genomic profiles. In general, these data were safe to share at the time they were created, but today some could reveal research participants' identity and privacy information. Such risks continue to increase over time, activated by new techniques, new knowledge, and new databases. We recommend NIH consider this new type of risk when updating the GDS policy. Human research participants typically provide their personal biological materials and data in order to facilitate biomedical discovery. Barriers to unfettered access undermine this goal. This risk should be considered alongside other risks. Likewise, the magnitude of risks to individual research participants should be considered in the context of other routinely experienced privacy risks (e.g., location tracking) rather than as isolated risk to be mitigated at all costs. We have considered potential mitigations of the privacy risks from the functional genomics data described above (unpublished draft manuscript attached). In particular, participants should be made aware of risks of leakage or future re-identification. We also encourage more projects with unrestricted sharing of data, where research participants understand, acknowledge, and accept the risk that they could be re-identified, with all the potential consequences of this decision.

GDS and DMS data sharing timelines. Whether the continued use of earlier submission expectations for human genomic data in the GDS Policy (e.g., submission of human data within three months of data generation) is needed, or whether timelines should be harmonized with the DMS Policy expectations (i.e., sharing of data no later than the time of publication or at the end of the performance period, whichever comes first), as described in the Notice.

We request that NIH allow the delayed release of data subsets for the purpose of allowing technology evaluation and assessment. For example, the Critical Assessment of Genome Interpretation (CAGI) has over the past decade organized blinded community experiments aimed at objectively evaluating the quality of computational methods for interpreting human genomic variation. As the accumulation of genomic data continues to increase, the lack of ability to evaluate their impact threatens to undermine

the future and promise of genomic medicine. Computational methods offer a potentially powerful approach for exploring genomic data, but their reliability and clinical utility have not been established. Each CAGI experiment starts with participants being provided with prepublication genetic data on which they are asked to infer phenotypes. Predictions are subsequently evaluated by independent assessors who have access to the experimental or clinical outcomes. Such experiments have shown that several computational methods have major utility for research and clinical applications. Our analysis has additionally provided a basis for the development of improved methods, as well as for more calibrated and powerful use in clinical settings. Such findings would not have been possible without the availability of prepublication data to share with predictors. Similarly, recent breakthroughs in protein structure prediction, made possible through another community experiment (CASP), would not have been possible without the availability of prepublication data. To enable such research, the GDS should explicitly allow delayed public release of datasets being actively used in such assessments.

Uploaded File: https://osp.od.nih.gov/wp-content/uploads/rfi2021_scidash/uploads/XiBlQapLmX.pdf

Description: Unpublished draft manuscripts

Email: brenner@compbio.berkeley.edu

ID: 1934

Submit date: 2/28/2022

I am responding to this RFI: On behalf of myself

Name: Emily Bonkowski, CGC

Type of Organization: Nonprofit Research Organization

Role: Scientific researcher

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

Genetic data has been shown to be easily identifiable, particularly in rare disease and in underrepresented, isolated, unique, and Indigenous populations. I cannot speak to what policy would be acceptable as there may not be one policy that meets the needs of all potential participants and groups, but would implore the Institute to design a policy alongside these groups, rather than simply collect their input, to develop a policy.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

While data linkage increases the power of the data, it makes it harder to oversee and its use harder to regulate. If there are specific groups that oppose the eventual GDS Policy, data should *not* be allowed to be linked in this proposed way. Risks include secondary use not outlined at the time of the original research consent, study of stigmatized conditions or characteristics to make broad generalizations from the data set, and diminished autonomy of study participants. Though linkage may drive development of new therapeutics, it would also drive profitability of the data set's use by for-profit entities. Profit sharing should be considered in any GDS policy; participants are a valuable resource and should benefit from profit that is made from research, including from data linkage (which will likely lead to more profit, both intellectually and monetarily).

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Future use of data under GDS Policy should be a part of the data consent. IRBs should take a conservative stance on data submission, as it cannot be taken back once entered into repositories. Trained genetics professionals (i.e. genetic counselors) and dynamic consent systems should be utilized as part of this process. Specific items for comprehension should be agreed upon (e.g. my data may be

shared, my data may be used for other purposes including xyz, my data in some form may end up being sold or informing development of profitable endeavors, I will not benefit from my data being shared, etc.). Additionally, the typical model for informed consent in the US prioritizes the Western principle of autonomy above all else. This approach ignores other ethical and cultural frameworks for decision making. In particular, Indigenous peoples have sovereignty over their data and samples. One person's individual consent and eventual data sharing and linkage could lead to whole unique populations being implicated genetically. Historically, this has not gone well (see treatment and study of Havasupai tribe). Unique populations around the world are at risk for having their data used in ways that, individually may seem permissible, as a group, is not. Individual consent should not replace group consent, and unique groups should prescribe how consent should be handled. Participants should also have a way of knowing if or how their data has been used over time. Having worked directly with patients clinically and participants on a research basis, this is a priority for participants.

Data management and sharing principles for NIH-supported resources

a. **Any aspect of the principles described for Data Submission.**
b. **Any aspect of the principles described for Data Access.**
c. **Any aspect of the principles described for Data Security.**

While data sharing among NIH-supported resources and research will help a robust resource of data, push innovation forward, and help level disparities between well-sourced resources and those that have fewer resources, it is not without its risks. Again, unique or vulnerable populations will be particularly impacted by wide data sharing, and sharing cannot be undone. They receive no benefit and are most at risk for exploitation and harmful generalization.

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 (e.g., with consent for future use and to be shared broadly, as well as IRB review of risks associated with submitting data to NIH), even when data are de-identified.**
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 ("Implementation of the NIH Genomic Data Sharing Policy for NIH Grant Applications and Awards," NOT-OD-14-111).**
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.**

If smaller scale studies are in identifiable rare disease patients or unique or vulnerable populations, they should *not* fall under the GDS policy of that policy is to freely share and link data. While exempting "small scale studies" from GDS policy is perhaps intended to lessen data sharing burden for researchers, having the policy *not* apply would actually serve to protect those participants. This framing centers the participant.

Email: esbonkowski@gmail.com

ID: 1935

Submit date: 3/1/2022

I am responding to this RFI: On behalf of myself

Name: Janis Geary

Name of Organization: Arizona State University

Type of Organization: University

Role: Scientific researcher

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

NIH should work with Indigenous scholars and Tribal governments/leadership to ensure that de-identification processes respect Indigenous governance over identifiers of Indigeneity. Enabling researchers to link research that includes Indigenous identifiers is against the principles of CARE (https://www.gida-global.org/care). Currently, the variables listed under 45 CFR 164.514(b)(2) do not refer to identifiers of Tribal affiliations or Indigenous status.

Use of potentially identifiable information. The circumstances under which submission of data elements considered potentially identifiable to repositories under the GDS Policy would be acceptable, any additional protections (including for security) that would be warranted, and whether there is certain potentially identifiable information that would not be acceptable to submit.

The NIH should implement additional protections for data that is potentially identifiable for communities and not just individuals. For example, if there are identifiers that can link individuals with specific Tribal affiliations, this information could be used to do research that the Tribe is opposed to, is stigmatizing, or fails to account important context. An approach to data management that allows this type of research would be in violation of all of the CARE Principles of Collective Benefit, Authority to Control, Responsibility to Indigenous Peoples, and Ethics for research based on Indigenous Data.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

The CARE Principle of Authority to Control instructs that “Indigenous Peoples’ rights and interests in Indigenous data must be recognised and their authority to control such data be empowered”. This does not necessarily mean that data linkage must be forbidden outright, but that Indigenous governance must determine when and what data may be linked.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

Consent here seems to apply only to individuals who are contributing their personal data. Consent for data linkage needs to also consider community consent and Tribal consent. NIH should work with Tribal partners and Indigenous organizations to determine what consent language would be appropriate. Additionally, individual consent is not meaningful if participants are not made explicitly aware of risks to communities and Tribal populations when they consent to data linkage.

Data management and sharing principles for NIH-supported resources

a. Any aspect of the principles described for Data Submission.
b. Any aspect of the principles described for Data Access.
c. Any aspect of the principles described for Data Security.

The "Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research" refers only to the FAIR principles of data sharing. NIH should also support repositories and encourage use of repositories that are consistent with CARE principles for Indigenous Data Governance. The requirements of the GDS policy seem to preclude Indigenous Data Repositories from being NIH-Supported. NIH should either write the GDS policy so that Indigenous organizations and Tribes can be supported and named as acceptable repositories for Indigenous Data, or the NIH should make exceptions to the policy so that Indigenous data may be stored in non-supported alternative data management and sharing resources and still be considered to satisfy the GDS policy expectations.

Data sharing expectations under the GDS Policy. Whether there are other types of research and/or data that warrant the data processing level and timeline expectations established by the GDS Policy (e.g., sharing lower levels of processed data, not just those of sufficient quality to validate and replicate findings as in the DMS Policy).

The policy and requests for input seem to be based on the assumption that increasing the openness of data is naturally the best governance approach to increase the impact of generated data. However, studies of governance of shared resources (found in the literature on "knowledge commons") provide evidence that governance should be designed to support the intended outcomes the data resource was created to achieve. The NIH states that "Sharing research data supports the NIH mission and is essential to facilitate the translation of research results into knowledge, products, and procedures that improve human health". The intended outcome is not "shared data", but "knowledge products, and procedures that improve human health". The success of data sharing initiatives and governance of shared data should be evaluated against that goal. Instead, data sharing initiatives seem to be evaluated against whether or not they comply with a selected the governance approach (open data), meaning these initiatives are evaluated as successful if they create the most openly available data. However, in some situations open data is not the most effective governance approach to achieve the goal of improving human health. In the case of Indigenous data, and data about other similar populations, the threat of having data openly available for research that is not in the control of the source population will prevent participation in genomic research. Data that does not exist cannot have any impact on human health.

Creating non-open data governance approaches that enable these populations to participate confidently in genomic research does not lessen the impact of that data. Rather, it increases the quantity of data that exists and therefore contributes to the goal of improving human health. If the NIH has the desire to ensure equitable health impacts from genomics, they cannot use equal governance applied across populations with different data governance needs.

Email: jdgeary@asu.edu

ID: 1936

Submit date: 3/1/2022

I am responding to this RFI: On behalf of an organization

Name: Kevin Mcghee

Name of Organization: New York Genome Center

Type of Organization: Nonprofit Research Organization

Role: Institutional official

De-identification. The risks and benefits of expanding de-identification options, including adding the expert determination described at HIPAA 45 CFR 164.514 (b)(1) (the HIPAA Privacy Rule), as an acceptable method for de-identification under the GDS Policy, and whether other de-identification strategies exist that may be acceptable in lieu of HIPAA standards.

Broadening of de-identification criteria, as proposed, may provide substantial benefits as discussed in the RFI. NIH should provide guidance on the appropriate process for using statistical methods as an alternative to the HIPAA Safe Harbor standard. NIH should also consider the associated costs to grantees for data storage at closeout, data security for specialized data and DMP monitoring and compliance during the life of the award through closeout.

Data linkage. Whether the GDS Policy should permit data linkage between datasets that meet GDS Policy expectations (e.g., data obtained with consent for research use and de-identification), and whether the GDS Policy should support such linkages to datasets that do not meet all GDS Policy expectations (e.g., data may have come from a clinical setting, may not have been collected with consent, may retain certain potentially identifiable information). Feedback is also requested on risks and benefits to any such approaches.

NIH should provide more guidance concerning the types of data linkages that may raise concerns about risk of re-identification and require this additional level of consent, and how this new requirement will interact with the future generation of data from specimens that were collected prior to the effective date of this informed consent requirement.

Consent for data linkage. Whether data linkage should be addressed when obtaining consent for sharing and future use of data under the GDS Policy, as well as in IRB consideration of risks associated with submission of data to NIH genomic data repositories. And if so, how to ensure such consent is meaningful.

The proposal to explicitly address data linkage in the informed consent process should be accompanied by more detailed guidance to institutions than has been previously provided by NIH. NYGC is concerned that this adds another layer of complexity onto an informed consent process that may already be difficult for some research subjects to understand; as such, NIH should provide template language that can be incorporated into consent forms, along with guidelines for presenting this aspect of the consent process to research subjects. We also encourage the NIH to undertake systematic outreach to educate, inform, and engage the public to better understand the potential value and impact of data linkage and broad sharing of large genomic data resources as well as gathering concerns of public interest.