Comments of the World Privacy Forum

Regarding

Department of Health and Human Services, National Institute of Health Request for Information: Proposed Policy for sharing of data obtained in NIH-supported or conducted genome-wide association studies (GWAS).

Via email and electronic submission

NIH GWAS RFI Comments
National Institutes of Health
Office of Extramural Research
6705 Rockledge Drive
Room 350
Bethesda, MD 20892-7963

October 29, 2006

Re: NIH Genome Wide Association Studies Request for Information, 71 Federal Register 51626

The World Privacy Forum welcomes the opportunity to respond to the National Institutes of Health (NIH) request for information on the proposed policy for sharing of data obtained in the NIH supported genome-wide association studies (GWAS). 71 Federal Register 51626 (Aug. 30, 2006). In the August 30, 2006 Request for Information (RFI), the NIH defines a genome wide association study as "any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition." The NIH is proposing the development of a central GWAS data repository at the NIH that will "provide a single point of access to basic information about NIH-supported GWAS and to available genotype-phenotype datasets for GWAS." The NIH envisions that access to all NIH-supported GWAS datasets will be possible through this repository.

These comments focus primarily on the privacy implications of the proposed policy. Specific areas analyzed in these comments include issues related to identifiability, secondary uses of the genetic data, oversight, legal protections, and informed consent.

The World Privacy Forum is a non-profit, non-partisan public interest research organization. It focuses on in-depth research and analysis of privacy topics, including topics in medical privacy, financial privacy, and other aspects of privacy.

I. Identifiability

The NIH request for information (RFI) does not adequately describe the degree of identifiability for data in the GWAS repository. The indistinct discussion of the use of a "random, unique code" is insufficient at this stage to allow for a fair analysis for the following reasons.

First, the RFI does not describe the specific coding system for data, and because of this it is impossible to evaluate how coding will protect privacy. It is impossible to assess how the ambitious goals of the GWAS will be achieved with data that is coded in such a way that it cannot be linked to individual patients being followed longitudinally. Without clear, testable standards, institutions providing data may use coding schemes that do not adequately protect identifiers, that do not work at all, or that fail when used repeatedly.

Second, the collection of pedigree data may create additional privacy concerns. Additionally, the degree of protection available through the coding of pedigree data needs to be explained.

Third, the identifiability of genetic data is a moving target. Portions of an individual's genetic code that appear to be non-identifiable today may become identifiable tomorrow as a result of new technologies or other data repositories maintained by other researchers or, more ominously, by law enforcement agencies.

Fourth, as a patient visits different hospitals and participates in one or more research projects, the combined trail of data may support identification efforts that are not possible by looking at individual data repositories alone. Data from projects that may not be identifiable separately may become identifiable when combined. The GWAS repository needs to consider this possibility because databases of genetic data will proliferate.

The NIH needs to explain how the GWAS repository cannot become a source of information for aggressive law enforcement investigators, private litigants, marketers, or others who seek an individual (or multiple individuals) with particular genetic characteristics. The World Privacy Forum recommends that NIH provide a complete explanation of any coding scheme that will be required or used in connection with the GWAS repository. The explanation must permit an independent analysis of the technical adequacy of the coding scheme. Indeed, NIH should fund an independent review of any coding scheme that it uses or requires. NIH should also offer a technical explanation why the GWAS repository – either on its own or in connection with other information resources or technologies -- cannot be used by third parties to identify individuals.

We have another concern about identifiability. Coding identifiable information, if done properly, offers some protection for privacy. However, information that does not have any overt identifiers may nevertheless be capable of reidentification. The work of Carnegie Mellon Professor Latanya Sweeney offers overwhelming evidence on this point. We refer you in particular to B. Malin and L. Sweeney, *How (Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-identification to Evaluate and Design Anonymity Protection Systems*. 37 Journal of Biomedical Informatics 179-192 (2004),

<http://privacy.cs.cmu.edu/dataprivacy/projects/trails/dnaTrails.html>. We quote the paper's abstract here because it makes the point that removal or encryption of explicitly identifiable genetic information is not sufficient.

> The increasing integration of patient-specific genomic data into clinical practice and research raises serious privacy concerns. Various systems have been proposed that protect privacy by removing or encrypting explicitly identifying information, such as name or social security number, into pseudonyms. Though these systems claim to protect identity from being disclosed, they lack formal proofs. In this paper, we study the erosion of privacy when genomic data, either pseudonymous or data believed to be anonymous, is released into a distributed healthcare environment. Several algorithms are introduced, collectively called RE-Identification of Data In Trails (REIDIT), which link genomic data to named individuals in publicly available records by leveraging unique features in patient-location visit patterns. Algorithmic proofs of re-identification are developed and we demonstrate, with experiments on real-world data, that susceptibility to re-identification is neither trivial nor the result of bizarre isolated occurrences. We propose that such techniques can be applied as system tests of privacy protection capabilities.

We still have not exhausted identifiability concerns. Even if genetic sequences in the GWAS repository are not identifiable on their own or in connection with any existing or foreseeable database, law enforcement agencies may nevertheless seek the data. Consider the scenario in which a law enforcement agency is searching for an unknown suspect whose DNA sequence is known. If the agency searches a GWAS repository and finds a match, that would provide a major investigatory lead. The law enforcement agency then needs to go to the source of the data and find a way to obtain the original identifiable data to find the suspect.

The World Privacy Forum recommends that NIH's analysis of the identifiability of GWAS data go beyond the coding of identifiable data elements. All possibilities for identification and reidentification must be considered. A broad and independent review of all actual and possible identifiability issues is essential.

II. A Repository of Genetic Data Is Exposed to Secondary Use

Any collection of personal data, whether overtly or potentially identifiable, will be a magnet for secondary users and secondary uses. A repository of genetic information is no exception. Indeed, we already know that genetic data is of interest to and is actively used by numerous law enforcement agencies in different ways. An excellent article on the subject is by Mark Rothstein and Meghan Talbott: *The Expanding Use of DNA in Law Enforcement: What Role for Privacy*, 34 J.L.Med. & Ethics 153-164 (2005).

As genetic information continues to proliferate in medical, research, and other types of data compilations, law enforcement can be expected to intensify its interest and its demands regarding this type of data. Advances in identification technology will only add to the attractiveness of the data.

Personal information in government files is especially vulnerable to secondary uses. For example, any information held by NIH appears to be disclosable to any component of the Department of Health and Human Services (HHS) pursuant to the provision of the Privacy Act of 1974 that allows disclosure of information to any department employee who has a need for the information in the performance of his or her duties. 5 U.S.C. §552a(b)(1). One HHS agency without a health research function that might have a particular interest in genetic or pedigree data is the Office of Child Support Enforcement. The HHS Office of the Inspector General, with its law enforcement activities, is another.

Federal records are vulnerable in other ways. Privacy Act records can be disclosed for a variety of routine uses. Routine uses applicable to all HHS systems of records include those listed in Appendix B to 45 CFR 5b. The agency-wide routine uses are:

> (1) In the event that a system of records maintained by this agency [stet] or carry out its functions indicates a violation or potential violation of law, whether civil, criminal or regulatory in nature, and whether arising by general statute or particular program statute, or by regulation, rule or order issued pursuant thereto, the relevant records in the system of records may be referred, as a routine use, to the appropriate agency, whether federal, or foreign, charged with the responsibility of investigating or prosecuting such violation or charged with enforcing or implementing the statute, or rule, regulation or order issued pursuant thereto.
>
> (2) Referrals may be made of assignments of research investigators and project monitors to specific research projects to the Smithsonian Institution to contribute to the Smithsonian Science Information Exchange, Inc.
>
> (3) In the event the Department deems it desirable or necessary, in determining whether particular records are required to be disclosed under the Freedom of Information Act, disclosure may be made to the Department of Justice for the purpose of obtaining its advice.
>
> (4) A record from this system of records may be disclosed as a ``routine use'' to a federal, state or local agency maintaining civil, criminal or other relevant enforcement records or other pertinent records, such as current licenses, if necessary to obtain a record relevant to an agency decision concerning the hiring or retention of an employee, the issuance of a security clearance, the letting of a contract, or the issuance of a license, grant or other benefit.
>
> A record from this system of records may be disclosed to a Federal agency, in response to its request, in connection with the hiring or retention of an employee, the issuance of a security clearance, the reporting of an investigation of an employee, the letting of a contract, or the issuance of a license, grant, or other benefit by the requesting agency, to the extent that the record is relevant and necessary to the requesting agency's decision on the matter.
>
> (5) In the event that a system of records maintained by this agency to carry out its function indicates a violation or potential violation of law, whether civil, criminal or regulatory in nature, and whether arising by general statute or particular program statute, or by regulation, rule or order issued pursuant thereto,

the relevant records in the system of records may be referred, as a routine use, to the appropriate agency, whether state or local charged with the responsibility of investigating or prosecuting such violation or charged with enforcing or implementing the statute, or rule, regulation or order issued pursuant thereto.

(6) Where Federal agencies having the power to subpoena other Federal agencies' records, such as the Internal Revenue Service or the Civil Rights Commission, issue a subpoena to the Department for records in this system of records, the Department will make such records available.

(7) Where a contract between a component of the Department and a labor organization recognized under E.O. 11491 provides that the agency will disclose personal records relevant to the organization's mission, records in this system of records may be disclosed to such organization.

(8) Where the appropriate official of the Department, pursuant to the Department's Freedom of Information Regulation determines that it is in the public interest to disclose a record which is otherwise exempt from mandatory disclosure, disclosure may be made from this system of records.

(9) The Department contemplates that it will contract with a private firm for the purpose of collating, analyzing, aggregating or otherwise refining records in this system. Relevant records will be disclosed to such a contractor. The contractor shall be required to maintain Privacy Act safeguards with respect to such records.

(10)-(99) [Reserved]

(100) To the Department of Justice or other appropriate Federal agencies in defending claims against the United States when the claim is based upon an individual's mental or physical condition and is alleged to have arisen because of activities of the Public Health Service in connection with such individual.

(101) To individuals and organizations, deemed qualified by the Secretary to carry out specific research solely for the purpose of carrying out such research.

(102) To organizations deemed qualified by the Secretary to carry out quality assessment, medical audits or utilization review.

(103) Disclosures in the course of employee discipline or competence determination proceedings.

We reproduce the complete list of agency-wide routine uses to emphasize the number and scope of disclosures that are allowed for all HHS Privacy Act data. Subsection (b) of the Privacy Act itself lists additional disclosures -- including some for law enforcement -- that are permitted for personal data in all systems of records. In short, any personal data in a HHS system of records is vulnerable to more than a dozen different types of disclosure.

Several of the HHS-wide routine uses also authorize disclosure to law enforcement with few substantive and procedural protections. Further, routine uses for a specific system of records can authorize additional disclosures. If NIH establishes a system of records for the GWAS repository, it may be able to limit the number of routine uses that allow disclosure of the data. However, it does not appear that NIH would readily be able to disclaim the Appendix B routine uses applicable to all HHS systems of record or the statutory provisions authorizing disclosure of Privacy Act records. Thus, no matter how narrowly NIH defines routine uses for the GWAS system, the records could still be widely disclosed.

The above analysis assumes that the Privacy Act of 1974 will apply to the GWAS repository. If the GWAS information is not maintained in a Privacy Act system of records – and it is not apparent from the RFI whether a system of records is either contemplated or necessary – then the records might be maintained without being subject to any statutory privacy controls at all. If the Privacy Act does not apply, the result may be worse because it would mean that the GWAS repository might be used and disclosed for secondary purposes without even the minimal procedural and substantive limitations imposed by the Act. It would also mean that data subjects might be without any statutory remedies if their data were used in unexpected, improper, or outrageous ways.

The World Privacy Forum observes in passing that NIH is perhaps the only health care provider of any size that is not subject to the HIPAA privacy rule. The HIPAA rule may not be relevant here because the records under the GWAS program may not constitute protected health information (PHI) in the hands of NIH and because the HIPAA rules offer few substantive or procedural protections against law enforcement or national security disclosures. Nevertheless, the HIPAA rule does not apply to NIH because the Secretary of HHS chose not to apply the rule to the NIH. By seeking and maintaining an exemption from HIPAA so that the NIH is not accountable for privacy as are other health care providers, the NIH is not in a position to argue that it particularly trustworthy when it comes to protecting the privacy of health data subjects. Indeed, NIH activities need to be scrutinized for privacy consequences to a greater degree than other federal agencies and private institutions that are subject to the HIPAA privacy rule.

The proposed NIH policy cannot be fully or fairly evaluated without more information about the status of the GWAS repository under the Privacy Act of 1974. The NIH should publish for comment at the earliest possible stage a draft of a Privacy Act system of records notice for the repository. If there is to be no system of records notice, then the NIH should describe what privacy protections – including but not limited to limitations on use and disclosure; remedies for data subjects; and transparency – will apply to records maintained outside the purview of the Privacy Act of 1974.

The World Privacy Forum recommends and requests that the NIH republish the RFI with a full explanation of the applicability of the Privacy Act of 1974 and of other privacy consequences of the GWAS data collection activity. In addition, the NIH should prepare and publish a Privacy Impact Assessment in accordance with the E-Government Act of 2002 and OMB's implementing memorandum (M-03-22). It is important that the public be provided with a draft Privacy Act System of Records Notice and a Privacy Impact Assessment at the earliest possible opportunity, that the public be permitted to comment on privacy documents, and that public comments be considered before plans for the repository proceed further.

III. The GWAS Repository Needs Stronger Protection Against Secondary Use

In the hands of the NIH, the GWAS repository may be highly vulnerable to disclosure for the reasons already discussed. However, a comparable database in the possession of a government grantee or contractor may have a much higher level of protection against compelled disclosure

through a certificate of confidentiality. In general, certificates of confidentiality authorize researchers to resist compulsory legal demands (e.g., subpoenas and court orders) for identifiable research information about individuals. By providing a defense against compelled disclosure, certificates provide a defense against legal obligations to disclose records to law enforcement agencies, private litigants, and others who may have an interest in the records for purposes unrelated to the purpose for which the records were compiled. One statute that establishes a certificate program is 42 U.S.C. § 241.[1] The NIH should be familiar with certificate programs since it administers one.[2] It is surprising and disappointing that the proposed policy does not refer to certificates of confidentiality.

The degree of protection provided by a certificate of confidentiality is uncertain. One deficiency is that a certificate may not protect against *voluntary* disclosures by the record keeper. Other ways of limiting disclosures may also be available. Contracts are among the instruments that might limit the ability of a record keeper to make voluntary disclosures of personal information.

The GWAS repository should be fully protected by a certificate of confidentiality. The World Privacy Forum expresses no opinion whether a government agency can qualify for a certificate of confidentiality under existing certificate programs. If NIH can obtain a confidentiality certificate or equivalent protection for the GWAS repository, then it may be an appropriate custodian. If not, then the repository may be uniquely vulnerable to secondary use in the hands of NIH, and virtually any qualified institution outside of government that obtains a certificate of confidentiality would be a better guardian of the privacy of data subjects. Contractual or equivalent agreements between NIH and an external, non-governmental data repository could also provide an additional level of protection against inappropriate use of data.

We also observe that the data in the possession of NIH will be subject to request under the Freedom of Information Act. While identifiable data may not be disclosable under the Act, the status of data without identifiers or that is not overtly identifiable is less certain.

The World Privacy Forum recommends that a GWAS repository only be maintained by an institution that qualifies for and actually obtains a certificate of confidentiality providing statutory protection against compelled disclosure of data. Any institution maintaining the repository must also formally commit to protecting the data against voluntary disclosures as well. If NIH cannot meet this standard, then it should provide for the maintenance of the repository, if at all, by a non-governmental organization that has the expressed intention, willingness, and wherewithal to aggressively defend the data against all demands for

---

[1] 42 U.S.C. § 241(d) ("The Secretary [of Health and Human Services] may authorize persons engaged in biomedical, behavioral, clinical, or other research (including research on mental health, including research on the use and effect of alcohol and other psychoactive drugs) to protect the privacy of individuals who are the subject of such research by withholding from all persons not connected with the conduct of such research the names or other identifying characteristics of such individuals. Persons so authorized to protect the privacy of such individuals may not be compelled in any Federal, State, or local civil, criminal, administrative, legislative, or other proceedings to identify such individuals."). Other statutes that provide for certificates of confidentiality or the equivalent include: 42 U.S.C. § 242m(d); 42 U.S.C. § 299c-3(c); 42 U.S.C. § 290aa(n); 42 U.S.C. § 3789g(a); 42 U.S.C. § 10604(d); and 44 U.S.C. § 3501 note.

[2] See the NIH's Certificates of Confidentiality Kiosk at <http://grants.nih.gov/grants/policy/coc/index.htm>.

secondary use and disclosure not expressly consistent with the narrowly-defined purpose of the repository.

The World Privacy Forum further recommends that all data contributed to the GWAS repository come from activities that are themselves subject to a certificate of confidentiality. If data in the repository receives legal protection but equivalent data maintained by the provider of the information does not, then those who seek data need only direct their efforts to obtaining the data from the supplier of the data.

Whether certificate of confidentiality programs as presently constituted provide sufficient protection against secondary use of research data is not readily apparent. Some protection is, however, better than none. The World Privacy Forum recommends that NIH also sponsor an independent review of the adequate of certificate programs covering genetic research data. NIH should publish the results of that review, along with any recommendations for administrative actions or additional legislation.

IV. Oversight

The RFI describes standards and oversight mechanisms for ensuring that GWAS data is maintained, used, and disclosed in accordance with the policy that governs the repository. However, it is far from clear how these mechanisms will work.

> Investigators using GWAS data must stipulate that that they will:
>
> • Use the data only for the approved research use;
> • Protect data confidentiality;
> • Follow all applicable laws and any 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 or share any of the data elements from datasets obtained from parties; and
> • Provide annual progress reports on research.

It is not clear what these stipulations mean. What does it mean to "protect data confidentiality"? This phrase is worthless as a privacy standard or as a definition of the obligations of investigators. What are the elements of data confidentiality that must be protected? Must an investigator go to court to fight all subpoenas for research data? If so, must the investigator take each legal dispute to the highest available court? Can data be disclosed in accordance with state or federal laws that are unrelated to approved research use? What if applicable law or institutional policies conflict with the rules of the repository? Can a scientific auditor obtain access to identifiable data? Can an investigator store the data in a laptop, use a university computer, or hire an outside contractor? Must investigators require research associates to sign confidentiality agreements? What security standards apply to the investigator and to the data? Who will review the annual progress reports to make sure that the use of the data is consistent with the rules? What does it mean not to "share" data? What are datasets obtained from parties

and why are the restrictions on selling and sharing only applicable to those datasets and not other data or derivative data?

It is possible to continue listing questions of this type at great length. The point should be clear. The casual listing of vague phrases designed to reassure the world that privacy and other standards will be met is not reassuring at all. Indeed, the lack of specificity only reinforces suspicions that NIH's interest in privacy protections may only be skin deep.

The description of the Data Access Committees is also insufficient to assess whether the committees will have sufficient independence and authority to function effectively. How will these committees be able to confirm that proposed research is consistent with institutional constraints? Accepting stipulations from researchers – or even from institutions – is not likely to be sufficient. In general, committees composed principally of researchers who approve requests for data from other researchers may not offer the best method of protecting patients. The shortcomings of the existing institutional review board process are well known, and any mechanisms established to oversee access to and use of GWAS data must do better.

More attention must also be paid to the protections in place for data contributed to the repository. The proposed policy states that investigators contributing data "should verify that appropriate data security, confidentiality, and privacy measures are in place for the protection of GWAS participants." Like the stipulations for investigators who seek data from GWAS, the standards for data contributors are vague, incomplete, and certain to be interpreted differently by different institutions. The NIH needs to specify in detail precisely what measures are "appropriate" before it will accept data for the repository.

The World Privacy Forum recommends that NIH greatly expand the description of the obligations to protect privacy and security that it intends to impose on both providers and users of the GWAS database. The descriptions must take into account any different obligations of federal researchers who operate under policies and laws that may be overtly inconsistent with the standards. The NIH should fully describe the membership and operations of Data Access Committees. In addition, the NIH should publicly commit sufficient resources to the oversight of the GWAS program, including the use of audits and a regular review of annual reports submitted by investigators. The NIH should also state what sanctions will be applied to investigators and institutions that fail to follow the requirements or that fail to protect the privacy of individuals.

V. Legal Protections

The NIH's intent in protecting individuals is applauded, notwithstanding the current deficiencies in implementation of that intent. Aside from the issues already discussed, an additional type of protection is essential. The goal of privacy protection is to protect individuals. The NIH policy and all implementing documents signed by data suppliers, data users, and others should expressly state that the purpose of the data restrictions is to protect individuals and that individuals are intended third-party beneficiaries of the agreements. The goal is to make it easier for individuals to sue investigators or others who misuse data or undermine the privacy or other interests of data subjects. Without establishing a firm basis for a legal remedy for individuals, the

enforceability of the NIH policy will be uncertain. The Department avoided this issue in the HIPAA health privacy rule by leaving the third-party beneficiary issue unresolved. The NIH should not follow the Department's lead on this issue. By creating the GWAS repository, the NIH is taking actions that expose data subjects to additional risks. NIH should also take all appropriate steps to provide individuals with the ability to use existing remedies when needed. If those entrusted with data violate the standards and individuals are harmed, those individuals should have enforceable legal rights.

The World Privacy Forum recommends that the NIH expressly provide that data subjects are intended third-party beneficiaries of the legal, technical, and administrative protections established for the GWAS repository.

VI. Informed Consent

It is unclear from the policy what role informed consent will play in the contribution of data to the GWAS repository. One part of the policy states that inclusion of data must be consistent with the initial informed consent process of study participants. However, that is not the same as saying that the participants must have affirmatively agreed to the sharing of their data with the GWAS repository. If a study collects patient data under a privacy waiver from an IRB, there may be no consent at all. Given the long-term risks to privacy that arise when a patient's information becomes part of the GWAS repository, no patient data should be included in the repository without the express informed consent of the data subject. This also means that the policy governing GWAS must address what happens to data if the data subject withdraws consent. Even if information is not identifiable in the GWAS repository, the contributing institution will be able to identify it using the coded information and key to that information that it holds. As long as it is possible for a chain of identifiability to be reconstructed across databases, the full rights of the data subject must be preserved.

VII. Conclusion

The World Privacy Forum is grateful for the opportunity to comment on the policy for the GWAS repository. The policy established for GWAS may have widespread influence on other comparable data activities established by private companies and academic institutions. If existing laws do not establish a fair balance of the needs of researchers and the privacy and other interests of data subjects, then one result may be the identification of a need for greater and firmer regulation of genetic databanks.

Respectfully submitted,

Pam Dixon
Executive Director,
World Privacy Forum
www.worldprivacyforum.org