Comments of the World Privacy Forum
to
The National Institutes of Health regarding *Request for Information on Proposed Updates and Long-Term Considerations for the NIH Genomic Data Sharing Policy*, NOT-OD-22-09

Via electronic submission to: https://osp.od.nih.gov/rfi-updating-the-nih-genomic-data-sharing-policy

National Institutes of Health (NIH)
9000 Rockville Pike
Bethesda,
Maryland 20892

28 February 2022

The World Privacy Forum welcomes the opportunity to respond to the National Institutes of Health's (NIH) *Request for Information on Proposed Updates and Long-Term Considerations for the NIH Genomic Data Sharing Policy*, Notice Number NOT-OD-029, November 30, 2021, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-029.html.

The World Privacy Forum (WPF) is a nonprofit, non-partisan 501(c)(3) public interest research group. WPF focuses on multiple aspects of privacy, with health privacy being among our key areas of work. We publish a large body of health privacy information, including guides to HIPAA; reports and FAQs for victims of medical identity theft; and materials on genetic privacy, precision medicine, electronic health records, and more. We testify before Congress and federal agencies, and we regularly submit comments on HIPAA and related regulations. WPF's Executive Director co-chairs a data governance working group at the World Health Organization (WHO) and is co-chair of a UN Statistics data governance working group. For more about our work and our reports, data visualizations, testimony, consumer guides, and comments, see http://www.worldprivacyforum.org.

Regarding the NIH Request for Information regarding its Genomic Data Sharing Policy, we recognize that NIH is taking reasonable approaches to address current issues facing the use and maintenance of genetic information in health research. The World Privacy Forum thinks that

NIH should take a longer-range view of the problems of genetic data sharing because of ongoing developments in information technology that pose new challenges, and legitimate public demands for privacy protections. Going forward, we anticipate that data from many domains will be gathered, synthesized, and utilized in the research context, including the genetic / genomic context.

In an insightful discussion about the new NIH Data Sharing Policy, the prominent bioethicist Dr. Mark A. Rothstein wrote that "Deidentified data, big data, and access to data by researchers not subject to federal research regulations are new informed consent issues."[1] WPF agrees. The new NIH data sharing policy is vitally important to the debate regarding big genomic data (and its complex implications). If the data sharing policy is to be successful, and we hope that it is, we believe that NIH policy will need to incorporate more aspects of the evolving research data ecosystem circa 2022 and address the new issues and challenges in a forward-looking way.

We offer these comments with that longer-range view in mind, and we hope that NIH will do more to anticipate future developments and challenges in its current work, including the impacts of advances in big genomic data [2] as well as issues regarding access to this data, for both research and non-research purposes.

I. Big genomic data and its impacts on deidentification in today's world – and tomorrow's

We understand from the RFI that NIH is well aware of the direction of deidentification and reidentification in the research context. While acknowledging that this is well-trodden ground, we nevertheless note in these comments the emerging and challenging confluence of some new and some known issues, propelled in part by advances in technology, and in part by the impacts of the Covid-19 pandemic.

A. Big genomic data

We are not far from the time when relying on deidentification as a means of privacy protection for genetic data will be impossible as a practical matter. In the past, the dissemination of genetic data was only a trickle by today's standards, and past levels of computing power facilitated deidentification as a powerful tool for privacy protection. However, that time is passing by, and rapidly so in the area regarding genetic / genomic data. Today, increased computing power (particularly machine learning techniques) combined with the wide availability of genomic and non-genomic data sets from the public and private sector has advanced data analysis to the point that deidentification does not have the same utility as it once did. These capacities, combined

---

[1] Mark A. Rothstein, *Informed Consent for Secondary Research Under the New NIH Data Sharing Policy* 3 May, 2021. J.L.Med.&Ethics,49 (3), Forthcoming, http://dx.doi.org/10.2139/ssrn.3838976

[2] "Big genomic data" is considered to be an important forthcoming area of clinical and research work, to the point that new physicians and researchers are being trained in the field. See: C.K. Rubanovich, C. Cheung, J. Mandel et al, *Physician preparedness for big genomic data: A review of genomic medicine education initiatives in the United States*. Human Molecular Genetics, 2018 Aug 1;27 (R2):R250-R258. doi: 10.1093/hmg/ddy170 See also: A. Zimani et al, *Increasing genomic literacy through national genomic projects*, Front Genet. 2021 Aug 12;12:693253. doi: 10.3389/fgene.2021.693253, eCollection 2021.
PMID: 34456970

with the much broader dissemination of genetic data both within and outside of HIPAA-covered entities, has created the availability of "genetic big data."

There are several aspects of genetic big data that are important to consider. Certainly, one aspect is that genetic big data is particularly difficult to deidentify.[3] This is a simple statement on its face, yet behind it lurks an entire sea-change that has led us to what Rothstein has accurately described as new problems.

First, as a simple matter of fact, there are many programs and proposals aimed at amassing and integrating increasing amounts of genomic data for research purposes. Sequence data from trial participants has provided valuable but generally limited phenotypic data in the past. However, this data, combined with data from large biobanks such as FinnGenn[4], the UK Biobank[5], and the Nebraska Biobank[6], among others, provides rich material for genome-wide association studies (GWAS). These efforts are being enabled in new ways by platforms that are specifically designed to aggregate clinical and genomic data from research participants in order to facilitate the sharing of deidentified data with researchers.[7]

Clinical data research stores such as biobanks are not new to NIH or others. However, utilizing these data stores with genetic data held by HIPAA-covered entities (such as hospitals) plays an increasingly important role in genomic research. Many patients already have some genomic data in their electronic health record (EHR). The era when many, if not all, individuals will have complete genomes as part of their health record is not far off.

This brings us to our second point. Even though over time, increasing stores of EHR genomic data have been available, a significant obstacle to utilizing the genomic data from patient records locked into EHRs has been incompatibility between the electronic health records systems often used in hospital settings and those used in research settings. In the past few years, however, there has been substantial progress toward integrating precision medicine efforts into EHRs.

As of late 2021, Epic Systems' "App Orchard Gallery" now includes at least one platform that provides a way to synthesize genetic / genomic data from molecular labs with clinical information from patients' EHRs, creating much easier access to genomic data, and in a dominant EHR platform in the US.[8] New coding efforts are also bridging this gap, for example, the Minimal Common Data Elements initiative (mCODE), which establishes common data

---

[3] See M.A. Rothstein, *Is deidentification sufficient to protect health privacy in research?* American Journal of Bioethics 10, no. 9, 2010: 3-11. See also: B. Malin and L. Sweeney, *How (not) to protect genetic data in a distributed network using trail re-identification to evaluate and design anonymity protection systems.* Journal of Biomedical Informatics 37, No. 3, 2004: 179-192.

[4] FinnGenn, Finland https://www.finngen.fi/en

[5] UK Biobank, https://www.ukbiobank.ac.uk

[6] Nebraska Biobank, University of Nebraska Medical Center. https://www.unmc.edu/research/biobank.html

[7] *Seven Bridges launches Unified Patient Network to facilitate clinical research network with aim to advance Precision Medicine and improve patient care*, Contify Life Science News, 2 Dec. 2021.

[8] *2bPrecise: Precision health platform available on EPIC App Orchard*, Wireless News, 20 November 2021.

standards for oncology clinicians and researchers, is built on the FHIR standard, and is being implemented at Vanderbilt-Ingram and other cancer centers.[9]

In practice, these advances in records and data integration mean that many more EHRs contain genomic data. EHRs, when held by HIPAA-regulated entities, are lower on the risk continuum than genomic data aggregated in unregulated environments. However, EHR records containing genomic data are more difficult – if not impossible – to truly deidentify. Also, there is an emerging problem with EHRs that contain genomic data; that is, the newly-established HIPAA interoperability and patient access rules from 2020 that facilitate the sharing of patients' EHR with requesting parties, whether or not the parties are covered entities under HIPAA, and whether or not the parties are valid health researchers conducting a valid study.

With a few clicks, patients can send their health data – and increasingly, their genomic data -- in a nearly frictionless manner to other health care providers, to qualified researchers, to themselves as a backup record, and unfortunately, to a host of other parties that are primarily seeking to monetize the data for other values (such as marketing), and are not primarily conducting health research. Even those that are conducting health research may well not be doing so under the auspices of the Common Rule, which is a serious risk today.

Because the interoperability rule is so new and still rolling out, we are monitoring early developments. Thus far, we are seeing early signs of potential issues regarding the qualifications of entities requesting access to patients' EHRs.[10] NIH needs to assess these specific risks early and determine what steps need to be taken to address them. We propose some steps in these comments, particularly in installing requirements for robust data use agreements.

B. Secondary uses of genomic data

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, even if that repository is intended for research purposes. This is something WPF said in comments to NIH in 2006 regarding its RFI for its GWAS repository policy.[11] The risks we mentioned then have proven to be true, and even more so today. Here, we briefly note that even though NIH is conducting its activities for research purposes, the NIH genomic repositories will continue to be of high interest to secondary parties and uses, including access by non-research parties, and research that is not subject to the Common Rule, among other issues. It is worth putting every available remedy in place. We discuss some of the risks below.

---

[9] mCODE Initiative, https://mcodeinitiative.org "The initiative to create a core cancer model and foundational HER data elements."

[10] *The New Healthcare Interoperability Rules: A risk and compliance perspective*, Association of Healthcare Internal Auditors, January 2021. https://ahia.org/AHIA/media/WhitePapers/PWC_Healthcare-Interoperability-Risk-Compliance-White-Paper_Updated.pdf

[11] NIH Genome Wide Association Studies, RFI, Comments of the World Privacy Forum, 29 October 2006. http://www.worldprivacyforum.org/wp-content/uploads/2006/10/WPF_NIH_RFIGWAS10292006fs.pdf

1. Law enforcement and other non-research access to genomic data held by HHS / NIH

Genetic data is of interest to and is actively used by numerous law enforcement agencies in different ways.[12] 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 data stores or repositories is especially vulnerable to secondary uses. Assuming that the Privacy Act of 1974 (Privacy Act) will apply to genomic data held by federal agencies, we offer by way of example that 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 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.

Several of the routine uses that apply to all of HHS also authorize disclosure to law enforcement with few substantive and procedural protections. Further, routine uses under the Privacy Act for a specific system of records can authorize additional disclosures. If NIH establishes a system of records for its GWAS repositories and other genomic data intended for research purposes, 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 GWAS and other systems containing genomic data, the records could still be widely disclosed, including for non-research purposes.

The NIH already administers a Certificate of Confidentiality program,[13] so we understand that NIH already has knowledge that 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.[14]

---

[12] 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).

[13] *Certificates of Confidentiality*, National Institutes of Health. https://grants.nih.gov/policy/humansubjects/coc.htm

[14] 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

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 and Data Use Agreements, which we discuss at length in Section II of these comments, are among the instruments that might limit the ability of a record keeper to make voluntary disclosures of personal information. We mention Certificates of Confidentiality here to note that they will work hand in hand with an updated approach to Data Use Agreements.

2. Secondary access to private sector, commercial genomic repositories

Additional access to genomic records can occur in today's version of unregulated private sector genomic holdings. This is an area of potentially great vulnerability for genomic data. We do not expect NIH to attempt to control genomic collections held outside of NIH. However, we understand that the compilation of multiple genomic data stores is becoming the norm in the big genomics data environment. We note that any access to existing commercial, private sector genomic repositories needs to be attended to with great care.

If NIH-funded researchers are seeking data from commercial genomic repositories, such as Direct to Consumer genetic testing companies, ideally, research subjects would give meaningful consent for such use directly to NIH, instead of NIH relying on weak and potentially unethical privacy and confidentiality practices in the DTC genetic testing environment. Practices in the private sector vary widely, and we suggest that the NIH establish its own guidelines for the use of this data type.

The well-publicized identification of the Golden State Killer using DNA evidence that was collected for law enforcement purposes from private sector, Direct-to-Consumer DNA companies is an important example here.[15] The relevance of that case is that protection of the privacy of genetic information is not simply a matter of individual privacy, for the simple reason that an individual's DNA can lead to the identification, classification, and categorization of relatives of that individual. We do not object to the arrest of the Golden State Killer. Our argument is regarding the *access* to the genomic data; there could easily be *other uses* of familial DNA that have other, unwelcome consequences for individuals, families, and society at large.

The availability of so much DNA information may make it difficult for those responsible for its maintenance to resist pressures to violate the terms of collection and maintenance for the DNA information. That appeared to be the case in the Golden State Killer utilization of genomic data.[16] We remain deeply concerned about how "genomic big data" in private sector and commercial hands will be utilized in the future, including for research purposes. There are few guardrails in place for entities not subject to the Common Rule. Genetic information about

---

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.

[15] See, e.g., Paige St. John, *The untold story of how the Golden State Killer was found: A covert operation and private DNA*. (LA Times, Dec. 8, 2020), https://www.latimes.com/california/story/2020-12-08/man-in-the-window.

[16] Id.

Americans and others, particularly in conjunction with the pandemic, has been widely collected, and this has created more demand for data of all kinds.

It is foreseeable that as time goes on, entities not subject to HIPAA privacy and security rules could hold significant quantities of individual-level genomic information, and this data may be utilized for many other unrelated purposes including human subject research not subject to the Common Rule. We support research use of the data under the Common Rule. In order to protect public trust as well as individuals' and groups of individuals' genomic data so that it is only utilized for valid health research, we urge the NIH to consider leakages and usages that are emerging and pose significant challenges to ethical use, patient trust, and transparency in uses of genomic data, among other issues.

C. New data sets, including utilization of private sector data

Concerns about the limits of deidentification of personal information go beyond repositories of DNA.[17] The deidentification of "anonymized" personal information is threatened by the ongoing increase in the availability of new data sets, including private sector data sets made available for research purposes. This data may be acquired and utilized apart from any NIH agreement, as such data sets are seen as "contextual" and have evaded privacy controls thus far. Data brokers that sell consumer information have been around for decades. However, the pandemic has created an extraordinary uptick in the creation of new consumer data sets, as well as increasing demand for "contextual patient data" using data sets from social media, geolocation data sets, telecommunication data sets, and even retail datasets. This, combined with the ever-expanding capabilities of computers and artificial intelligence has in its totality created a new environment for human subject research.

Even in cases where it appears that reidentification is impossible, there may be repositories of data that allow some parties in possession of those repositories to reidentify data that others cannot. A1998 statement about anonymization by Professor Latanya Sweeney is even more true today than when she first uttered it: "I can never guarantee that any release of [deidentified] data is anonymous, even though for a particular user it may very well be anonymous."[18] NIH needs to prepare for the future and recognize the looming failure of deidentification amidst the preponderance of new datasets, both in aggregate, and those containing microdata or identifiable data.

The Office of the Privacy Commissioner of Canada recently appeared before the Canadian Parliament to testify about the problems with the Canadian government's collection and use of aggregate cellphone data for public health purposes. In his statement, the Commissioner addressed problems with the application of Canada's privacy law and principles to the use of data for public health purposes; interactions between the Office of the Privacy Commissioner of Canada and the Public Health Agency of Canada; the broadened uses of deidentified data by the

[17] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization* (August 13, 2009). UCLA Law Review, Vol. 57, p. 1701, 2010, U of Colorado Law Legal Studies Research Paper No. 9-12, https://ssrn.com/abstract=1450006

[18] National Committee on Vital and Health Statistics, Subcommittee on Privacy and Confidentiality, *Proceedings of Roundtable Discussion: Identifiability of Data* (Jan. 28, 1998).

public and private sectors, and the need for law reform to govern such use.[19] His statement contained notable discussions of the challenges, including the problems of utilizing consent in the big data context.

These are the kinds of issues that NIH will need to address in its data sharing policies. New patterns of public sector data use during the pandemic have opened up a perceptible rift in public trust regarding data used in public health research contexts. When "research data" involves data of the public, when they did not specifically consent to those research uses, there is a shift today toward less trust of those additional uses.

D. Data accuracy and data accuracy requirements in light of expansive new data uses (and potential re-identification)

We do not expect NIH to take responsibility for uses of genetic information that it cannot control or influence. At the same time, we must not lurch toward premature policies in our haste to utilize all data sets, even those used without appropriate guardrails. One of those guardrails involves data accuracy in the context of AI and big genomic analysis. In regulated human subject research, accuracy is attended to with extraordinary detail. But high levels of accuracy are not required for all data sets of consumer information, and today, these broader "contextual" data sets are being used in combination with other more regulated data sets in the health context.[20] In this time of expanded uses of new and unregulated data sets, accuracy may be the first casualty.

To offer an example, data sets containing credit scores, neighborhood scoring, and other demographic or purchase history information[21] are increasingly sought by health researchers and clinicians to provide context for their health research activities.[22] This effectively combines protected data of known accuracy with unprotected data of unknown accuracy.[23] A marketer using a private sector data set may be happy enough to use data suggesting a fifty-fifty chance the consumer (or group of consumers) of having a particular disease. The cost to the marketer of being wrong may be small, and the consequences to the individual narrow. Contrast this example with an employer using that same dataset, faced with hiring someone with a fifty-fifty change of an expensive disease. That employer may not be willing to take the risk given that the costs of being wrong could easily be greater. The consequence to the unhired consumer would be significantly greater as well. But what would be the consequence of human subject research that relies on consumer data sets of unquantified accuracy? We posit that the consequences could be

---

[19] *Statement of Daniel Therrien, Privacy Commissioner of Canada*, Appearance before the Committee on Access to Information, Privacy, and Ethics and their Study of the Collection and Use of Mobility Data by the Government of Canada. 7 February 2022, Ottowa, Ontario. https://priv.gc.ca/en/opc-actions-and-decisions/advice-to-parliament/2022/parl_20220207/

[20] J. Margolis, *How consumer data (not more clinical data) will fix healthcare*, MedCity News, 9 April 2018. https://medcitynews.com/2018/04/consumer-data-not-clinical-data-will-fix-healthcare/

[21] P. Dixon and B. Gellman, *The Scoring of America*, World Privacy Forum, 2 April 2014. https://www.worldprivacyforum.org/2014/04/wpf-report-the-scoring-of-america-how-secret-consumer-scores-threaten-your-privacy-and-your-future/

[22] *Assessing attitudes to lifestyle data and health research*, Consumer Data Research Center, 5 October 2017. https://www.cdrc.ac.uk/assessing-attitudes-lifestyle-data-health-research/

[23] Price II, William Nicholson, *Problematic Interactions between AI and Health Privacy,* (March 3, 2021). 2021 Utah L. Rev. 925, U of Michigan Public Law Research Paper No. 21-014. https://ssrn.com/abstract=3797161

significant. While the combined data may be regarded as probabilistic, there are still ethical issues with the interpretation of this data by end-users who may not understand the nuances of the differences in accuracy.

We recognize that controlling inaccurate consumer data sets and the resulting analysis derived from those data is not fully within the capability of NIH to provide. Statutory rules require Congress to act, and we do not see any immediate prospect for congressional action barring combined uses of genomic data and private sector consumer data sets.

One response is to ensure that there are firm rules regarding the accuracy of any datasets – genomic or contextual -- that are used in human subject research. Even if a data set is used for "contextual purposes," it needs to be accurate and have proven and verifiable accuracy.

Another tool to consider is to use one or more analytical methodologies for protecting the identity of data subjects, including during the research process. These may have some applications to genetic data. There is no escaping from the truth that in the United States, there are only a small number of individuals with *Progeria, Gitelman Syndrome*, and other similarly rare genetic conditions. It will be practically impossible to protect individuals with many rare conditions (and their families) from reidentification in many circumstances.[24]

It will also be the case that some law enforcement actions and some activities by commercial companies and foreign governments are beyond the scope of anything that NIH can do through rules and guidance. NIH, however, can and should take stronger actions for any activities that fall within its influence, recognizing that the proliferation of sources of DNA, and the greater interest and use of DNA data – including when combined with contextual data of unknown accuracy -- will only increase pressure on the research community.

II. Data Use Agreements as a key administrative tool for protecting genomic data

The most promising administrative tool to protect the privacy of DNA (and other health) data is through ***data use agreements***. Data use agreements may take the form of contracts, memoranda of understanding, or other instruments. Data use agreements should be comprehensive**,** detailed, and strictly enforceable. We will address enforcement again later in these comments.

We recognize that data use agreements are not a new idea for NIH, and some of our suggested standards may reflect current practices already in place. Nevertheless, we offer these suggestions as a more comprehensive set of requirements for future data use agreements, no matter what form they take.

[24] Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, *Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models*, 10 Nature, 3069 (2019), https://www.nature.com/articles/s41467-019-10933-3.pdf

A. Chain of Custody

NIH should require a strict **chain of custody arrangement** such that all who have access to any individual level health data (whether deidentified in principle or otherwise) should be required to sign a standard data use agreement. This includes not just the principal investigator for a research project but each individual working for the project who has the ability to access the data. If the data goes to a repository or to another researchers, the recipient must also sign the agreement or, in the case of repositories, operate under comparable conditions. The name of each signatory should be accessible to the public through the Internet. The goal here is to emphasize personal responsibility and to facilitate transparency.

All data transfers should be publicly reported as well. It is important that people are not subject to human subject research without knowing that this is the case. This must be a fundamental ethic.

B. Reidentification

Each recipient of data under a data use agreement must agree not to reidentify or attempt to reidentify any information received under the agreement. Each recipient must take reasonable steps to prevent any related party from reidentifying or attempting to reidentify any information received under the agreement. Each recipient of data must agree to notify their own management as well as NIH of: 1) any attempt by any person subject to the data use agreement to reidentify data obtained through the agreement; and 2) any use of the data for a reidentification effort by anyone else.

C. Further Use

Each recipient of data under a data use agreement must agree not to further use or disclose information received under the agreement except in accordance with that data use agreement.

D. Security

The agreement must require each recipient of the data to maintain the data in accordance with a written and public security policy posted on the Internet that states a commitment that data will be encrypted at all times whether at rest or in motion. A security policy should also require that the recipient of data maintain reasonable physical, administrative, and technical safeguards to protect against improper data transfers as well as reidentification of personal information. Some technical security details may, of course, be withheld from the public. NIH should provide a model security policy that each recipient of the data can formally adopt, or each recipient can establish their own security policy that is as strict or stricter than the NIH model.

E. Third Party Beneficiary

Each individual who is the data subject of any individual level record should be expressly designated as a third party beneficiary of the data use agreement. Under current law, a data subject may be unable to sue relying upon an ordinary contract or other agreement involving the source of

data and the use of that data. The data subject is not a party to the contract and ordinarily lacks privity – an adequate legal relationship – to the contract or agreement. The goal of a third party beneficiary clause is so that a data subject can enforce their interest in confidentiality by relying on the obligations in a contract or agreement.[25] We suspect that lawyers for the research community will complain loudly about any third part beneficiary clause because it offers some real prospect of enforcement.

F. Penalties

The consequences for anyone violating the terms of a data use agreement should be severe. For researchers, a penalty should include a ban on access to any NIH data for a period of years, and a severe penalty would be a lifetime in case of attempts to reidentify data for gain or commercial use; in cases of gross negligence; or for repeated violations. The penalties should be sufficient that all researchers will take notice. All penalties should be publicly reported as well.

III. HIPAA standards for deidentification

We are concerned about continuing reliance on the two existing HIPAA standards for deidentified patient data. Both standards are troublesome. The NIH RFI references the HIPAA deidentification standard, and this is somewhat unfortunate.

HHS developed the safe harbor deidentification method *decades* ago. In the interim, we've seen the availability of vast new amounts of personal data from multiple sources, and that is separate and apart from the growth of private DNA data banks by direct-to-consumer (DTC) companies and others. It is not a secret in the statistical and research communities that the safe harbor method is out-of-date and needs to be strengthened. One consequence of these developments is that older judgments about what constitutes deidentified data are increasingly obsolete, a trend that will continue indefinitely. Therefore, we believe that NIH's reliance on the HIPAA safe harbor standard is an unfortunate choice today.

HIPAA's other method, the expert determination method, has problems as well. There are no standards for who constitutes an expert. As a result, the expert determination method suffers from the possibility that anyone seeking to deidentify data may be able to find a "hired gun" willing to endorse a marginally effective deidentification method. Other problems with the expert method are the lack of a requirement for publishing the methodology used for making judgments; the lack of a clear standard for assessing the risk of reidentification other than the vague *very small* standard in the rule; and the absence of any effective oversight or enforcement for the expert determination method.

We recognize that NIH is part of the Department of Health and Human Services, and that NIH has limited powers to deviate from the published regulation. We suggest NIH should push back internally to pressure the Department to revisit the safe harbor standard. Further, we believe that there are enough differences between standard health data and DNA data that NIH can propose a

[25] For more on this subject, see Robert Gellman, "The Deidentification Dilemma: A Legislative and Contractual Proposal", 21 *Fordham Intellectual Property, Media & Entertainment Law Journal* at text accompanying notes 109-114 (2010), https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=1277&context=iplj.

separate, modern safe harbor method for patient data that includes DNA. NIH should also consider offering more guidance for those using the expert method for deidentification of DNA data. Additional guidance from NIH does not have to be in the form of a regulation, but additional direction for the deidentification of DNA data seems within the realm of possibility. We would like to see all deidentification efforts for patient data of all types supplemented with data use agreements that could serve to block the gaps, existing or future, created by the shortcomings of all deidentification methods.

IV. Consent

Consumer consent in the privacy arena is an increasingly troubled concept today. Entire industries with tens of billions of dollars of revenue rely in large measure on some form of consumer consent. They succeed only because consumers do not understand what they consent to; because they present consent choices in a confusing way that benefits those seeking consent rather than consumers; and because those seeking consent in the United States are not obliged to present meaningful choices.

In a presentation at the Institute for Bioethics, Health Policy, and Law at the University of Louisville, Mark Rothstein noted that identifiable data sets may be linked with publicly accessible information such as vital statistics, military service records, employment records, financial and consumer information, educational records, travel information, social media postings, and government records. He asked: "Should these possible uses by third party health researchers be disclosed in the informed consent process?[26]

A recent study by Consumer Reports illustrates problems in the context of direct-to-consumer companies that offer DNA testing and related services. Their testing found that the privacy policies associated with this testing provided that when consumers opt-in to research uses of their data, that "many are providing third-party access not only to their DNA but also to other types of data the company has about you, which can include information about your relatives and family history." [27]

The World Privacy Forum supports broadly NIH's goal of "maximizing scientific advances and public benefit by sharing genomic data and associated phenotypic data." Still, we think that NIH can do more to provide consumers with more information so that consumers have real choices the reflect their own interests and concerns. We observe that, over decades, polling has consistently shown that consumers want to be asked for consent to allow their health records to be used for research. That is not the policy under HIPAA.

---

[26] Mark A. Rothstein*, Informed Consent for Data Sharing*, Presentation, Institute for Bioethics, Health Policy, and Law, University of Louisville School of Medicine. https://www.cdrc.ac.uk/assessing-attitudes-lifestyle-data-health-research/ See also The Journal of Law, Medicine, and Ethics, *Special Issue on Unregulated Health Research Using Mobile Devices*, edited Mark A. Rothstein and John T. Wilbanks, Spring 2020.

[27] Catherine Roberts, *The Privacy Problems of Direct-to-Consumer Genetic Testing* (Consumer Reports, 2022), https://www.consumerreports.org/dna-test-kits/privacy-and-direct-to-consumer-genetic-testing-dna-test-kits-a1187212155/.

The consent process used to obtain authority to use current data for future DNA research does not reflect what consumers really want. Still, we do not think that it is practical to ask consumers for specific consent for each future research uses of their data. There is a significant tension here between personal rights and the public interest. This tension is not easily resolved.

There are better alternatives than collecting consents for DNA research that are the moral equivalent of what consumers find everyday from commercial private sector companies. We offer several suggestions.

First, before being asked for consent for unidentified and unknown linkage studies, consumers should be educated about what they are asked to consent to. This type of education can be easily done today through websites, videos, phone apps, or other media.

Second, consumers should have to pass a test showing that they understand the choices they face. Only a few questions will be needed to accomplish this purpose, and the testing process will take no more than a few minutes.

Third, educational and testing material should be prepared by a neutral party and should be even-handed in identifying the benefits and the risks. We do not want anyone's thumb on the scale.

Fourth, consumers should have better assurances that data they make available for research purposes will not be used against them in a court, by police, in the economic marketplace, by employers, by educators, or by others for commercial purposes. We recognize that Certificates of Confidentiality provide some degree of protection. While we agree that Certificates of Confidentiality are necessary to assure consumers, they are not sufficient by themselves. The "loophole" that allows some disclosures required by federal, state, or local laws indicates that the certificates do not offer a fully sufficient range of assurances. We recognize the limits that NIH may face in issuing stronger certificates, but NIH could supplement existing restrictions with improved data use agreements as suggested above.

The risks of failure to provide meaningful consent may seem remote today, but they are real. It may be noteworthy that in the context of COVID-19, those demanding personal rights have won their share of battles over public health and the public interest. Whether these types of conflicts will spill over into health research activities remain to be seen. Now that public health has lost much of its positive image, we expect additional push back and scrutiny.

We observe that the personal views of members of Congress on health research access to records are likely not significantly different than consumers at large. If asked, majorities will offer the same opinion that consumers should be able to consent before their health records are disclosed for research. Those opinions will only become stronger when consents are totally open ended and cover all types of unknown future research.

If the consent issue ever came to an open vote in a legislature, there is a good chance that the research community would be unhappy about the result. We think that when legislatures eventually rein in Direct-to-Consumer DNA testing companies, scientific researchers could be caught up in the same restrictions. More effort today to adopt better consent policies will do

much to distinguish scientific from commercial activities and to avoid future legislative restrictions. NIH can do more to disambiguate its activities from the "research" that is now being conducted quite apart from Common Rule guardrails.

V. Institutional Research Boards, Privacy Boards, or Equivalent Bodies

It is difficult to generalize about institutional review boards (IRBs), privacy boards, or equivalent bodies. Some do an excellent job. But many are simply indifferent to privacy or are incapable of protecting the privacy of research subjects. We wonder as well whether many have the capability of properly examining the details of deidentification methodologies. The World Privacy Forum believes that the role, adequacy, and function of IRBs need a broad reassessment. That assessment is, obviously, far beyond the scope of what NIH seeks to do today. We raise this matter because we think that the shortcomings and inconsistencies of IRBs are well known within the research community. IRBs provide tremendous "cover" for research activities, but how much protection of privacy results is less certain.

As we discussed regarding certificates of confidentiality, IRBs seem necessary but not sufficient to protect consumers. Very few seem willing to raise the fundamental problems with IRBs or to look down the road toward refurbishing or strengthening IRBs in the health research process. We urge NIH to find a way to start the discussion about the future of IRBs.

VI. Enforcement by NIH

The discussion in these comments has already considered some aspects of enforcement of deidentification policies. All of this being said, we think that there is a need for more focused oversight of deidentification activities. All of the enforcement methods in use today with respect to identification and deidentification have significant shortcomings. For the most part, potential enforcers look the other way. We think that NIH should and could take more steps on enforcement, which we discuss further below.

A. Formal audit, enforcement, and policy office at NIH

Our suggestion to better future-proof the challenging issues of deidentification, reidentification, audit, and enforcement is that there should be a formal audit, enforcement and policy office at NIH. This office should have a broad assignment of overseeing, investigating, and auditing deidentification activities of researchers, of repositories, and of IRBs. The office should have authority to investigate complaints about deidentification matters from the public or from researchers.

B. Oversight for data use agreements, authority to audit, imposition of penalties

This office should also be tasked with overseeing data use agreements, including those involving deidentified data. Data use agreements are useless if no one is looking to see if researchers are complying with them. The office can be assigned to impose and enforce penalties on researchers and other who violate data use agreements. Data use agreements should also acknowledge the role of this office and require those who sign the agreements to cooperate with the office.

C. Reviewing expert determinations under HIPAA standard

Another task should be reviewing expert determinations under the HIPAA standard to make sure that those determinations are reasonable. In all cases where deidentification methods have been in place for a period of years, the office should be tasked with assessing whether the circumstances, technology, and availability of consumer information from other sources bring the original judgment into question. This type of activity could be conducted with the assistance of those engaged in deidentification activities, and it could include public hearings or conferences.

D. Education, training, and guidance for IRBs regarding deidentification / reidentification

The office might also provide expert advice to IRBs that lack the ability to assess deidentification issues presented to them. These assignments for a deidentification oversight office may not exhaust the list of useful assignments, but they are a good start.

VII. Conclusion

The World Privacy Forum again thanks NIH for the opportunity to offer these comments. We recognize the complexity of these issues, and hope that some of our suggestions will be useful. We are concerned that legitimate research will be harmed by the issues we have raised in these comments. Consumers and patients are losing faith in public health, and they are losing trust in non-transparent, non-consensual research using their data, from mobile data to EHR data.

We urge the NIH to look further toward the future in order to guide how to best design new safeguards today. It will be neither simple nor easy. However difficult tackling the issues may be -- from big genomic data to deidentification methods and effects to unregulated research, and to the lure of genomic repositories to non-research uses -- the difficulties are far outweighed by the benefits of addressing these issues earlier, rather than later.

Respectfully submitted,

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