Genome-Wide Association Studies (GWAS)

NIH Points to Consider

NIH Points to Consider for
IRBs and Institutions in their Review of Data Submission Plans for Institutional Certifications Under NIH's Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies
(GWAS)

INTRODUCTION

Under the National Institute of Health (NIH) Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS) (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html), institutions are responsible for certifying that plans for the submission of genotype and phenotype data from GWAS to the NIH meet the expectations of the policy. The purpose of this document is to assist Institutional Review Boards (IRBs) and/or, as appropriate, Privacy Boards in their review, and institutions in their certification, of investigator applications and proposals involving the submission of GWAS data to the NIH under this policy.[1]

This information is being provided in two parts: Part I provides information about the:

a) policy;
b) benefits of broad sharing of GWAS data through a central data repository at NIH;
c) risks associated with the submission and subsequent sharing of such data; and
d) safeguards that will be in place at NIH to protect the data.

Part II is intended to provide specific points to consider for institutions and IRBs in their review and certification of an investigator's plans for submission of data to the GWAS data repository, including the adequacy of consent forms for data submission. The NIH recognizes the complex and evolving nature of the ethical issues related to this policy and will issue additional guidance as may be needed on the GWAS website at http://grants.nih.gov/grants/gwas/index.htm.

[1] The NIH recognizes that this review and certification process goes beyond regulatory requirements under 45 CFR part 46 as outlined in an August 2004 policy guidance of the Office for Human Research Protections entitled 'Guidance on Research Involving Coded Private Information and or Biological Specimens.' Following discussions with NIH staff, OHRP advised NIH that the GWAS repository does not currently involve human subjects research because the data being submitted will be collected solely for other research studies, and because the data will be coded and the identity of individuals from whom the data were obtained will not be readily ascertainable to the investigators maintaining the repository. This determination also means that IRB review and approval of the submission of GWAS data to dbGaP is not required under the regulations. Nonetheless, for the reasons outlined in this document, NIH, as a policy matter, will not accept data into the GWAS repository without the appropriate certifications from the institution and verification by an IRB and/or Privacy Board that the submission criteria stipulated in the policy have been met.

This document will be updated when the *Draft NIH Policy for Genomic Data Sharing* is finalized.

INSTITUTIONAL CERTIFICATION

The NIH will accept GWAS data into the NIH GWAS data repository after receiving appropriate certification by the responsible Institutional Official(s) of the submitting institution that they approve submission to the NIH GWAS data repository. The certification should assure that:

- The data submission is consistent with all applicable laws and regulations, as well as institutional policies;
- The appropriate research uses of the data and the uses that are specifically excluded by the informed consent documents are delineated;
- The identities of research participants will not be disclosed to the NIH GWAS data repository; and
- An IRB and/or Privacy Board, as applicable, reviewed and verified that:
  - The submission of data to the NIH GWAS data repository and subsequent sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;
  - The investigator's plan for de-identifying datasets is consistent with the standards outlined in the policy;
  - It has considered the risks to individuals, their families, and groups or populations associated with data submitted to the NIH GWAS data repository; and
  - The genotype and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46.

PART I: BACKGROUND INFORMATION

A. NIH Policy for Sharing of GWAS Data

1. Data types to be shared through the GWAS Policy

The NIH GWAS policy facilitates the sharing of large datasets containing coded,[2] de-identified[3] genotype and phenotype data obtained in NIH supported or conducted research. The GWAS policy applies to data obtained prospectively as well as to studies using existing specimens and phenotype data. A key element of the NIH GWAS policy is the expectation that data from NIH-supported GWAS be deposited into the NIH GWAS data repository, currently designated as the **d**ata**b**ase of **G**enotypes **a**nd **P**henotypes (dbGaP), at the National Center for Biotechnology Information (NCBI), a component of the National Library of Medicine (NLM). The data submitted for inclusion in the NIH GWAS data repository will be coded and de-identified by the submitting investigator, but the investigator may retain the key to the code that would link to specific individuals. NCBI will never receive the code or any other information that would enable the identification of the individuals who are the source of the data.

The genotype data will likely consist of the measurement of a large number of single nucleotide polymorphisms (SNPs). Presently GWAS are assessing 300,000 to 1 million SNPs within each individual sample. The SNP pattern provides information unique to an individual (except identical siblings). Because SNP patterns vary among ethnic groups, determination of the likely ethnicity of a participant is possible and comparisons of individual SNP patterns can enable the recognition of family relationships. Identification of a specific individual through GWAS data in the NIH GWAS data repository will require comparison with a SNP pattern from another identifiable DNA sample from the same person. It is anticipated that technological and analytical capacity available to the public is likely to enhance the feasibility of SNP pattern identification in the future.

[2] *Coded* means that any identifying information (such as name or social security number) that would enable the investigator to readily ascertain the identity of the individual to whom the private information or specimens pertain 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. From http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.htm

[3] *De-identified*, for purposes of this document, means that the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 CFR 46.102(f)), the 18 identifiers enumerated at section 164.514(b)(2) of the HIPAA Privacy Rule are removed and the submitting institution has no actual knowledge that the remaining information could be used alone or in combination with other information to identify the subject of the data.

This document will be updated when the *Draft NIH Policy for Genomic Data Sharing* is finalized.

The phenotype data deposited in the NIH GWAS data repository may include information about disease status and characteristics that are not individually identifiable; however, some characteristics may be shared within families or common among population subgroups.

2. Essential role of Institutional Officials and IRBs and/or Privacy Boards in implementation of the policy

The nature of GWAS information about participants and the broad data distribution goals of the NIH GWAS data repository highlight the importance of IRBs and institutions in reviewing plans for data submission, as well as the adequacy of the informed consent process and documents through which the data were obtained. Because the genotype and phenotype information generated about individuals will be substantial and, in some instances, sensitive (such as 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 must be protected.

In order to minimize risks to study participants, data submitted to the NIH GWAS data repository will be de-identified and coded using a random, unique code. Data should be de-identified according to the following criteria:

the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 C.F.R. 46.102(f)); the 18 identifiers enumerated at section 45 C.F.R. 164.514(b)(2) (the HIPAA Privacy Rule) are removed; and the submitting institution has no actual knowledge that the remaining information could be used alone or in combination with other information to identify the subject of the data.

Institutional Officials of the submitting institution and IRBs and/or Privacy Boards play a key role in making sure that the submission of data to the NIH GWAS data repository is consistent with the NIH GWAS policy. The NIH will only accept GWAS data into the NIH GWAS data repository after receiving appropriate certification by the responsible Institutional Official(s) of the submitting institution that they approve submission to the NIH GWAS data repository.

The certification should assure that:

- The data submission is consistent with all applicable laws and regulations[4] as well as institutional policies;
- The appropriate research uses of the data and the uses that are specifically excluded by the informed consent documents are delineated;[5]
- The identities of research participants will not be disclosed to the NIH GWAS data repository; and
- An IRB and/or Privacy Board, as applicable, reviewed and verified that:

[4] Applicable federal regulations may include HHS human subjects protection regulations (45 CFR Part 46), FDA human subjects protection regulations (21 CFR Parts 50 and 56), and the Health Insurance Portability and Accountability Act Privacy Rule (45 CFR Part 160 and Part 164, Subparts A and E).

[5] Applicable federal regulations may include HHS human subjects protection regulations (45 CFR Part 46), FDA human subjects protection regulations (21 CFR Parts 50 and 56), and the Health Insurance Portability and Accountability Act Privacy Rule (45 CFR Part 160 and Part 164, Subparts A and E).

- The submission of data to the NIH GWAS data repository and subsequent sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;
- The investigator's plan for de-identifying datasets is consistent with the standards outlined in the policy;
- It has considered the risks to individuals, their families, and groups or populations associated with data submitted to the NIH GWAS data repository; and
- The genotype and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46.

B. Benefits of Broad Sharing of GWAS Data through an NIH Central Data Repository

1. Nature of genome-wide association studies (GWAS)

The scientific aim of GWAS is to discover genetic factors that contribute to the development, progression, or treatment options for a particular disease or trait (such as high blood pressure or obesity). GWAS are particularly powerful for the study of common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses, where the individual genetic contributions to the disease are expected to be relatively weak. When combined with clinical and other phenotypic data, analysis of whole genome information offers the potential for increased understanding of basic biological processes affecting human health and improvement in the prediction of disease and treatment options.

GWAS are possible because of the development of new technologies that can quickly and accurately analyze whole-genome samples for SNPs. To carry out a GWAS, researchers study large groups of individuals, some of whom have the disease being studied and some with similar characteristics who do not have the disease. Each person's entire genome is scanned to identify the specific SNPs at an appropriate number of marker sites along the chromosomes (depending on the population being studied, this can range from about 375,000 to 1 million markers). If certain genetic variations are statistically found to be more frequent in people with the disease than in people without the disease, the variations are said to be "associated" with the disease. The associated genetic variations can serve as guides to the region of the human genome where the genetic contributor to a disease may reside.

2. Reasons for making data accessible to multiple investigators

The NIH is promoting and facilitating the sharing of data generated by individual GWAS because the volume of data that will be generated in even one study is far greater than any individual or small group of collaborators can fully explore and because of the potential to gain important scientific knowledge and tools through the analysis of aggregated GWAS data. The NIH GWAS data repository enhances the NIH's capacity to make GWAS data available to a wide range of scientific investigators in order to facilitate genetic research and enable research discoveries for the benefit of the public health.

GWAS are most informative when the study population is large. The larger the population, the greater the statistical power to determine that observed associations are real and not due to chance. Although the costs associated with whole genome analysis have been decreasing and are expected to continue to decline over time, the costs in terms of research resources (in terms of participant samples and funding) are high because of the large number of study samples required to produce high quality data. The very nature of GWAS allow the data to be used to address multiple research hypotheses. Given the resources involved and the potential

for public benefit, it is prudent to create a database that facilitates the use of these data to address as many hypotheses as are ethically appropriate.

C. Risks Associated with Submission and Broad Sharing of GWAS Data

The main concerns associated with submitting data to the NIH GWAS data repository are those entailed with other genetic research, i.e., those relating to participant privacy and confidentiality. Privacy and confidentiality concerns associated with wide data sharing through the NIH GWAS data repository stem from the nature and magnitude of the genotype and phenotype data involved; the storage of those data in a central, Federal government repository; and the distribution of these data for secondary research. Described below are risks associated with submission and broad sharing of GWAS data. Section D describes measures to minimize such risks. As in the review of any research, it is important to consider any risks in the context of the protections in place to minimize those risks as well as in the context of the expected benefits of the proposed research.

Risks of Identification. The NIH GWAS database will NOT contain information that is typically used to identify individuals such as name, address, telephone number, birth date or social security number. Although the NIH-held data will be coded and the NIH will not hold direct identifiers to individuals whose data are included within the NIH GWAS data repository, the agency recognizes the personal and potentially sensitive nature of the genotype-phenotype data. Additionally, technologies available within the public domain today, and technological advances expected over the next few years, make the identification of specific individuals from raw genotype-phenotype data feasible and increasingly straightforward. For example, someone might be able to compare information in the GWAS database with genotype or phenotype information obtained from other, unrelated activities and be able to identify the individual who is the source of the data (or a blood relative of that individual). If data come from a discrete population (e.g., one small community), it could be more straightforward to cross classify individuals on several variables and make inferences about the source of a given sample.

In addition, discussions are occurring in the scientific community and among privacy experts about the uniqueness of individual genome-wide data and the possibility that in the future such data may by itself become identifiable. See NHGRI's Workshop on Privacy, Confidentiality and Identifiability in Genomic Research (http://www.genome.gov/19519197 and as further discussed in Science, 3 Aug. 2007, vol. 317, p. 600).

The NIH is committed to the protection of research participant privacy and the preservation of the confidentiality of individual-level data submitted to the NIH GWAS data repository. The NIH is, therefore, implementing a number of measures to protect the confidentiality and security of all data submitted to the NIH GWAS data repository (see below). However, as in any system of protections, there are limitations to the protections afforded by these measures.

Risks Associated with Inadvertent or Inappropriate Use or Disclosure of Individually Identifiable Information. The NIH GWAS data repository will not contain individually identifiable information and, therefore, such data cannot be released to secondary users. However, the primary GWAS study may involve individually identifiable information. Submitting institutions should understand that potential harms to research participants or their family members can occur if individually identifiable information is inadvertently or inappropriately used or disclosed. These harms could include denial of employment or insurance of a research participant (or a

relative). Other harms that may occur from inadvertent or inappropriate disclosure or use of individually identifiable information include psychosocial harms, such as stress, anxiety, stigmatization, or embarrassment resulting from inadvertent disclosure of information on family relationships, ethnic heritage, or potentially stigmatizing conditions.

Risks Associated with FOIA. The datasets submitted to the NIH will be maintained in an NIH data repository and will, thereby, become U.S. government records that are subject to the Federal Freedom of Information Act (FOIA). As an agency of the Federal government, the NIH is required to release government records in response to requests under the federal Freedom of Information Act (FOIA), unless the records are exempt from release under one of the FOIA exemptions. The NIH believes that release of unredacted GWAS datasets in response to a FOIA request would constitute an unreasonable invasion of personal privacy under FOIA Exemption 6, 5 U.S.C. § 552 (b)(6). Therefore, among the safeguards that the NIH foresees using to preserve the privacy of research participants and confidentiality of genomic data in the NIH GWAS data repository is the redaction of individual-level genotype and phenotype data from any disclosures made in response to FOIA requests and the denial of requests for unredacted datasets. It is important to note, however, that FOIA affords requesters an opportunity to contest an agency's determination.

Risks Associated with Law Enforcement Access. The NIH will not possess direct identifiers within the NIH GWAS data repository, nor will the NIH have access to the link between the data keycode and the identifiable information that may reside with the primary investigators and institutions for particular studies. However, it is conceivable that law enforcement agencies could request access to the de-identified genotype and phenotype data within the NIH GWAS data repository and, for example, search for matches to DNA specimens collected for forensic purposes.[6] While expected to be rare, such requests may be fulfilled by the NIH. Law enforcement officials might then seek to compel disclosure of identifying information from the institution holding the identifying information. However, the release of identifiable information from the institution holding the identifying information may be protected from compelled disclosure if a Certificate of Confidentiality is or was obtained for the original study.

Risks to Specific Populations, Groups, and Communities. Medical research has already shown that some populations demonstrate a higher predisposition to develop certain medical diseases or disorders than others. GWAS will provide insight into how certain variants contribute to health and disease and will also increase knowledge of how genetic variants differ in frequency between and among populations. Genetic variants associated with physical disorders, diseases, and behavioral traits are expected to be found. Genetic associations are determined by statistical frequencies and causative variants will be found in all populations with differing frequencies. Higher or lower frequencies that contribute to observed health patterns, particularly those that tend to be viewed negatively, can lead to genetic stereotypes that can stigmatize all members of a population group whether they possess a given genetic variant or not. In the absence of genetic non-discrimination laws, such information may also affect the insurability or employability of populations or groups. Persons sharing ethnic heritage may similarly be affected by results obtained from sharing of GWAS data.

Return of Individual Research Results. For reasons explained later in this document, the return of individual research results to participants from secondary GWAS is expected to be a rare occurrence. Nevertheless, as in all research, the return of individual research results to participants must be carefully considered because

[6] Law enforcement officials routinely obtain DNA specimens as part of their investigative work and collect DNA from convicted offenders. Every state has established a DNA database, and these databases are linked through the Federal Combined DNA Index System (CODIS) program.

the information can have a psychological impact (e.g., stress and anxiety) and implications for the participant's health and well-being. While clinically valid and meaningful results may have a positive impact on an individual's health, harms can occur if unvalidated research results are provided back to participants or used for medical decision-making. The ethical protections for GWAS data that have been developed to address these and other issues are discussed in the next section.

D. Ethical Protections for the GWAS Data

The NIH has developed policies and procedures to promote the ethical conduct of GWAS research and to minimize risks to research participants. The NIH acknowledges that the practical and ethical questions relevant to the NIH GWAS Policy are the subject of considerable discussion in the research community. The NIH remains committed to participating in the on-going dialog on these topics and to addressing the evolving scientific, ethical and societal issues within the NIH GWAS policy and practices as appropriate.

Operating Policies. NIH is establishing policies and procedures for the NIH GWAS data repository that address, among other matters, the privacy of GWAS research participants and confidentiality of their data, the interests of participants, families and groups, data access procedures, and data security mechanisms. They will be reviewed periodically and updated as necessary by the GWAS oversight bodies discussed later in this document.

De-identification of Data. Before data are submitted to NIH, submitting investigators will be expected to de-identify the data according to the following criteria: 1) the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 CFR § 46.102(f)); and 2) the following identifiers enumerated at section 164.514(b) (2) of the HIPAA Privacy Rule are removed:

1. Names.
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census: a. The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people. b. The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Telephone numbers.
5. Facsimile numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.

9. Health plan beneficiary numbers.
10. Account numbers.
11. Certificate/license numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification

In addition, the submitting institution should have no actual knowledge that the remaining information could be used alone or in combination with other information to identify the individuals who are the subject of the information. In reviewing data submission plans, the relevant IRB and/or Privacy Board should consider the extent to which the genotype and other phenotype information associated with the participants could be used to identify an individual or his or her family members by matching the genotype/phenotype datasets to other sources of information.

Coding of Data. Before data are submitted to the NIH GWAS data repository, submitting investigators will be expected to assign a random, unique code to the data to protect participant privacy and confidentiality. As a further protection, submission of GWAS data must be accompanied by a written certification by the submitting institution stating that the identities of research participants will not be disclosed to the NIH GWAS data repository.

Certificates of Confidentiality. Prior to submitting GWAS data to the NIH GWAS data repository, investigators and their IRBs may want to determine whether a Certificate of Confidentiality has been obtained for their research or, if one has not been obtained, to consider whether or not it would be appropriate to do so. Certificates of Confidentiality may provide an additional safeguard with regard to compelled disclosure in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level, of information that could be used to identify individual research participants. Certificates of Confidentiality are issued to help achieve research objectives and promote participation in research. They can be granted for studies collecting genetic and other information that, if disclosed, could have adverse consequences for participants or damage their financial standing, employability, insurability, or reputation. Further information on when Certificates of Confidentiality may be appropriate and application instructions, can be obtained at the NIH Certificate of Confidentiality kiosk: http://grants2.nih.gov/grants/policy/coc/

NIH GWAS Data Repository Security Measures. To secure the data, the NIH GWAS data repository will include multiple tiers of data security such as sequential firewalls, independent networks, and encryption based on the content and level of risk associated with the data. All data and information will be submitted to a high

security network within NIH through a secure transmission process. Details on security measures can be found on the NCBI website, http://www.ncbi.nlm.nih.gov.

Controlled Access to Individual Data. Access to individual-level genotype and phenotype data will be tightly controlled. Individual genotype and phenotype data will only be available for research through a controlled access procedure. Only basic descriptive information about each GWAS study, such as the measures that it used, and the composition of the study population will be publicly available. Selected aggregate statistical calculations[7] will also be made publicly available.

Assuring Appropriate Data Use. Researchers eligible for access to individual-level data include, but are not limited to, qualified investigators from academic institutions and commercial organizations, both domestic and foreign. Researchers will have to apply for access to data included in the NIH GWAS data repository through the submission of a Data Access Request that will include a brief description of the proposed research use. Requests will be approved by a researcher's home institution and then routed to an NIH Data Access Committee (DAC). Each funding Institute and Center at NIH may choose to form a DAC to review incoming requests for datasets within their purview, but consistency and transparency of the GWAS data access process will be assured through common principles and operational mechanisms. A DAC consists of federal staff with expertise in relevant scientific disciplines and ethical issues related to protecting the privacy of research participants and the confidentiality of their data. Outside experts may be consulted as necessary. DACs review requests for access to determine that the proposed use of a dataset is scientifically and ethically appropriate and does not conflict with any constraints or informed consent limitations identified by the submitting institution. If a data request raises concerns related to privacy and confidentiality, risks to populations or groups, or other concerns, the relevant DAC may consult with other experts as appropriate. Only after approval by the relevant DAC will data be available for download in a secure and encrypted format by a recipient investigator.

Investigators and institutions seeking data from the NIH GWAS data repository will submit to the NIH a Data Access Request along with a Data Use Certification that will stipulate a number of protections for research participants. Both the Data Access Request and the Data Use Certification must be co-signed by the investigator and by the appropriate designated Institutional Official to document their joint agreement to follow NIH policy for the use of GWAS data obtained from the NIH GWAS data repository. The Data Use Certification will stipulate that, subject to applicable law, the investigator and institution will:

- Use the data only for the approved research;
- Protect data confidentiality;
- Follow appropriate data security protections;
- Follow all applicable laws, regulations and local institutional policies and procedures for handling GWAS data;
- Not attempt to identify individual participants from whom data within a dataset were obtained;
- Not sell any of the data elements from datasets obtained from the NIH GWAS data repository;

[7] The particular calculations for a given dataset may vary by program (based on data content and program direction).

- Not share with individuals other than those listed in the request any of the data elements from datasets obtained from the NIH GWAS data repository;
- Agree to the listing of a summary of approved research uses within the NIH GWAS data repository along with his or her name and organizational affiliation;
- Agree to report violations of the GWAS policy to the appropriate DAC;
- Acknowledge the GWAS policy with regard to publication and intellectual property; and
- Provide annual progress reports on research using the GWAS dataset.

The recipient investigator will be expected to protect the data by following best practices for data security posted on the NIH GWAS data repository website at http://www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf, or other dataset-specific recommendations as detailed for a given GWAS within the repository. In addition, progress reports will be reviewed by the relevant DAC to verify continued appropriate use of the data.

Withdrawal of Consent. The NIH GWAS data repository has developed policies with regard to removal of individual data records if consent is withdrawn. Submitting investigators and their institutions may request removal of data on individual participants from the data repository in the event that a research participant withdraws consent. However, data that have already been distributed for approved research use will not be able to be retrieved.

Return of Research Results. The NIH anticipates that GWAS will generate an unprecedented number of associations between particular genetic loci and diseases, or conditions or treatments. These associations constitute the first step in a multistep process between uncovering the mechanism of action of a genetic locus and developing therapies or diagnostics that can be used in patient care. Initial findings will need to be confirmed and validated by further research before their potential clinical significance is understood. In addition, many statistical challenges in this area of research must be overcome in order to avoid false positive or false negative results. As in any research, harms may result if individual research findings that have not been clinically validated are returned to subjects or are used for clinical decision-making prematurely.

The return of individual findings from studies using data obtained from the NIH GWAS data repository is expected to be rare, because secondary investigators will not be able to return individual research results directly to a participant as neither they nor the NIH GWAS data repository will have access to the identities of participants. Moreover, data obtained from secondary studies are not expected to have immediate implications for the health of individual participants for the reasons mentioned above. If a secondary investigator does generate clinically valid results of immediate clinical significance, he or she can only facilitate their return by contacting the contributing investigator who holds the key to the code that identifies the participants. In such cases, the contributing 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.

If they have not already done so, contributing institutions and their IRBs may wish to establish policies for determining when it is appropriate to return individual findings from research studies.

Oversight of GWAS Activities. The NIH has established policies for oversight of the NIH GWAS data repository and for monitoring GWAS data use practices. They include a review process for NIH GWAS activities to ensure ongoing, high-level NIH oversight and regular input from public representatives, including those with expertise in bioethics, privacy, data security, and appropriate scientific and clinical disciplines.

The governance and oversight structure for the NIH GWAS data repository and for monitoring GWAS data use practices provides oversight tailored to the specific role involved and involves a number of oversight committees. One of these groups, the Research Participant Protection and Data Management Steering Committee, will include among its members the chairs of all Data Access Committees at the NIH as well as appropriate staff from NIH policy and oversight offices (e.g., the Office of Science Policy and the Office of Human Subjects Research). This committee will work to promote consistent and robust participant protections across relevant NIH programs. In order to maintain GWAS policy consistent with evolving technological and ethical considerations, the NIH Director will solicit recommendations on the policy from external experts representing public and scientific stakeholders through the Advisory Committee to the Director. The governance and oversight mechanisms for GWAS activities are further explained in the GWAS Policy.

This document will be updated when the *Draft NIH Policy for Genomic Data Sharing* is finalized.

PART II: DATA SHARING PLANS, INSTITUTIONAL CERTIFICATION, AND POINTS TO CONSIDER REGARDING INFORMED CONSENT

A. Data Sharing Plans

Competing GWAS applications will be expected to include a GWAS data sharing plan as part of the Research Plan, or to provide an appropriate explanation as to why submission to the NIH GWAS repository is not possible. Data sharing plans are expected to describe how the expectations of the policy will be met, including the consistency of the informed consent for submission to the NIH GWAS data repository and subsequent sharing, how informed consent will be obtained (for prospectively collected samples and data), and how data will be subsequently de-identified in accord with the specific criteria for data submission. IRBs should be cognizant of the GWAS data sharing plans at the time of IRB review of the application in order to assess their appropriateness for a specific dataset and to provide the relevant analysis called for within the policy under the institutional certification expectations.

B. Institutional Certification

Institutions submitting GWAS data to the NIH GWAS data repository are responsible for certifying that data submission plans meet the following expectations defined in the GWAS policy:

- The data submission is consistent with all applicable laws and regulations[8] as well as institutional policies;
- The appropriate research uses of the data and the uses that are specifically excluded by the informed consent documents are delineated;
- The identities of research participants will not be disclosed to the NIH GWAS data repository; and
- An IRB and/or Privacy Board, as applicable, reviewed and verified that:
    - The submission of data to the NIH GWAS data repository and subsequent sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;
    - The investigator's plan for de-identifying datasets is consistent with the standards outlined in the policy;
    - It has considered the risks to individuals, their families, and groups or populations associated with data submitted to the NIH GWAS data repository; and
    - The genotype and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46.

[8] Applicable federal regulations may include HHS human subjects regulations (45 CFR Part 46), FDA human subjects regulations (21 CFR Parts 50 and 56), and the Health Insurance Portability and Accountability Act Privacy Rule (45 CFR Part 160 and Part 164, Subparts A and E).

This document will be updated when the *Draft NIH Policy for Genomic Data Sharing* is finalized.

C. Points to Consider Regarding Informed Consent

The NIH recognizes that the issues related to determining the appropriateness of informed consent for submission of data to the NIH GWAS data repository and subsequent sharing for research are quite complex. The GWAS policy applies to genome-wide association research utilizing genetic materials and data collected both prospectively and retrospectively and the applicable considerations regarding informed consent may vary depending upon which type of study is being proposed.

Prospective Studies. For prospective studies, in which GWAS are included within the study design at the time research participants provide their consent, the consent form and process must comply with the requirements of 45 C.F.R. Part 46 and any other applicable law. From an ethical standpoint, the informed consent process and document should make it clear that participants' DNA will undergo genome-wide analysis and that genotype and phenotype data will be shared for research purposes through the NIH GWAS data repository.

Retrospective Studies. For retrospective studies performed using existing genetic materials and previously collected data, the NIH anticipates considerable variation in the extent to which future genetic research and data sharing have been addressed within the informed consent documents. In all such cases, IRBs are expected to determine whether the initial consent under which existing genetic materials and data were obtained is consistent with the submission of data to the NIH GWAS repository and the sharing of that data in accord with the GWAS policy.

The NIH anticipates that for studies that propose to use pre-existing data or samples, IRBs may conclude in some cases that the original consent is not adequate for submission to the GWAS data repository and subsequent sharing for research. In these cases, the IRB may decide that it is appropriate and necessary for the investigator to seek explicit consent of the research participants for submission to the NIH GWAS repository and subsequent sharing. Programmatic consideration to requests from investigators for funding to support efforts to seek re-consent from participants will be provided on a case-by-case basis. It should be noted that the criteria for a waiver of consent under 45 CFR part 46 are inapplicable to such IRB considerations since the GWAS database does not currently involve human subjects research. The criteria that are expected to be applied in making the determination that submission is consistent with the consent are set forth in the GWAS policy and explained in this document.

The IRB also may determine that re-consent is not feasible or appropriate for a given study. Moreover, the IRB may determine that it cannot verify that the other three criteria described in the policy[9] have been met for submission to the NIH GWAS repository. In all these cases, the researcher's data sharing plan should explain the IRB's determination that submission to the GWAS repository is not appropriate. NIH Institutes and Centers will consider these issues on a case-by-case basis when making programmatic decisions to fund GWAS studies for which the submission criteria cannot be met.

---

[9] As outlined elsewhere in this document, in addition to verifying that submission to the GWAS repository and subsequent sharing for research purposes is consistent with the informed consent of study participants from whom the data were obtained, the IRB is also expected to verify that:

- The investigator's plan for de-identifying datasets is consistent with the standards outlined in the policy;
- It has considered the risks to individuals, their families, and groups or populations associated with data submitted to the NIH GWAS data repository; and
- The genotype and phenotype data to be submitted were collected in a manner consistent with 45 C.F.R. Part 46.

This document will be updated when the *Draft NIH Policy for Genomic Data Sharing* is finalized.

The following points to consider may be helpful to IRBs in determining the consistency of existing consents with the NIH GWAS data sharing policy, as well as to investigators in preparing new consent documents for this purpose. They are not intended to be proscriptive, nor are they all of the issues that may be appropriate for IRBs to consider in specific scenarios. Each research project and consent document is unique and local IRBs are in the best position to evaluate the potential benefits and risks of data submission and the consistency of consent with submission to the NIH GWAS data repository.

Scope of Written Consent.

Is the informed consent consistent with the anticipated research activities under the NIH GWAS policy? For instance:

- Does the consent form either allow or preclude:
  - *genetic research or analysis?*
  - *future use and broad sharing of the participant's coded phenotype and genotype data for research?*
  - *submission of the participant's coded phenotype and genotype data to a government health research database for broad sharing to qualified investigators?*
- Does the consent form have any restrictions, such as:
  - *types of subsequent research using the participant's phenotype and genotype data?*
  - *location of such research?*
  - *types of medical conditions or diseases studied?*
  - *duration of storage and use of phenotype and genotype data?*
  - *limitations on who can use the participant's phenotype and genotype data (e.g. some consents may state that only non-commercial researchers can use the data)?*

For studies that are found to be acceptable for submission to the NIH GWAS data repository, the certification provided to the NIH should delineate the appropriate research uses of the data and any uses that are specifically excluded by the informed consent documents.

Potential Benefits

Does the consent form discuss that potential benefits may accrue broadly to the public through the advancement of science and understanding of health and disease, rather than resulting in direct benefits to individuals?

Risks

Does the consent form discuss risks associated with genetic or genomic research? Are these risks consistent with the risks involved in GWAS activities? For example:

- *Does the consent form discuss risks of broad sharing of phenotype and genotype data?*
- *Does the consent form discuss privacy risks of data sharing (e.g., the possibility that the coded data may be released to members of the public, insurers, employers, and law enforcement agencies)?*
- *Does the consent form discuss the risks of computer security breaches relevant to maintaining data in an electronic format?*
- *Does the consent form discuss relevant risks to relatives or identifiable populations or groups?*

Return of Research Results

Does the consent form include a discussion of whether or not research results will be returned to subjects, and under what conditions? Are those representations consistent with the GWAS policy that research results may only be returned in rare instances following established procedures at the contributing institutions?

Privacy and Confidentiality Protections

Does the consent form address how individual privacy and data confidentiality will be protected? Is the manner in which privacy and confidentiality measures are described consistent with the NIH GWAS policies?

Withdrawal of Consent

Does the consent form address whether a subject can withdraw his/her phenotype and genotype data from research use? Is this language consistent with GWAS policies?

Commercial Use

Does the consent form allow for or preclude commercial use of the subject's phenotypic and genotypic data? If specific restrictions are specified, they should be included within the institutional certification to the NIH.

Other

Is there any other information in the consent form that is inconsistent with the information provided about the NIH GWAS data repository and the GWAS policies and procedures?

Other Issues to Consider

- Does the study involve children? If so, has the IRB considered the appropriateness of the continued maintenance and sharing of the data when the child reaches the legal age of consent?
- Does the study involve proxy consent? If so, are there any special ethical issues that should be considered?

This document will be updated when the *Draft NIH Policy for Genomic Data Sharing* is finalized.

- Does the study involve vulnerable populations, and if so, have any special ethical concerns related to the study population been addressed?
- Have any special cultural considerations or requirements been addressed with regard to the study population (e.g., the need for tribal consent from Native American populations)?
- Are any issues of group harm relevant and have they been considered?

GLOSSARY

The human **genome** is all the DNA contained in an organism or a cell, including both the DNA comprising chromosomes within the nucleus and the DNA in mitochondria.

A **single-nucleotide polymorphism (SNP)** is a variation in a DNA sequence that results when a single nucleotide (A,T,C,or G) in the genome sequence is replaced by another.

Phenotype data are data on health conditions, behavioral characteristics, or measurable or observable traits (such as blood pressure, alcohol consumption, cholesterol, or eye color) that are obtained during physical or psychological examination and maintained in a medical or research record. Phenotype data may also include information about medical treatments, drug tolerance, and family medical history as well as responses to questionnaires.

A **genome-wide association study (GWAS)** is a study of genetic variation across the entire human genome that is designed to associate genetic variations (SNPs) with traits (such as blood pressure or weight) or with the presence or absence of a disease or condition. This type of study is a comprehensive measurement of all or nearly all variation in all human chromosomes, and sometimes mitochondrial DNA as well. GWAS typically involve hundreds of markers, rather than, for example, studies of candidate genes or targeted chromosomal regions. 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 r2 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.

dbGaP ("**d**ata**b**ase of **G**enotypes **a**nd **P**henotypes") is a central data repository at the National Center for Biotechnology Information (NCBI), a branch of the National Library of Medicine (NLM) at NIH.

NIH Data Access Committees (DAC) are oversight committees composed of federal staff with expertise in a variety of areas such as the relevant scientific disciplines, research participant protection and privacy issues. The DAC will review requests for research access to GWAS data to determine whether the proposed use of the data is scientifically and ethically appropriate and to confirm that the proposed research use does not conflict with any constraints or informed consent limitations identified in institutional certifications.

NIH GWAS data sharing policy ("GWAS policy") is the policy that GWAS data obtained with NIH support should be shared through a central repository when such data sharing is compatible with the consent provided by the participant. It can be found at http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.htm.