Points to Consider for Institutions and Institutional Review Boards in Submission and Secondary Use of Human Genomic Data under the National Institutes of Health Genomic Data Sharing Policy

Introduction

Under the National Institutes of Health (NIH) Genomic Data Sharing (GDS) Policy,[1] institutions and their Institutional Review Boards, privacy boards, or equivalent bodies (hereafter IRBs) are responsible for assuring NIH that plans for the submission of genomic and phenotypic data from research studies to NIH-designated data repositories meet the expectations of the Policy. The purpose of this document is to assist IRBs in their review of, and institutions in their certification of, investigator applications and proposals involving the submission and access of human genomic data under the NIH GDS Policy.[2]

Part I: The NIH GDS Policy and Institutional Responsibilities

A. The NIH GDS Policy

The NIH GDS Policy facilitates the sharing of large-scale genomic data (e.g., data from genome-wide association studies (GWAS),[3] single nucleotide polymorphism (SNP)[4] arrays, and genome sequence, transcriptomic, metagenomic, and epigenomic data) as well as phenotypic and other associated data generated in NIH-funded research.[5] A key element of the NIH GDS Policy is the expectation that data from NIH-funded human genomic research will be submitted to an NIH-designated data repository, such as the NIH database of Genotypes and Phenotypes (dbGaP).

B. Essential Role of Institutional Officials and IRBs in Implementing the NIH GDS Policy

IRBs and institutions have an important role to play under the NIH GDS Policy in reviewing data sharing plans for consistency with the NIH GDS Policy, as well as the adequacy of the informed consent process and documents used to obtain consent for the generation and secondary research use of the data. Because the volume of genomic and phenotypic data will be substantial and potentially sensitive (e.g. data related to the presence or risk of developing particular diseases or conditions and information regarding family relationships or ancestry), the confidentiality of the data and the privacy of participants should be protected (see Part III.B for more information on risks).

---

[1] NIH GDS Policy, see https://osp.od.nih.gov/wp-content/uploads/NIH_GDS_Policy.pdf.

[2] NIH recognizes that this review and certification process goes beyond regulatory requirements under 45 CFR part 46 as outlined in an October 2008 policy guidance from the Office for Human Research Protections (OHRP) entitled "Coded Private Information or Specimens Use in Research, Guidance (2008)." For the reasons outlined in this document, NIH, as a policy matter, will not accept human data into a data repository without the appropriate certifications from the institution and verification by an IRB, privacy board, or equivalent body that the submission criteria stipulated in the NIH GDS Policy have been met.

[3] GWAS is a study of genetic variation across the entire human genome that is designed to associate genetic variations with traits (such as blood pressure or weight) or with the presence or absence of a disease or condition. To meet the definition of a GWAS, the density of genetic markers and the extent of linkage disequilibrium should be sufficient to capture (by the r^2 parameter) a large proportion of the common variation in the genome of the population under study, and the number of samples (in a case-control or trio design) should provide sufficient power to detect variants of modest effect.

[4] A Single Nucleotide Polymorphism (SNP) is a variation in a DNA sequence that results when a single letter (A, T, C, or G) in the genome sequence is replaced by another.

[5] The NIH GDS Policy applies to competing grant applications submitted to NIH for the January 25, 2015, receipt date or after; proposals for contracts submitted to NIH on or after January 25, 2015; and NIH intramural research projects generating genomic data on or after August 31, 2015.

NIH will accept data into an NIH-designated data repository only after receiving appropriate certification by the Institutional Signing Official[6] of the submitting institution.

Part II: Data Sharing Plans, Institutional Certification, and Points to Consider Regarding Informed Consent

A. Data Sharing Plans

NIH expects all extramural investigators[7] proposing to generate large-scale human or non-human genomic data using NIH funding to include a genomic data sharing plan in the funding application. Intramural investigators[8] are expected to submit a genomic data sharing plan to their Scientific or Institute and Center (IC) Director prior to the start of research.

The data sharing plan should describe how the expectations of the NIH GDS Policy will be met and denote the type(s) of data to be submitted, which data repository(s) data will be submitted to, the appropriate uses of the data (i.e. Data Use Limitation), and the data sharing timeline. An IRB assurance of the data sharing plan should also be included, as well as any request for an exception to submission. The NIH Guidance for Investigators in Developing Genomic Data Sharing Plans provides expectations and examples of genomic data sharing plans for human and non-human research.[9]

B. Institutional Certification

An Institutional Certification[10] stipulating the appropriate secondary uses of data submitted to an NIH-designated repository should be provided by the Institutional Signing Official(s) of the submitting institution during the Just-in-Time pre-award process (or the start of research for NIH intramural investigators) when genomic data generation is proposed. The purpose is to assure that submission of data to an NIH-designated data repository is consistent with the NIH GDS Policy and with the informed consent of the original study participants. As part of the process to develop the Institutional Certification, the IRB should review the proposal for data submission and sharing included in the funding application. With respect to the nature of this IRB review, NIH defers to the institution submitting the data to determine what is appropriate. However, IRB review may be conducted in a manner consistent with the expedited review procedure described by 45 CFR 46.110.[11]

It is important that the submission of human genomic data to NIH-designated repositories be consistent with any local, state, or federal laws or regulations as well as any specific to the participants' community,

---

[6] An Institutional Signing Official is generally a senior official at an institution who is credentialed through NIH eRA Commons system and is authorized to enter the institution into a legally binding contract and sign on behalf of an investigator who has submitted data or a data access request to NIH

[7] Extramural investigators are composed of scientists, clinicians, and other research personnel affiliated with more than 3,100 organizations, including universities, medical schools, hospitals, and other research facilities located in all 50 states, the District of Columbia, Puerto Rico, Guam, the Virgin Islands, and points abroad.

[8] Intramural investigators are NIH scientists who conduct research and training activities in NIH laboratories on its campuses in the Bethesda (including the NIH Clinical Center), Rockville, Frederick, and Baltimore, Maryland, areas; Research Triangle Park, North Carolina; Detroit, Michigan; Phoenix, Arizona; and the Rocky Mountain Laboratories, Montana.

[9] For additional guidance on the *National Institutes of Health Guidance for Investigators in Developing Genomic Data Sharing Plans*, see https://osp.od.nih.gov/wp-content/uploads/NIH_Guidance_Developing_GDS_Plans.pdf.

[10] Institutional Certification forms are available at https://osp.od.nih.gov/scientific-sharing/institutional-certifications/.

[11] http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.110

population, or group. If the research involves tribal populations, the Institutional Signing Official(s) should consider tribal laws and regulations, and whether consultation with tribal communities may be appropriate.

Fillable Institutional Certification Forms[7] are available on the NIH Office of Science Policy website and a sample Institutional Certification form is provided in Appendix A of this document.

C. Considerations Regarding Consent

NIH recognizes that the issues related to determining the appropriateness of participants' consent for submission of human genomic data to NIH-designated data repositories and subsequent sharing for research are complex and may vary depending on the proposed research and, in particular, whether the specimens were collected after January 25, 2015. Under the NIH GDS Policy, NIH expects explicit consent will have been obtained to use research and clinical specimens and cells lines and strongly encourages investigators seeking consent to include consent for future research use and broad sharing of genomic and phenotypic data generated from such specimens. The *NIH Guidance on Consent for Future Research Use and Broad Sharing of Human Genomic and Phenotypic Data Subject to the NIH Genomic Data Sharing Policy*[12] and frequently asked questions (FAQs)[13] related to consent for broad sharing can be found on the NIH Office of Science Policy website.

The NIH National Human Genome Research Institute (NHGRI) has created an online Informed Consent Resource for genomics research. In addition to discussion about the basic elements of informed consent in the context of genomics research, it also provides information regarding other considerations for informed consent with particular relevance to genomics research, such as the type of informed consent (broad or specific), potential benefits and risks to research participants, and data and sample sharing.[14] Examples of consent forms used in genomics research and model consent language are also available through the NHGRI resource.[15]

Part III: Considerations for Sharing of Genomic Data

A. Benefits of the Broad Sharing of Genomic Data through an NIH-Designated Data Repository

Data sharing supports the mission of NIH, and NIH promotes and facilitates the sharing of genomic data because the data can be used to address multiple research hypotheses and can be aggregated in analyses of complex questions. In addition, access to genomic data from research studies facilitates validation of the original studies' findings and helps to ensure the integrity and transparency of NIH-funded research. NIH-designated data repositories (e.g. dbGaP) provide a central location for the registration of studies and access of data.

As of October 2018, NIH has provided over 5,600 investigators access to 1,025 studies, resulting in over 2,460 peer-reviewed publications contributing significant advances to a wide range of fields such as

---

[12] For additional information about the *NIH Guidance on Consent for Future Research Use and Broad Sharing of Human Genomic and Phenotypic Data Subject to the NIH Genomic Data Sharing Policy*, see https://osp.od.nih.gov/wp-content/uploads/NIH_Guidance_on_Elements_of_Consent_under_the_GDS_Policy_07-13-2015.pdf.

[13] For FAQs related to consent for broad sharing, see https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing-faqs/.

[14] NHGRI Special Informed Consent Considerations, see https://www.genome.gov/27559024/informed-consent-special-considerations-for-genome-research/.

[15] NHGRI Consent Forms Examples and Model Consent Language, see https://www.genome.gov/27559023/informed-consent-sample-consent-forms/.

cancer, mental health, addiction, cardiovascular disease, and computational biology. For example, access to NIH genomic data enabled researchers to identify a previously unknown association between Parkinson's disease and the immune system[16] which may offer new targets for gene therapy trials and drug development. The NIH Office of Science Policy website provides further statistics regarding the sharing and use of human genomic data obtained from NIH-designated data repositories.[17]

B. Risks Associated with the Submission and Broad Sharing of Human Genomic Data

Concerns associated with broad data sharing largely stem from the nature and extent of the genomic and phenotype data involved and the distribution of the data to Approved Users for secondary research. As in the review of any research, it is important to consider any possible risks in the context of the protections put in place to minimize those risks, as well as in the context of the expected benefits of the proposed research. Several risks and the NIH GDS Policy provisions to mediate those risks, are discussed below.

1. Risks of identification
Individual-level genomic data

Currently available and emerging technologies make the re-identification of specific individuals from raw genomic data increasingly feasible. For example, some research has demonstrated that data and other information in publicly accessible resources can be compared with genotypic or phenotypic information obtained from other sources to re-identify the individual who is the source of the data.[18,19] Risks of re-identification of research participants may be increased among small and easily identifiable populations; therefore, it may be appropriate to consider de-identifying research data from these populations at the community level.[20] Although the feasibility of using genomic data to re-identify an individual through matching with other data or information is increasingly recognized, the likelihood that a participant's data will be used to re-identify them is anticipated to be very small, but it is unknown.

The NIH GDS Policy stipulates that human data submitted to NIH-designated data repositories, such as dbGaP, are to be coded[21] and de-identified by the submitting investigator, and the key to the code that links the data to specific individuals held by the institution. In order to minimize the risk that research participant identities could be readily ascertained, data should be de-identified

---

[16] Hamza, TH, et al. Common genetic variation in the HLA region is associated with late-onset Parkinson's disease. Nature Genetics. 2010 Sep; 42(9): 781-5. doi: 10.1038/ng.642. Epub 2010 Aug 15. See http://www.nature.com/ng/journal/v42/n9/full/ng.642.html.

[17] https://osp.od.nih.gov/scientific-sharing/facts-figures/.

[18] Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013 Jan 18;339(6117):321-4. doi: 10.1126/science.1229566. See http://www.sciencemag.org/content/339/6117/321.long.

[19] Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. Science. 2018 Oct 11. pii: eaau4832. doi: 10.1126/science.aau4832. See http://science.sciencemag.org/content/early/2018/10/10/science.aau4832.long .

[20] More information about de-identification of genomic data from American Indians and Alaskan Natives is available from the National Congress of American Indians Policy Research Center at http://www.ncai.org/prc.

[21] *Coded* means that any identifying information (such as name or social security number) that would enable the original submitting investigator to ascertain readily the identity of an individual has been replaced with a number, letter, symbol, or combination thereof (i.e., the code), and a key to decipher the code exists, enabling linkage of the identifying information to the private information or specimens. See https://www.hhs.gov/ohrp/regulations-and-policy/guidance/research-involving-coded-private-information/index.html.

by standards consistent with both HIPAA[22] and the Common Rule.[23] NIH-designated data repositories must protect the data according to the appropriate federal standards for information protection. Only qualified investigators (e.g. tenure-track professors, senior scientists) may request access to human genomic data. During the Data Access Request (DAR) process, investigators and their institutions agree to adhere to the Data Use Certification (DUC) Agreement[24] and the Genomic Data User Code of Conduct[25], both of which state that users may not use the requested datasets, either alone or in concert with any other information, to identify or contact individual participants from whom data and/or samples were collected. Users also agree to implement the *NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy*[26], in addition to their own institution's IT security practices and policies.

Genomic Summary Results

NIH employs an unrestricted access model for genomic summary results (GSR)[27] from most studies, in line with the distinct risks and benefits related to this type of data relative to individual-level genomic data. However, NIH acknowledges that it is possible that privacy and confidentiality risks related to unrestricted access to GSR[28] may be heightened for study populations from isolated geographic regions or with rare traits. It is also possible that certain study populations may be more vulnerable to group harm due to potential for stigma related to traits being studied or other participant protection concerns. In addition, for studies that include data on potentially stigmatizing traits, the outcomes of any privacy breach could conceivably cause greater harm to research participants than is likely under most circumstances. Therefore, institutions submitting datasets to NIH-designated data repositories should indicate in the genomic data sharing plan and the Institutional Certification if GSR from incoming studies should be designated as "sensitive" and provided only through controlled-access procedures. In such cases, GSR will then be accessible only in conjunction with access to individual-level data and any Data Use Limitations (DULs) attached to use of the individual-level data will apply. When determining the appropriate access model for GSR from studies under their purview, institutions should consider whether the study includes potentially vulnerable populations (e.g., small sample sizes, isolated or identified geographic regions, Native Americans/Alaska Natives or other indigenous populations, rare disease communities) or potentially stigmatizing traits. The institution's considerations should reflect the perspectives of the study population(s) who participated in the research. When possible, consultation with communities and study populations that are involved in the research may be appropriate to determine their perspectives about the balance of privacy concerns relative to the priority that many communities may have to support broad data sharing in order to advance research.

---

[22] For additional information about HIPAA, see http://www.hhs.gov/ocr/privacy/.
[23] For additional information about the Common Rule, see, https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html.
[24] https://osp.od.nih.gov/wp-content/uploads/Model_DUC.pdf.
[25] https://osp.od.nih.gov/wp-content/uploads/Genomic_Data_User_Code_of_Conduct.pdf.
[26] For additional information about the *NIH Security Best Practices for Controlled-Access Data Under the Genomic Data Sharing (GDS) Policy*, see https://osp.od.nih.gov/wp-content/uploads/NIH_Best_Practices_for_Controlled-Access_Data_Subject_to_the_NIH_GDS_Policy.pdf.
[27] For the purposes of the NIH GDS Policy, genomic summary results are defined to include those provided by a study's investigator, if any, as well as summary statistics that may be computed by the relevant NIH-designated data repository across all non-sensitive studies with data included in that repository. Genomic summary results include systematically computed statistics such as, but not limited to: 1) frequency information (e.g., genotype counts and frequencies, or allele counts and frequencies), and 2) association information (e.g., effect size estimates and standard errors, and p-values). These values may be defined and calculated using scientifically relevant subsets of research participants included within study populations (e.g., disease, trait-based, or control populations).
[28] http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000167

2. Psychosocial and other harms

Certain data that include potentially stigmatizing genetic, phenotypic, behavioral, or social traits (e.g., mutations associated with neurological or psychological disorders) may merit particular consideration during IRB review of proposals for data submission. Harms (e.g., stress, anxiety, stigmatization, or embarrassment) to individuals, groups, or populations may potentially arise from the disclosure of such data. For example, some populations demonstrate a higher predisposition to develop certain diseases or disorders than others are generally known to do. Higher or lower frequencies of genetic variants that contribute to observed health patterns within these populations might be used to discriminate against or otherwise stigmatize any member of the population group, whether they possess a given genetic variant or not. Additionally, some types of research (e.g. studies of ancestry) may be considered objectionable to certain populations or groups. The IRB should consider delineating the appropriate parameters for use of the data through the use of DULs[29] that could minimize the potential for harm to individuals and their families, groups, or populations.

Note that the Genetic Information Nondiscrimination Act and the Affordable Care Act prohibit the use of genetic information in health insurance or employment decisions.[30]

3. The Federal Freedom of Information Act (FOIA)

Genomic and associated phenotypic data submitted to an NIH-designated data repository become U.S. government records subject to FOIA. NIH is required to release government records in response to requests under FOIA, unless certain exemptions apply, one of which is if the release of the records would result in an unwarranted invasion of personal privacy. Additionally, Section 2013 of the 21st Century Cures Act, P.L. 114-255, enacted Dec. 13, 2016, provides NIH with new authority to exempt certain information collected or used during the course of biomedical research from disclosure. Because of the potential risk to personal privacy due to the nature and volume of genomic data held in NIH-designated data repositories, NIH intends to deny any FOIA request for such data. However, it is possible that NIH's decision to withhold genomic data could be challenged in court. A similar concern exists for research data held by grantees who are subject to state-level freedom of information laws.

4. Potential for Access by Law Enforcement

Law enforcement agencies could conceivably seek to compel disclosure of de-identified genomic data held by a submitting institution or within an NIH-designated data repository to search for matches to DNA specimens collected for forensic purposes. Certificates of Confidentiality protect against compelled disclosures of "identifiable, sensitive information"[31] in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level and may provide an additional safeguard for participants. Investigators and institutions conducting studies collecting or using genetic and other information that, if disclosed, could have adverse consequences for participants such as compromising their financial standing, employability,

[29] For additional information about Data Use Limitations, see https://osp.od.nih.gov/wp-content/uploads/standard_data_use_limitations.pdf.
[30] Genetic Information Nondiscrimination Act (GINA) of 2008. See http://www.genome.gov/24519851.
[31] NIH considers individual level human genomic data to be "identifiable, sensitive information." For more information, see https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-109.html.

insurability, or reputation, may request a Certificate of Confidentiality from the NIH if they have not already been issued a Certificate of Confidentiality by NIH. [32]

Effective October 1, 2017, all research that was commenced or ongoing on or after December 13, 2016 and is funded by NIH is automatically issued a Certificate of Confidentiality by NIH. Additionally, NIH has issued a Certificate of Confidentiality for dbGaP, which protects data stored in dbGaP and all copies of that data. NIH also encourages investigators and institutions submitting or accessing large-scale human genomic datasets in NIH-designated data repositories to seek a Certificate of Confidentiality as an additional measure to prevent compelled disclosure of any personally identifiable information they may hold if they have not already been issued a Certificate of Confidentiality by NIH.

C. Assuring Appropriate Secondary Use of Genomic and Phenotypic Data

NIH has established policies for the oversight of NIH-designated data repositories and for monitoring the secondary use of controlled-access genomic and phenotypic data, to protect the privacy of research participants and the confidentiality of their data. Qualified investigators, both domestic and foreign, are eligible to request controlled-access data in NIH-designated data repositories through the submission of a DAR that includes a brief description of the proposed research use and an attestation to comply with the DUC Agreement[20], Genomic Data User Code of Conduct[21], and the *NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy*.[22] Requests for data must be approved by an investigator's institution before a review by NIH Data Access Committees (DACs). Decisions to grant access are made based on whether the request conforms to the NIH GDS Policy and program specific requirements or procedures (if any). In particular, all data uses proposed for NIH genomic data must be consistent with the DULs proscribed for the dataset by the submitting institution and identified on the public website for NIH-designated data repository. NIH DACs consist of federal employees with expertise in bioethics, privacy, data security, and appropriate scientific and clinical disciplines. Consultants with specific expertise may be invited to meetings or to provide written consultation.

Only after approval by the relevant NIH DAC will data be available to investigators in an encrypted format using secure file transfer technology. The governance and oversight structure for NIH-designated data repositories and for monitoring genomic data use are further explained in the NIH GDS Policy and on the NIH Office of Science Policy website.[33]

1. Protections for research participants

Investigators and institutions seeking access to data from an NIH-designated data repository agree to a Genomic Data Use Code of Conduct[20] and the DUC Agreement[19] that stipulates the terms and conditions of data access including a number of protections relating to the security and use of research participant data. Both the DAR and the DUC Agreement must be co-signed by the investigator and the responsible Institutional Official to document their joint agreement to follow NIH policy for the use of data accessed through NIH-designated data repositories.

Investigators approved to use controlled-access data are expected to protect the confidentiality of the data by following best practices for data security in addition to any other dataset-specific recommendations as detailed for a given genomic research study. Annual progress updates on

[32] For additional information about Certificates of Confidentiality, see http://grants.nih.gov/grants/policy/coc/.
[33] GDS Policy Oversight. See https://osp.od.nih.gov/scientific-sharing/policy-oversight/.

data use will be reviewed by the relevant NIH DAC to verify continued appropriate use of the data.

The NIH GDS Policy prohibits investigators who download unrestricted-access data from NIH-designated data repositories from attempting to identify individual human research participants from whom the data were obtained, and are expected to acknowledge in all oral or written presentations, disclosures, or publications the specific dataset(s) or applicable accession number(s) and the NIH-designated data repositories through which the investigator accessed the data.

Submitting investigators and their institutions may request removal of data on individual participants from NIH-designated data repositories in the event that a research participant withdraws from the study or part of the study, or does not wish their individual data to be included within the data available for sharing. However, data that have already been distributed to Approved Users for research will not be able to be retrieved.

NIH expects that any users of NIH genomic summary results (GSR) will: 1) complete the review of a responsible genomic data use informational module prior to accessing the information; 2) not use GSR to re-identify individuals or generate information that could allow participant's identities to be readily ascertained; and 3) use GSR to promote scientific research or health.

3. Return of individual research results

The return of individual research results to participants from secondary studies is expected to be a rare occurrence as neither investigators who access data nor the data repository will have access to the identities of participants. Moreover, secondary research using data in repositories is rarely expected to have immediate implications for the health of individual participants.

If a secondary investigator does generate potentially clinically actionable results of immediate clinical significance, he or she can only facilitate their return by contacting the investigator who originally submitted the data and holds the original key to the code that identifies the participants. In such cases, the submitting investigator would be expected to comply with all applicable laws and regulations and consider the benefits and risks associated with the return of individual research results to participants and follow established institutional procedures (e.g., consultation with and approval by the IRB) to determine whether return of the results is appropriate and, if so, how it should be accomplished.