A Practical Path Toward Genetic Privacy in the United States

April 2020

FUTURE OF PRIVACY FORUM

PRIVACY ANALYTICS
an IQVIA company

ABOUT THE FUTURE OF PRIVACY FORUM

The Future of Privacy Forum (FPF) is a catalyst for privacy leadership and scholarship, advancing responsible data practices in support of emerging technologies. FPF is based in Washington, DC, and includes an advisory board comprising leading figures from industry, academia, law, and advocacy groups.

ABOUT PRIVACY ANALYTICS

Privacy Analytics, an IQVIA company, allows healthcare organizations to quickly and easily apply a responsible de-identification methodology that ensures individual privacy and legal compliance.

Authors:
Carson Martinez, Policy Fellow, Future of Privacy Forum
Elizabeth Jonker, CIPP/C, Privacy Analytics

Acknowledgments
This paper benefited from contributions and editing support from Rachele Hendricks-Sturrup, FPF Health Policy Counsel; Katelyn Ringrose, FPF Policy Fellow; Kelsey Finch, FPF Policy Counsel; Khaled El Emam, Founder of Privacy Analytics; Luk Arbuckle, Chief Methodologist of Privacy Analytics; Shweta Kumar, FPF intern; and Joshua Cornwell, FPF intern.

For questions about this publication, please contact Rachele Hendricks-Sturrup at rhendrickssturrup@fpf.org.

TABLE OF CONTENTS

INTRODUCTION

The volume of genetic data[1] is growing—some estimates predict between 100 million and 2 billion human genomes will be sequenced by 2025 worldwide.[2] While the number of people whose genomes have been sequenced is relatively small in comparison with the total patient population, genetic sequencing is becoming faster and less expensive, and shared databases of genetic data are increasingly proliferating. Genetic data represents an important and burgeoning subsection of personal data that is being produced by both consumer-facing and healthcare-directed companies for genealogical or health purposes. The collection and processing of genetic data have been shown to accelerate biomedical discoveries and to aid in personalized medicine. As genetic data continues to be generated at a growing rate, data storage requirements are likely to be enormous as distribution moves toward cloud-based infrastructure.[3] Ensuring security and privacy will be increasingly important to maintain trust, particularly in light of the risks presented by privacy breaches involving patient information.[4]

Genetic data is typically stored with or linked to identifying or quasi-identifying records (e.g. electronic health records, consumer profiles, socio-demographic information, and other clinical information).[5] The most common approach for protecting genetic data today

[1] Genetic information is defined under the Health Insurance Portability and Accountability Act (HIPAA) as: "with respect to an individual, information about:

(i) The individual's genetic tests;
(ii) The genetic tests of family members of the individual;
(iii) The manifestation of a disease or disorder in family members of such individual; or
(iv) Any request for, or receipt of, genetic services, or participation in clinical research which includes genetic services, by the individual or any family member of the individual."

See: '45 CFR 160.103 - Definitions.' (LII / Legal Information Institute) <https://www.law.cornell.edu/cfr/text/45/160.103> accessed 6 March 2018.'45 CFR 160.103 - Definitions.' (LII / Legal Information Institute) <https://www.law.cornell.edu/cfr/text/45/160.103> accessed 6 March 2018.; For the purposes of this paper, we define genetic data as data that concerns information about an individual's inherited or acquired genetic characteristics, as well as phenotypic characteristics which can be inferred based on specific genetic characteristics, derived from the sequencing or analysis of human DNA, RNA, and chromosomes. Sequencing is typically accomplished through gene sequencing, exome sequencing, and whole genome sequencing (WGS). The analysis of human DNA includes targeted diagnostics, population-based screening tests, large-scale platforms, and other genetic testing techniques.

[2] Zachary D. Stephens and others, 'Big Data: Astronomical or Genomical?' (2015) 13 PLOS Biol e1002195.

[3] ibid.

[4] Identity Theft Resource Center, 'ITRC Data Breach Report 2016' (2017) <http://www.idtheftcenter.org/images/breach/2016/DataBreachReport_2016.pdf> accessed 4 May 2017.Identity Theft Resource Center, 'ITRC Data Breach Report 2016' (2017) <http://www.idtheftcenter.org/images/breach/2016/DataBreachReport_2016.pdf> accessed 4 May 2017.

[5] Simson Garfinkel, 'De-Identification of Personal Information' (2015) NISTIR 8053. Simson Garfinkel (n 21). Simson Garfinkel (n 20). Simson Garfinkel (n 19).

is de-identification of information that accompanies it (e.g. “accompanying personal data”). In the US, the Health Information Portability and Accountability Act (HIPAA) Privacy Rule protects individual privacy by limiting the use and disclosure of individuals' protected health information (PHI)[6] when held by a covered entity or business associate[7] and specifies the circumstances under which PHI is considered de-identified. After PHI has been de-identified, there are no restrictions on how the information may be disclosed by covered entities or business associates. According to the HIPAA Privacy Rule, genetic data and its accompanying personal data must be de-identified when utilized for purposes other than treatment, payment, or healthcare operations. De-identification may be achieved by either reliance on an expert or via the Safe Harbor Method, which has become the standard practice due to the ease of its implementation. The Safe Harbor Method requires the removal of specific identifiers for data to qualify as de-identified, including data points like: names, email addresses, social security numbers, and more. Genetic data itself is not explicitly specified as one of the 18 specific identifiers that must be removed for data to be considered de-identified, and therefore it may be possible to release such data for the purposes of analysis under the HIPAA Privacy Rule.[8] On the other hand, biometric identifiers are specified as one of the 18 identifiers that must be removed under HIPAA Safe Harbor, and there is growing interest in utilizing DNA typing methods for biometric purposes.[9]

While the relevant US federal agencies have yet to rule on whether they consider genetic data itself to be personal information, there is an increasing belief among academics and privacy professionals that genetic data presents unique potential privacy risks, as it remains largely unaltered during one's lifetime, and often implicates both the individual's family members and future generations, and poses significance for particular cultural groups and individuals. Further, genetic data contains information related to a variety of factors such as ethnic heritage and disease predispositions, among other distinguishing traits. Due to these unique characteristics, many have argued that genetic data should warrant a higher level of privacy protection than traditional health information.[10] Others

[6] According to the HIPAA Privacy Rule, PHI is information, including demographic information that relates to past, present, or future physical or mental health status or condition. Provision of healthcare, or payment for the provision of healthcare in any form (written, electronic, or oral) that identifies the individual or for which there is a reasonable basis to believe can be used to identify the individual. See 45 CFR 160.103, 164.502.

[7] Covered entities are defined as health plans, health care clearinghouses, health care providers who transmit PHI, and business associates, where applicable. Business associates are organizations with whom covered entities share health information to help carry out their activities and functions. See: 45 CFR 160.102.

[8] A 2013 amendment to the HIPAA Privacy Rule incorporated genetic information into the definition of “health information.” Health and Human Services Department. s.l. : Federal Register, Jan 25, 2013, pp. 5565-5702. 78 FR 5565.

[9] National Institute of Standards and Technology, ‘DNA Biometrics’ <https://www.nist.gov/programs-projects/dna-biometrics> created March 11, 2010, updated July 13, 2017.; ‘IBIA I Biometric Technologies I DNA’ <https://www.ibia.org/biometrics-and-identity/biometric-technologies/dna> accessed 16 July 2018.

[10] Genetic exceptionalism is the concept that genetic information should be treated differently from other health information for the purposes of data access and permissible use. While we do not engage in the genetic exceptionalism debate in this paper, we considered whether or not genomic data should be

have argued that research involving sequencing of so-called “anonymous” genetic data should be considered human subjects research,[11] thus requiring more nuanced consent and scrutiny by institutional review boards (IRBs).[12]

Researchers in the past 10 years have begun to document several theoretical ways in which the privacy of genetic data, whose accompanying personal data has been de-identified, may be compromised through re-identification;[13] the methods of such re-identification attacks are also evolving and becoming increasingly feasible.[14] Genetic privacy has thus emerged as a legitimate, yet challenging concern for both individuals and their families. In light of these studies, it is increasingly likely that new techniques will be required to ensure that genetic data does not become readily identifiable and to keep pace with the emergence of new re-identification risks for genetic data.

However, applying more advanced de-identification techniques may not be functional in practice—as the strength of the de-identification techniques applied can have a negative impact on a dataset’s utility. There is often a tradeoff between the level of privacy protection afforded, and the level of utility in the data, a concept well understood in statistical disclosure control communities[15] and increasingly acknowledged by genetic communities as scientific norms and capabilities evolve.[16] Researchers require access to robust data for genetic studies, and further de-identifying genetic data could greatly thwart research, especially if phenotypic data needs to be analyzed alongside the

protected differently from other PHI due to the challenges posed by de-identification. For literature related to genetic exceptionalism, see: Emily Darraj and Brian Mcelyea, ‘Security and Privacy of Genomic Data’ (Northrop Grumman Corporation 2017).; Jennifer Kulynych and Henry T Greely, ‘Clinical Genomics, Big Data, and Electronic Medical Records: Reconciling Patient Rights with Research When Privacy and Science Collide’ (2017) 4 Journal of Law and the Biosciences 94.; & Amy L. McGuire and others, ‘Confidentiality, Privacy, and Security of Genetic and Genomic Test Information in Electronic Health Records: Points to Consider’ (*Genetics in Medicine*, 1 July 2008) <https://www.nature.com/articles/gim200876> accessed 9 January 2018.

[11] Human subjects research is defined as research involving “an individual about whom an investigator...conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information.” See: Protection of Human Subjects, 45 Code of Federal Regulations (CFR) Part 46 2009 4.

[12] Amy L. McGuire and Richard A. Gibbs, ‘No Longer De-Identified’ (2006) 312 Science 370.

[13] Yaniv Erlich and Arvind Narayanan, ‘Routes for Breaching and Protecting Genetic Privacy’ (2014) 15 Nature Reviews Genetics 409.

[14] Ewan Birney, Jessica Vamathevan and Peter Goodhand, ‘Genomics in Healthcare: GA4GH Looks to 2022’ [2017] bioRxiv 203554.; Ruichu Cai and others, ‘Deterministic Identification of Specific Individuals from GWAS Results’ [2015] Bioinformatics btv018.; Arif Harmanci and Mark Gerstein, ‘Quantification of Private Information Leakage from Phenotype-Genotype Data: Linking Attacks’ (2016) 13 Nature Methods 251.

[15] George T. Duncan, Sallie A. Keller-McNulty and S. Lynne Stokes, ‘Disclosure Risk vs. Data Utility: The R-U Confidentiality Map’ (Los Alamos National Laboratory 2001) LA-UR-01-6428. https://pdfs.semanticscholar.org/aae7/2110a204e0db8eaab9a9941c7a3b2ecef354.pdf

[16] Jalayne J. Arias, Genevieve Pham-Kanter and Eric G Campbell, ‘The Growth and Gaps of Genetic Data Sharing Policies in the United States’ [2014] Journal of Law and the Biosciences lsu032.; Jill O Robinson and others, ‘It Depends Whose Data Are Being Shared: Considerations for Genomic Data Sharing Policies’ [2015] Journal of Law and the Biosciences lsv030.

genetic information.[17] Other privacy-enhancing technologies will be needed to address the unique risks posed by genetic data, while maintaining utility and promoting data sharing.

This paper begins by detailing the current regulatory framework for how protected health information (PHI) is currently de-identified. **Section I** specifically walks through the de-identification methods set out in the HIPAA Privacy Rule and explains how genetic data fits into the regulation. **Section II** highlights the major re-identification research challenging the appropriateness of current de-identification practices for genetic data. **Section III** walks through the benefits and challenges posed by new technological innovations in this space, including privacy enhancing technologies like secure computation and differential privacy. And finally, **Section IV** describes the need for governance mechanisms - including access controls and data use agreements - in addition to de-identification throughout the information management life cycle of genetic data to minimize the risk of re-identification, highlighting model guidance documents and policies from organizations currently collecting and sharing genetic data.

I. REGULATORY REQUIREMENTS FOR DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION (PHI)

De-identification is the process of dissolving the relationship between a dataset and an individual, via a variety of approaches, algorithms, and tools, so that the disclosure of information from that dataset cannot reasonably be linked back to an identified individual.[18] How clearly any particular data point can be associated with an individual lies on a continuum—from data that explicitly identifies an individual (such as a name or photo) to completely anonymous data that is not and could not be associated to a specific individual (such as a statistical figure), with many shades of grey in between.[19]

Some types of information readily identify an individual and are known as “direct identifiers” (such as a name, social security number, or email address). Other types of information may identify an individual only when combined with other information (such as a date of birth, home ZIP code, or medical condition); these are known as “indirect identifiers” or “quasi-identifiers.” In order to reduce the risk that an individual can have their privacy violated by being identified in a dataset, both types of identifiers must be addressed.

[17] Jules Polonetsky, Omer Tene and Kelsey Finch, ‘Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification’ (2016) 56 Santa Clara Law Review 593.
http://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=2827&context=lawreview

[18] L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control* (Springer-Verlag 2001).

[19] Polonetsky, Tene and Finch (n 16). Polonetsky, Tene and Finch (n 17). Polonetsky, Tene and Finch (n 16). Jules Polonetsky, Omer Tene and Kelsey Finch, ‘Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification’ (2016) 56 Santa Clara Law Review 593.

A number of techniques are used alone or in combination to remove or manipulate identifiers in a dataset, including: limiting the amount of data released (minimization/sampling); suppressing, generalizing, or aggregating fields or records; and statistically degrading the data by adding random noise (perturbation).[20] Each de-identification technique has disadvantages and advantages, dependent upon the type of data being manipulated, to whom the data will be disclosed, and how the data will be used. The disclosure of personal information can occur in two distinct forms: identity disclosure and attribute disclosure. Identity disclosure is the process by which an individual is assigned to a particular record in a dataset; attribute disclosure is the process by which a characteristic is revealed or inferred about a specific individual in a dataset without knowing which specific record belongs to the individual.[21]

For genetic data, one example of a possible identity disclosure could be the identification of a particular individual in a genetic dataset that also contains state residence. An individual could be singled out in the dataset if that individual has a rare disease that only affects a limited number of individuals worldwide. Knowing the geographic location could make an individual unique under these circumstances (he/she is the only individual in that location with that disease). Attribute disclosure results from an adversary learning private information without necessarily identifying a particular individual. An example of attribute disclosure would be someone learning that people of a certain demographic group carry the gene for a particular inherited disease, such as Huntington's disease. As our knowledge of genetics is still growing, genetic data that does not disclose attributes about individuals or groups today may do so in the future. The harms related to each of these types of disclosures (both perceived and true) can vary from reputational, to opportunity, to financial harms. Many individuals fear that information derived from their genetic data may be misused or abused, specifically for discrimination purposes.[22] Additionally, these disclosures and subsequent harms could result in group privacy concerns, as the information that may be derived from an individual's genetic data may extend to the individual's family members and future generations.

The HIPAA Privacy Rule

The HIPAA Privacy Rule defines information as de-identified when "there is no reasonable basis to believe that the information can be used to identify an individual."[23]

[20] Salinger Privacy, 'Demystifying De-Identification: An Introductory Guide for Privacy Officers, Lawyers, Risk Managers and Anyone Else Who Feels a Bit Bewildered.'

[21] Khaled El Emam, *Guide to the De-Identification of Personal Health Information* (CRC Press (Auerbach) 2013).; Attribute disclosure may occur with or without identity disclosure; identity disclosure can lead to attribute disclosure, and vice-versa.

[22] E. Clayton, "Ethical, legal, and social implications of genomic medicine," New England J. Med., vol. 349, pp. 562–569, 2003.; M. Rothstein and P. Epps, "Ethical and legal implications of pharmacogenomics," Nature Rev. Genetics, vol. 2, pp. 228–231, 2001.

[23] Office for Civil Rights, 'Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule' (Department of Health and Human Services, 2012).

As is the case for the different federal and provincial health privacy laws in Canada and the EU, HIPAA only concerns itself with identity disclosure. The HIPAA Privacy Rule does not address the identifiability risks from attribute disclosure, because protections against attribute disclosure could potentially destroy data utility.[24] For the purposes of this paper, we thus mainly are concerned with identity disclosure.

The HIPAA Privacy Rule presupposes that data properly de-identified through the Safe Harbor Method or the Expert Determination Method pursuant to Sections 164.514(b) and(c) of the HIPAA Privacy Rule does not reveal individuals' identities connected to the data, and is therefore not subject to the regulation.[25] When data is de-identified, it is no longer considered PHI and thus could be released publicly without individual consent.

Under the Safe Harbor Method, data is considered de-identified when it has been stripped of 18 enumerated identifiers, including names, telephone numbers, email addresses, biometric identifiers, and social security numbers, among others.[26] The Safe Harbor Method also requires that the covered entity or business associate does not have knowledge that the PHI, alone or in combination with other information, could link a

[24] Khaled El Emam, *Guide to the De-Identification of Personal Health Information* (CRC Press (Auerbach) 2013).

[25] Office for Civil Rights (n 22).

[26] The 18 identifiers under the Safe Harbor Method include:

(1) "Names;
(2) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000;
(3) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, 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) Vehicle identifiers and serial numbers, including license plate numbers;
(6) Fax numbers;
(7) Device identifiers and serial numbers;
(8) Email addresses;
(9) Web Universal Resource Locators (URLs);
(10) Social security numbers;
(11) Internet Protocol (IP) addresses;
(12) Medical record numbers;
(13) Biometric identifiers, including finger and voice prints;
(14) Health plan beneficiary numbers;
(15) Full-face photographs and any comparable images;
(16) Account numbers;
(17) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section "Re-identification"]; and
(18) Certificate/license numbers."

See: ibid.

particular individual to the disclosed health information. The Privacy Rule also requires that covered entities and business associates adopt reasonable administrative, technical, and physical safeguards to protect PHI from unauthorized access, use, or disclosure.[27] While the Safe Harbor Method is hailed for its convenience, it does not come without its own issues, including the risk of reidentification.[28]

The Expert Determination Method, commonly referred to as the "statistical standard," requires that a person with appropriate knowledge make the judgment and have strong assurances that the risk of re-identification is "very small," by applying statistical and scientific principles and methods that render information not individually identifiable.[29] In contrast to the Safe Harbor Method, which is applied in the same manner to any dataset regardless of its characteristics, the Expert Determination Method allows for the fitting of the de-identification method to the risks associated with the specific dataset being analyzed —i.e., that incorporates the context of the data sharing into the risk assessment framework. The Expert Determination method has been criticized as being costly, time-limited, and difficult to obtain as the number of experts available to do such technical determination is small.[30]

While de-identification will never eliminate all risks arising from personal data sharing, it offers a method to liberate genetic data, while at the same time protecting privacy. PHI that has been de-identified through the Expert Determination Method or the Safe Harbor Method is considered not to involve human subjects under the Common Rule,[31] and thus informed consent is not required for the use and/or release of such de-identified data. Once information has been de-identified through the Safe Harbor or Expert Determination methods, the data is not subject to the HIPAA Privacy Rule, and thus the HIPAA Privacy Rule does not limit how a covered entity or a business associate may use or disclose it.

Genetic Data and the HIPAA Privacy Rule

Today, genetic data is used across the clinical care continuum from predictive to diagnostic testing of patients. Among many other applications, genetic data is used to assess the likelihood and extent of a therapeutic response, the possibility of treatment side effects, and the risks of drug interactions.[32] Genomics has become important for

[27] 45 C.F.R. §164.530(c).

[28] Sweeney, Latanya et al. "Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study." *Technology Science* vol. 2017 (2017): 2017082801. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6344041/>

[29] ibid; Office for Civil Rights (n 22).

[30] Privacy Analytics, 'Safe Harbor Versus Expert Determination' 2015 <https://privacy-analytics.com/de-id-university/blog/hipaa-safe-harbor-vs-expert-determination/> accessed December 5, 2019. See also, The HITRUST De-Identification Framework <https://hitrustalliance.net/de-identification/> accessed on December 5. 2019.

[31] 45 CFR 46: Protection of Human Subjects ('The Common Rule') 1991 (Code of Federal Regulations).

[32] Leslie P. Francis, 'Genomic Knowledge Sharing: A Review of the Ethical and Legal Issues' (2014) 3 Applied & Translational Genomics 111.

screening in oncology and obstetrics, and increasingly is impacting the practices of non-geneticist clinicians and being integrated into routine clinical care.[33] The clinical applications of genetic data, however, do not raise many privacy questions; the HIPAA Privacy Rule specifically permits covered entities and business associates to use identifiable genetic data and its accompanying personal data for treatment, payment, and healthcare operations without individual consent.[34] Covered entities and businesses also are permitted, but not required, to use and disclose genetic data without an individual's authorization when sharing that information with the individual and for one of the twelve specified public interest purposes, such as when required by law or for public health activities.[35]

In addition to these permitted uses and disclosures, genetic data is being generated and used for a myriad of research purposes.[36] The increasing production and availability of genetic data is providing researchers with a rich resource for investigation. Technological advancements to process genetic data also have led to advances in biomedical research and science. Further, genetic data research is shifting from individual-level to populations-level research; this emergence of population-level research is consequently increasing the amount of genetic data required for such research. These shifts also are being accelerated not only by researchers themselves, but also by initiatives from the US federal government, such as the NIH All of Us Research Program.[37] These projects are creating the movement toward sharing more genetic data and unlocking genetic data sources that were previously restricted.[38]

When genetic data held by a covered entity or business associate is disclosed or shared for purposes outside of permitted uses and disclosures specified by the HIPAA Privacy Rule—such as research or other secondary uses—covered entities and business associates must either de-identify the data or obtain consent from the individual for such use. Genetic data–while not addressed explicitly by the statute–presumably would fall into the category of PHI. As such, genetic data could be de-identified through either the Safe Harbor Method or the Expert Determination Method, though the Safe Harbor Method might be considered the most commonly used approach to de-identify data

[33] Joel B. Krier, Sarah S. Kalia and Robert C. Green, 'Genomic Sequencing in Clinical Practice: Applications, Challenges, and Opportunities' (2016) 18 Dialogues in Clinical Neuroscience 299.

[34] Office for Civil Rights (OCR), 'Guidance: Treatment, Payment, and Health Care Operations' (*HHS.gov*, 7 January 2009) <https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/disclosures-treatment-payment-health-care-operations/index.html> accessed 19 April 2018.

[35] 45 C.F.R. § 164.502(a)(1).

[36] Research is defined in the Privacy Rule as, "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge." See 45 CFR 164.501.

[37] 'National Institutes of Health (NIH) — All of Us' <https://allofus.nih.gov/> accessed 16 July 2018.

[38] C. Heeney and others, 'Assessing the Privacy Risks of Data Sharing in Genomics' (2011) 14 Public Health Genomics 17.;Jane Kaye and others, 'Data Sharing in Genomics – Re-Shaping Scientific Practice' (2009) 10 Nature reviews. Genetics 331.

today.[39] Although the Safe Harbor Method's list of 18 identifiers includes "biometric identifiers, including finger and voice prints" or "any other unique identifying number, characteristic, or code," the federal Department of Health and Human Services (HHS) Office of Civil Rights (OCR), which interprets and enforces the HIPAA Privacy Rule, has not issued guidance about whether or not genetic data should be considered a biometric, "other," or is itself an identifier for the purposes of the Safe Harbor Method.[40] Because research data de-identified according to the HIPAA Privacy Rule is considered not to involve human subjects, informed consent is not required for the use and/or release of such de-identified data. Thus, genetic data that is not associated with/linked to any of the 18 elements listed under Safe Harbor Method currently believed to constitute de-identified information.[41] This does not mean that the reveal of genetic data does not constitute a privacy risk.

While de-identification protects many types of health data by decreasing the probability of re-identification, genetic data poses unique privacy challenges: because genes have specifically identifiable elements, they offer the possibility for re-identification when large sets of an individual's genetic data, or sensitive portions thereof, are made available. In addition, for genetic data, the consequences of disclosures are not limited in scope or time as they can affect not only the individual identified, but also the relatives or descendants of the individual. Unlike other health data, such as heart rate or blood type, genetic data represents the unique combination and collection of traits that can partially or fully identify an individual.[42]

Today, it is generally recognized that while a person's genome is individuating and distinguishable from other people's genomes, it is not readily identifiable because identity is not readily ascertainable from the nucleotide bases themselves nor from the genotypic or phenotypic information derived thereof. Thus, there is still "practical obscurity"—interpreting genetic data currently is costly, labor-intensive, and requires highly skilled experts. However, as more and more genetic datasets are being produced and shared, the risk of re-identification has grown and the justification of genetic data's "practical obscurity" is becoming less reliable.

[39] The Expert Determination Method is less frequently used in comparison to the Safe Harbor Method, because the Expert Determination method is more expensive and there are too few experts available for hire. See: William Stead, 'Re: Recommendations on De-Identification of Protected Health Information under HIPAA' <https://www.ncvhs.hhs.gov/wp-content/uploads/2013/12/2017-Ltr-Privacy-DeIdentification-Feb-23-Final-w-sig.pdf>.

[40] Kulynych and Greely (n 9).

[41] Further, the federal regulators at the Office for Human Research Protections (OHRP) who oversee Common Rule compliance for the Department of Health and Human Services (HHS) have not challenged the assumption that genomes themselves constitute de-identified information. In part due to the absence of any private right to sue under the Common Rule or the HIPAA Privacy Rule, there has been little if any case law or judicial interpretation related to the "identifiability" of genomic data.

[42] Michelle Meyer, 'Re-Identification Is Not the Problem. The Delusion of De-Identification Is. (Re-Identification Symposium) | Bill of Health' <http://blogs.harvard.edu/billofhealth/2013/05/22/re-identification-is-not-the-problem-the-delusion-of-de-identification-is-re-identification-symposium/> accessed 5 January 2018.

Beyond HIPAA

While the US federal government has in practice treated genetic data as not readily identifiable when disclosed for research purposes, the National Institute of Health (NIH) has maintained that the disclosure of genetic data constitutes "a clearly unwarranted invasion of personal privacy" when disclosed via a Freedom of Information Act (FOIA) request.[43] In addition, the 21st Century Cures Act guards against inappropriate use of FOIA requests to gain access to genetic information of research participants in federally and non-federally funded research by allowing the Secretary of HHS to disqualify such research data from FOIA requests through Certificates of Confidentiality if: "(A) an individual is identified; or (B) there is at least a very small risk, as determined by current scientific practices or statistical methods, that some combination of the information, the request, and other available data sources could be used to deduce the identity of an individual."[44] Federal interpretations of the identifiability of genetic data thus seem to be in conflict: genetic data is treated as readily identifiable for the purposes of FOIA, but not necessarily readily identifiable under the HIPAA Privacy Rule.[45]

Further, in 2011, HHS published the Advanced Notice of Proposed Rulemaking (ANPRM), entitled "Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators" on whether genetic data should be considered identifiable. The 2011 ANPRM acknowledged that "there is an increasing belief that what constitutes 'identifiable' and 'de-identified' data is fluid" and that evolving technologies and the increasing accessibility of data could allow de-identified data to become re-identified.[46] While a full rule-making process did not occur, recent re-identification research highlights this concern and challenges our current assumptions about the effectiveness of de-identification of genetic data as a reliable means of privacy protection.

II. CHALLENGING DE-IDENTIFICATION OF GENETIC DATA

While de-identification is used today as a broad and generally accepted standard for protecting health data, recent re-identification attacks on genetic data, whose personal data has been de-identified (described in detail below), are challenging the assumption that current de-identification techniques are sufficient to protect this type of data.

[43] National Human Genome Research Institute (NHGRI), 'Privacy in Genomics' <https://www.genome.gov/27561246/privacy-in-genomics/> accessed 17 April 2018.

[44] National Institute of Health, 'Certificates of Confidentiality: Background Information' <https://humansubjects.nih.gov/coc/background> accessed 17 April 2018.

[45] Amy L. McGuire, 'Identifiability of DNA Data: The Need for Consistent Federal Policy' (2008) 8 The American Journal of Bioethics: AJOB 75.

[46] Human Subject Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, 76 Federal Register 143 (2011).

Re-identification is the process by which an adversary attempts to discern the identities that have been removed from de-identified data.[47] Re-identification risk can be described as the measure of how likely it is that an adversary can determine the identity and other personal information of an individual from a de-identified dataset.[48] Re-identification of genetic data can be achieved in a variety of ways, including:[49]

(*1) Matching the genetic data against a reference sample*, which involves directly comparing someone's genetic data to another set of genetic data from individuals who are from the same population as the original genetic sample. These two datasets can be used to confirm that two sets of genetic samples come from the same individual. The re-identification of such an individual will be dependent upon whether or not the reference sample data are personally identifiable.

(*2) Connecting genetic data to non-genetic databases,* which involves deductively linking genetic data and associated personal data—such as age, gender, ethnicity—with other databases, such as publicly available criminal, census, genealogy, health, and voter databases.

(3) Profiling from genetic data, which involves inferring physical traits and attributes from information encoded in the genome itself, such as gender, ethnicity, eye color, hair color, craniofacial characteristics, and height. These attributes can be combined to create a characterization of an individual, that by itself, may not lead to absolute identification, but if combined with other datasets, may lead to re-identification.

These techniques for re-identification of genetic data have been demonstrated in a number of recent research projects. The first attempts at re-identification of genetic datasets were performed by Homer *et al.* in 2008.[50] Researchers combined three datasets to identify an individual: a complex DNA mixture containing DNA from numerous individuals, a reference population, and the individual's genotype. Even when an individual's single nucleotide polymorphism (SNP) profile was aggregated with 1,000 or more other individuals, Homer *et al.* demonstrated that the individual could still be identifiable. This research showed that individual membership can be determined in summary-level allele frequency data when compared against a reference dataset. It has been suggested that a genetic sequence containing 30-80 independent SNPs (a typical human genome sequence contains 3 million SNPs) can uniquely identify an individual,[51] although today there is no scientific consensus on the minimum size of genetic sequence or number of SNPS necessary for re-identification.

[47] Simson Garfinkel (n 5).

[48] El Emam (n 23).

[49] William W. Lowrance and Francis S. Collins, 'Identifiability in Genomic Research' (2007) 317 Science 600.

[50] Nils Homer and others, 'Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays' (2008) 4 PLOS Genetics e1000167.

[51] Zhen Lin, Art B. Owen and Russ B. Altman, 'Genetics. Genomic Research and Human Subject Privacy' (2004) 305 Science (New York, N.Y.) 183. http://science.sciencemag.org/content/305/5681/183.

In 2013, Gymrek *et al.* re-identified 5 people from a DNA database without a reference database. This was done by analyzing the "1,000 Genomes" database, an anonymous, public DNA database[52] that only included individuals' year of birth and state residence (which is acceptably de-identified within the existing HIPAA framework). Researchers identified 34-68 short tandem repeats on the Y chromosome (Y-STRs) in the genomes of 10 individuals from the 1,000 Genomes database. Y-STRs distinguish male lines in families, as the Y chromosome is transmitted from father to son, as are surnames usually. After identifying the Y-STRs, they searched a publicly available genealogy database, which held 40,000 surnames and family pedigrees. After collecting those surnames and with state residence and birth year operating as indirect identifiers, the team queried other sources such as obituaries, genealogical websites, public demographic data, and internet record search engines for matches. From this cross-reference they were able to directly identify 5 of the 10 genomes and their complete pedigrees with high probability and to identify 50 related individuals from 10 genomes in the 1,000 Genomes Project.[53]

Another critical re-identification study published in the Public Library of Science (PLOS) Genetics in 2014 revealed that researchers could use the genome and computerized rendering software to "computationally predict" 3-D models of individual "faces" of particular genomes. Researchers analyzed 176 ancestry information markers (AIMs) to estimate individual genetic ancestry from DNA and map genes for genetically determined traits that vary between populations. Genetic ancestry through AIMs explained 9.6% of the total face variation. They also identified the sex of the individual, which explained 12.9% of the total face variation, and 76 single nucleotide polymorphisms (SNPs) located in 46 craniofacial candidate genes.[54] The combinations of genetic variations related to facial features create the predicted faces, which was mapped onto a face with 7,150 coordinates.[55]

Most recently in September 2017, Lippert *et al.* used physical information and genomic data obtained from whole-genome sequencing (WGS) to train a machine learning algorithm to identify people's traits based on their SNPs associated with face shape, height, weight, hair color, and skin color, and to reconstruct what an individual's face may look like.[56] Researchers reported that they could correctly re-identify individuals using the trait-predicted faces with a 74% accuracy rate. While Lippert et al. garnered much attention about the potential privacy implications that such a technique would

[52] '1000 Genomes | A Deep Catalog of Human Genetic Variation' <http://www.internationalgenome.org/> accessed 13 April 2018.'1000 Genomes | A Deep Catalog of Human Genetic Variation' <http://www.internationalgenome.org/> accessed 13 April 2018.

[53] Melissa Gymrek and others, 'Identifying Personal Genomes by Surname Inference' (2013) 339 Science 321.

[54] Other research on DNA-based facial modeling has been done with an even smaller set of SNPs (24): Peter Claes, Harold Hill and Mark D Shriver, 'Toward DNA-Based Facial Composites: Preliminary Results and Validation' (2014) 13 Forensic Science International: Genetics 208.

[55] Peter Claes and others, 'Modeling 3D Facial Shape from DNA' (2014) 10 PLOS Genetics e1004224.

[56] Christoph Lippert and others, 'Identification of Individuals by Trait Prediction Using Whole-Genome Sequencing Data' (2017) 114 Proceedings of the National Academy of Sciences 10166.

produce, the article also received a great deal of criticism from a number of authors claiming that the results reported were overstated, i.e., the algorithm analyzing the SNPs performed no better than an algorithm analyzing demographic values.[57] In addition to those listed above, other studies have also added to the growing consensus that genetic data are increasingly subject to re-identification through a variety of techniques.[58] In a 2018 report, the authors' note that over half of US individuals – approximately 60% in the case of individuals of European descent – could be identified using open genetic genealogy databases.[59] Although those findings do not depend on deidentification, they showcase the propensity of genetic data to serve as a powerful identifier.

These findings demonstrate that the re-identification of genetic data is possible, even when the accompanying personal data is aggregated or devoid of apparent identifiers as specified by the Safe Harbor Method, putting individual privacy at risk. While a reference sample for matching is currently required for successful re-identification and genetic data is still considered not readily identifiable, these studies demonstrate that re-identification from raw genetic data itself and from other indirect identifiers that the Safe Harbor Method does not require be removed may be possible in the near future. Other studies also suggest that the standard de-identification methods of suppression through the Safe Harbor Method and/or aggregation may not be well suited to raw genetic data, due to the complexity and uniqueness of such high-dimensional data.[60]

Together, the results of these studies have prompted many organizations to revisit their genetic data sharing policies and protocols. In 2008, the NIH and the Wellcome Trust Sanger Institute responded to the possibility of re-identification demonstrated by some of the studies mentioned (notably Homer *et al*.) by moving genetic summary results into

[57] For criticisms, see: Sara Reardon, 'Geneticists Pan Paper That Claims to Predict a Person's Face from Their DNA' (2017) 549 Nature News 139; Yaniv Erlich, 'Major Flaws in "Identification of Individuals by Trait Prediction Using Whole-Genome"' [2017] bioRxiv 185330; Antonio Regalado, 'Sorry, Your DNA Can't Predict Exactly What You Look like (Yet)' (*MIT Technology Review*) <https://www.technologyreview.com/s/608813/does-your-genome-predict-your-face-not-quite-yet/> accessed 21 November 2017; Peter Hess, 'Why Science Turned on the DNA Tycoon Raising Fears About Genetic Privacy' (*Inverse*, 4 October 2017) <https://www.inverse.com/article/36584-craig-venter-genome-dna-privacy-yaniv-erlich> accessed 21 November 2017.

[58] Other notable re-identification research includes: (1) Latanya Sweeney, Akua Abu and Julia Winn, 'Identifying Participants in the Personal Genome Project by Name' [2013] Available at SSRN <http://dataprivacylab.org/projects/pgp/1021-1.pdf> accessed 22 April 2014.; (2) Harmanci and Gerstein (n 13).; (3) Hae Kyung Im and others, 'On Sharing Quantitative Trait GWAS Results in an Era of Multiple-Omics Data and the Limits of Genomic Privacy' (2012) 90 The American Journal of Human Genetics 591.; (4) Eric E Schadt, Sangsoon Woo and Ke Hao, 'Bayesian Method to Predict Individual SNP Genotypes from Gene Expression Data' (2012) 44 Nature Genetics 603.; and (5) Kevin B. Jacobs and others, 'A New Statistic and Its Power to Infer Membership in a Genome-Wide Association Study Using Genotype Frequencies' (2009) 41 Nature genetics 1253.

[59] Erlich Y, Shor T, Pe'er I, Carmi S. See *Identity Inference of Genomic Data using Long-Range Familial Searches.* 6415, s.l. : Science, 2018, Vol. 362.

[60] Khaled El Emam, 'Methods for the De-Identification of Electronic Health Records for Genomic Research' (2011) 3 Genome Medicine 25.

controlled access portions of their data repositories, removing open web access to genetic datasets, and no longer publicly sharing aggregated, de-identified genetic data.[61]

However, in November of 2018, the NIH updated its policies for managing genetic summary results, announcing that it would now grant unrestricted access to genetic summary statistics for most NIH-funded studies.[62] Dr. Eric Green, Director of the Human Genome Research Institute, indicated that the decision to broaden access to genetic data was based on the continued publication of genetic summary results in scientific literature and aggregation of genetic data in public databases, as well as the continued absence of any reported, legitimate attempt at re-identification of de-identified study participants.[63] Millions of consumers have uploaded their DNA to direct-to-consumer genetic testing websites, and combining genomic data with genealogical data is being used to identify criminals based on sequencing biological data left behind at crime scenes. There has been increased pressure, especially since the arrest of the Golden State Killer to utilize genetic testing sites to achieve cold hits, and there have been at least four cold murder cases and one recent rape case solved through similar efforts.[64] However, as aforementioned, aggregated genetic data, including summary-level allele frequency data, still carries a theoretical risk of re-identification.[65]

Due to the potential re-identification both from the genetic data itself or from indirect identifiers that are not considered one of the 18 identifiers required to be removed by the Safe Harbor method, de-identification may pose a higher risk of disclosure for genetic data than currently believed. However, this does not necessarily mean that we should abandon de-identification entirely. As genetic datasets multiply, new methods of de-identification may be necessary to protect such high-risk data. Given how rapidly the landscape of re-identification risk is evolving for genetic data and the unique privacy implications that result, the path forward will require new technological solutions that protect the privacy of genomic data, while concurrently maintaining its analytic utility.

[61] Heeney and others (n 35). Heeney and others (n 36). Heeney and others (n 35). Heeney and others (n 34). Heeney and others (n 35). Heeney and others (n 30).C Heeney and others, 'Assessing the Privacy Risks of Data Sharing in Genomics' (2011) 14 Public Health Genomics 17.

[62] 'NOT-OD-19-023: Update to NIH Management of Genomic Summary Results Access' <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-023.html> accessed 16 January 2019.

[63] 'Protecting Participants, Empowering Researchers: Providing Access to Genomic Summary Results - Office of Science Policy' <https://osp.od.nih.gov/2018/11/01/provide-access-gsr/> accessed 16 January 2019.

[64] Christi Guerinni et. al, Should Police Have Access to Genetic Genealogy Databases? Capturing the Golden State Killer and other Criminals Using the Controversial New Technique, PLOS BIOLOGY (Oct. 2, 2018), <https://www.researchgate.net/publication/328038448_Should_police_have_access_to_genetic_genealogy_databases_Capturing_the_Golden_State_Killer_and_other_criminals_using_a_controversial_new_forensic_technique>; See also Katelyn Ringrose, A Cautionary Note: Genealogy Companies Need to Stop Giving Warrantless DNA Clues to Law Enforcement, Penn State Law Review (2019), <http://www.pennstatelawreview.org/wp-content/uploads/2019/11/Ringrose-Penn-Statim-FormattedAdobe.pdf>.

[65] Homer and others (n 48).

III. TECHNICAL SOLUTIONS

As genetic datasets and big data analytics tools continue to grow in quantity and sophistication,[66] breaches of medical data threaten to erode public confidence in the long-term security of personal information.[67] In order to counter these threats, many data processors are looking to employ privacy engineering solutions[68] to protect genetic data and promote data sharing.[69] While not all technical solutions will necessarily be applicable to genetic data,[70] many hold great promise. In this section, we will consider two prominent examples: differential privacy and secure computation.

Differential Privacy

Differential privacy is a mathematical concept that allows researchers to measure whether they can derive conclusions from a dataset while being unable to determine whether or not those conclusions are based on any individual's personal data.[71] The basic idea of differential privacy is that "the risk to one's privacy...should not substantially increase as a result of participating in a statistical database."[72] Differential privacy requires that the answer to any query be "probabilistically indistinguishable" from the original data with any one specific individual removed—i.e., a differentially private

[66] Vivien Marx, 'Biology: The Big Challenges of Big Data' (2013) 498 Nature 255.

[67] Dan Mangan, 'Health Data Breaches: What Do You Have to Lose?' *CNBC* (9 March 2016) <http://www.cnbc.com/2016/03/09/as-health-data-breaches-increase-what-do-you-have-to-lose.html> accessed 7 April 2016.

[68] *See* An Introduction to Privacy Engineering and Risk Management in Federal Systems, NIST, (Privacy engineering, according to NIST, means, "a specialty discipline of systems engineering focused on achieving freedom from conditions that can create problems for individuals with unacceptable consequences that arise from the system as it processes PII.") <https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8062.pdf>.

[69] Robinson, Jill O., et al. (n.6); See, for example, DNATX,< https://www.dnatix.com/faq/> accessed 23 September 2019.

[70] For example, blockchain technology, involving cryptographically secure distributed ledgers, has received great attention in recent years for its ability to allow permission-less, or "trust-less" exchanges of non-personal data or goods. However, some aspects of blockchain – e.g. the permanent, public, and de-centralized nature of data storage – may not be appropriate for data that is very personal, challenging to de-identify with confidence, or requiring accountability, credentialing, and oversight. As a result, blockchain technologies are not appropriate for genomic data. Additionally, blockchain applications remain mostly theoretical in the medical sector today. It remains to be seen whether future iterations of blockchain (e.g. "third generation" or "synchronous ledger technologies") may be useful for future solutions. See: Steve Wilson, 'Blockchain, Healthcare and Bleeding Edge R&D' (Constellation Research Inc., 10 October 2016) <https://www.constellationr.com/blog-news/blockchain-healthcare-and-leading-edge-rd> accessed 5 January 2018; & Stephen Wilson and David Chou, 'How Healthy Is Blockchain Technology?' [2017] HIMSS AsiaPac17, Singapore.

[71] Salinger Privacy (n 19).

[72] Cynthia Dwork, 'Differential Privacy' in Michele Bugliesi and others (eds), *Automata, Languages and Programming*, vol 4052 (Springer Berlin Heidelberg 2006) <http://www.springerlink.com/content/383p21xk13841688/> accessed 23 December 2011.

computation provides a randomized output that follows an almost identical probability distribution to the original with any one individual removed—meaning that the answer assumes that even if an adversary knows all the records except one in a dataset, the adversary still could not infer the information in the unknown record.[73] Differentially private solutions typically involve data perturbation to reach such conclusions. Although privacy laws and regulations concern themselves with identity disclosure, differential privacy also protects against learning something new about data subjects through attributes assigned to or inferred about them.[74]

How close the randomized output is to the original is determined by a privacy parameter $\varepsilon > 0$. Lower values of ε ("epsilon") imply stronger privacy guarantees, because more noise is added, making it more difficult to distinguish an individual's contribution to the dataset; higher values of ε imply weaker privacy guarantees. The appeal of differential privacy is the theoretical formulation on which it is based, allowing for formal, mathematical proofs of privacy protection or loss. The practical challenge with differential privacy is determining an appropriate parameter ε that will provide sufficiently low privacy risk and sufficiently useful and reliable data.[75] As there currently is no probabilistic measure of risk or a risk-based framework incorporating contextual parameters for ε, it is difficult to know how ε will meaningfully fit with current standards and guidelines.

Although an active area of research, there have been few real-world implementations of differentially private genomic data, and thus few examples of useful, valid, and safe results on which we can rely as use cases or examples. One example of how differential privacy is applied today is Google's Chrome browser, which collects user statistics (which users have opted-in to share) using differentially private randomized responses.[76] Another is a US Census Bureau application using synthetic data that plots a modified form of differentially private worker commute patterns.[77] Apple has also developed an opt-in differential privacy technique for macOS and iOS 10 users to gain insights into user

[73] Tianqing Zhu, *Differential Privacy and Applications*, vol 69 (1st edn, Springer 2017) <//www.springer.com/gp/book/9783319620022> accessed 5 January 2018.

[74] John M. Abowd and Lars Vilhuber, 'How Protective Are Synthetic Data?' in Josep Domingo-Ferrer and Yücel Saygın (eds), *Privacy in Statistical Databases* (Springer Berlin Heidelberg 2008) <http://link.springer.com/chapter/10.1007/978-3-540-87471-3_20> accessed 18 January 2016.; Josep Domingo-Ferrer and Jordi Soria-Comas, 'From T-Closeness to Differential Privacy and Vice Versa in Data Anonymization' (2015) 74 Knowledge-Based Systems 151.

[75] F. Dankar and K. El Emam, 'Practicing Differential Privacy in Health Care: A Review' (2013) 5 Transactions on Data Privacy 35; Chris Clifton and Tamir Tassa, 'On Syntactic Anonymity and Differential Privacy' (2013) 6 Trans. Data Privacy 161. http://www.tdp.cat/issues11/tdp.a124a13.pdf.

[76] Úlfar Erlingsson, Vasyl Pihur and Aleksandra Korolova, 'RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response', *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (ACM 2014) <http://doi.acm.org/10.1145/2660267.2660348> accessed 18 January 2016.

[77] A. Machanavajjhala and others, 'Privacy: Theory Meets Practice on the Map', *IEEE 24th International Conference on Data Engineering, 2008. ICDE 2008* (2008). http://www.cse.psu.edu/~duk17/papers/PrivacyOnTheMap.pdf

habits by employing differential privacy locally on the users' device before transmitting the data to Apple over an encrypted channel.[78]

While differential privacy may be a promising solution, it poses some challenges for genetic data. In applying differential privacy to protect against both identity and attribute disclosure, the data or results may be overly perturbed when applied to raw genetic data,[79] which can potentially weaken the usefulness of the data and blur the line between disease and health if done improperly.[80] Further, perturbation of genetic data can lead to a lack of trust as it is not possible to guarantee that nonsense data will not be outputted. Distorted genetic data that is outputted could thwart research results or even lead to misdiagnosis, as minute changes in nucleotide bases can lead to critical errors in gene analysis. Troublingly, a study incorporating genetic data considered the impact of using differentially private pharmacogenetics models to guide personalized warfarin dosing, based on genetic data and medical history, and found that there would be worse clinical outcomes for a significant number of patients in simulated clinical trials.[81]

Another challenge is that genetic data would need to be held in extremely large files with considerable amounts of noise, making it difficult to deploy efficient and scalable differential privacy protections. It also has been argued that the main factor of differential privacy, ε, cannot be strictly defined to a specific value for all purposes. The ε value represents a tradeoff between privacy and utility that would need to be determined by statistical experts on a case-by-case basis, given a variety of contextual factors (such as data type, research purpose, etc.). Further, it could be difficult to describe the context of differential privacy and how the process would enable the secure disclosure of their genetic data to research participants and patients who are considering providing an informed consent to share their data. These issues point to its current impracticality with genetic data.

Secure (Multi-Party) Computation

Secure (multi-party) computation is a type of cryptography allowing multiple parties, who want to pool data and compute a function without inter-party disclosures, to collaborate on fully encrypted data. The technique is theoretically thought of as the equivalent of sending encrypted data to a trusted third-party who would return the desired result without the need to decrypt the data, guaranteeing minimal information leakage. The

[78] Apple Inc., 'Apple Differential Privacy Technical Overview' <https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf>.

[79] Zaobo He, Yingshu Li and Jinbao Wang, 'Differential Privacy Preserving Genomic Data Releasing via Factor Graph', *Bioinformatics Research and Applications* (Springer, Cham 2017) <https://link.springer.com/chapter/10.1007/978-3-319-59575-7_33> accessed 12 December 2017.

[80] Lowrance and Collins (n 47).

[81] Matthew Fredrikson and others, 'Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing' (2014) <https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew> accessed 18 January 2016.

cryptographic primitives, or building blocks, to create secure computation protocols can come from a variety of techniques– including homomorphic encryption, garbled circuits, secret sharing, or others.

With appropriate contractual obligations and measures to ensure there are no leakages of personal information from the computational results themselves,[82] secure computation can be thought of in a risk-based framework as *protected* pseudonymous data[83] with a low risk of re-identification.[84] Secure computation is consistent with guidance provided by regulators as a means to protect personal health information while sharing encrypted data for the purposes of collaborative analysis.

There are very few real-world examples of secure computation, let alone in genetics. Linking for database matching or deduplication without sharing sensitive or personal information, known as secure linking,[85] is used by the Institute for Clinical Evaluative Sciences (ICES) for linking de-identified data (matching on insurance number, name, and date of birth) and was proposed for a human papillomavirus (HPV) vaccine initiative impact assessment.[86] Using a secure data collection system, which provided strong privacy and confidentiality assurances, a point prevalence study was conducted in 2014 to assess rates of antimicrobial resistant organisms (ARO) in long-term care homes in Ontario.[87] Another 2015 study in Estonia was conducted by linking a tax database and higher-education database to assess correlations between graduation from higher education and employment during that study period using secure computation.[88]

One notable pilot study from 2016 used secure computation in genetics for DNA-based prediction of HIV-related outcomes.[89] This study demonstrated some interesting points related to real-world use and interpretation of genetic data. Some variants provide clear, nearly deterministic results, whereas others are used in combination to determine genetic risk scores. In order to be useful, the authors found that it was necessary to

[82] Christine O'Keefe and James Chipperfield, 'A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems' (2013) 81 426.
[83] Pseudonymization is defined as a "specific kind of transformation in which the names and other information that directly identifies an individual are replaced with pseudonyms. Pseudonymization allows linking information belonging to an individual across multiple data records or information systems, provided that all direct identifiers are systematically pseudonymized." See: Simson Garfinkel (n 5).
[84] Luk Arbuckle and Khaled El Emam, 'Practical Applications of Secure Computation for Disclosure Control', Proceedings of Statistics Canada Symposium (2016).
[85] Khaled El Emam and Luk Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started* (O'Reilly 2013) ch Secure Linking."
[86] K. El Emam and others, 'A Protocol for the Secure Linking of Registries for HPV Surveillance' (2012) 7 PLoS ONE.
[87] Khaled El Emam and others, 'Secure Surveillance of Antimicrobial Resistant Organism Colonization or Infection in Ontario Long Term Care Homes' (2014) 9 PLoS ONE e93285.
[88] Dan Bogdanov and others, 'Students and Taxes: A Privacy-Preserving Social Study Using Secure Computation', *Cryptology ePrint Archive* (2015).
[89] Paul J. McLaren and others, 'Privacy-Preserving Genomic Testing in the Clinic: A Model Using HIV Treatment' [2016] Genetics in Medicine <http://www.nature.com/gim/journal/vaop/ncurrent/full/gim2015167a.html> accessed 11 February 2016.

accommodate both types of tests. Also, as interpreting genotyping results requires specific knowledge and training, the authors not only sought to extract relevant test data securely, but to provide meaningful reports to clinicians.

In 2017, a secure computation was performed on genomic data held in the cloud by a group of researchers at Stanford University in 2017.[90] Using a cryptographic "genome cloaking" technique, researchers were able to determine which patients at two medical centers with similar symptoms shared gene mutations and identify which gene mutation was responsible for four rare diseases, among other accomplishments. These analyses were done while keeping more than 97% of the genetic data completely encrypted.

Duality Technologies, a leading provider of privacy enhancing technologies, is currently attempting to further mature homomorphic encryption, the process by which data may be analyzed without the need for decryption.[91] While Duality initially focused on business transactions, their solutions scale linearly and could practically support millions of SNPs. In 2018, Duality co-won the iDash competition, an NIH funded event, for fastest computations on genetic data. Duality's run time on this challenge was 0.09 minutes leveraging 1.5GB of memory. For comparison, fellow competitor IBM ran for 23 minutes leveraging 8.6GB of memory.[92] Innovation in the arena of homomorphic encryption could potentially allow for deployment on cloud-based computing clusters, delivering low-cost, large-scale, and privacy-focused solutions.

IQVIA, a prominent health research company, launched E360 Genomics in 2019, a patented platform that allows clients to access aggregated data for research purposes.[93] IQVIA's techniques allow for the removal of phenotypic data while retaining genotypic information, as a methodology for protecting patient privacy while still allowing for great utility in the data itself.[94] IQVIA's process tokenizes patient level genetic variants consistently across all data sets, which preserves the data values and enables statistical analysis to be done on the tokens rather than the information itself. After analysis, variant tokens of interest can be de-tokenized for downstream analysis. While IQVIA's market is for research applications, the same tokenization technology could be applied to clinical applications.

However, while secure multi-party computation is advancing rapidly with more efficient and scalable methods,[95] as well as specialized hardware to accelerate computations,[96]

[90] Karthik A. Jagadeesh and others, 'Deriving Genomic Diagnoses without Revealing Patient Genomes' (2017) 357 Science 692.

[91] *See* Duality Technologies generally <https://duality.cloud/about-us/> accessed 24 September 2019.

[92] *See* iDash Competition Announcement <https://duality.cloud/duality-wins-idash-competition-fastest-computations-genomic-data/> accessed 24 September 2019.

[93] *See* IQVIA generally <https://www.iqvia.com/> accessed September 24 2019.

[94] See IQVIA's E360 Genomics Launch <https://www.iqvia.com/newsroom/2019/03/iqvia-launches-e360-genomics> accessed September 24 2019.

[95] Kurt Rohloff and David Bruce Cousins, 'A Scalable Implementation of Fully Homomorphic Encryption Built on NTRU' in Rainer Böhme and others (eds), *Financial Cryptography and Data Security* (Springer Berlin

current techniques are still difficult to implement and require longer computation times than conventional methods, which poses practical concerns for its applicability in genetics. As sharing of genetic data is needed for quick and accurate research, diagnosis, and treatment, adding friction to the sharing process through secure computation could slow down progress. Further, the storage of encrypted data requires a tremendous amount of memory, which can be extremely costly (a single sequenced human genome can require up to 300GB of storage).

IV. A PRACTICAL PATH FORWARD

Although differential privacy and secure multi-party computation show promise for protecting genetic data, the current applicability of these technologies is not yet fully developed. While these privacy tools may provide protections for genetic data that de-identification of the accompanying personal data cannot, they have costs, both in decreased data utility and added resource burdens, limiting their scalability at the present time. Further, as the amount of genetic data increases exponentially, adding the technical burden of these methods could ultimately slow or inhibit sharing and usage of genetic data.

While advanced technical solutions to the privacy risks posed by genetic data remain mainly in the research and development stage, in practice, organizations are currently focusing on privacy policies and data practices that outline privacy controls for genetic data (including, but not limited to de-identification), while treating genetic data as not readily identifiable. Leading organizations—such as the NIH,[97] International Cancer Genome Consortium (ICGC),[98] Wellcome Trust Sanger Institute,[99] UK BioBank,[100] and The

Heidelberg 2014) <http://link.springer.com/chapter/10.1007/978-3-662-44774-1_18> accessed 27 October 2015. ibid.

[96] David Cousins and others, 'SIPHER: Scalable Implementation of Primitives for Homomorphic Encryption' (Raytheon BBN Technologies 2015) Final Technical Report AFRL-RI-RS-TR-2015-252. http://www.dtic.mil/dtic/tr/fulltext/u2/a624640.pdf.

[97] The NIH funds a large amount of the health research conducted in the US, including genomic research. Researchers funded by the NIH are obligated through their funding agreements to share their data for secondary research purposes (See: National Institutes of Health, 'FINAL NIH STATEMENT ON SHARING RESEARCH DATA' (2003) <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>.). In order to protect the privacy of research subjects, the NIH has developed a comprehensive set of policies and procedures around genomic data sharing as outlined in their widely known Genomic Data Sharing Policy (GDS). See: National Institutes of Health, 'NIH Genomic Data Sharing (GDS) Policy' <https://gds.nih.gov/03policy2.html>.4/1/19 10:07:00 AM

[98] ICGC was launched in 2008 to coordinate a large number of cancer genome studies taking place around the world with an aim to gain a more complete understanding of the genomic changes related to individual cancers. ICGC shares genomic data with researchers for secondary studies in order to maximize the benefit to the public and to increase efficiency and productivity in research. See: International Cancer Genome Consortium, 'ICGC Goals, Structure, Policies and Guidelines' <https://icgc.org/icgc/goals-structure-policies-guidelines> accessed 6 April 2016.

[99] The Wellcome Trust is a charity in the UK which funds medical research. They are a leading advocate for open access to research data and organizers of the Fort Lauderdale meeting in January 2003 which solidified the research community's commitment to rapid release of genomic data. See: The Wellcome

Cancer Genome Atlas (TCGA)[101]—also highlight that de-identification practices of the accompanying personal data should not be used alone, but rather combined with a comprehensive privacy program and governance mechanisms. These governance mechanisms include access controls, data use agreements, and strong security protocols throughout the information management lifecycle that minimize the risk of re-identification of genetic data. Although there is some variability across organizations in the details of their privacy policies and how they have been structured, the strong similarities between them demonstrate a convergence of norms (see chart 1 of the Appendix for more details):

1. **Access Controls**: Access to genetic data is tiered depending upon the nature of the data and accompanying risk of re-identification. Data placed in open access repositories or on the web is publicly available without restriction, but the majority of such data is shared in aggregate or summary form, not at an individual level. Data in controlled access repositories is typically restricted to research use only, with research requests pre-approved by the organization prior to access being granted. Access to and use of controlled access data is subject to reviews by a governing body, such as a Data Access Committee.[102] Such reviews assess the requests for merit and/or conformity to any use limitations attached to the data. Governing bodies may provide authenticated users with broad access to the controlled access data or grant access to specific controlled access data sets based on the details of the proposed research (on a study-by-study basis).

2. **Contractual Controls:** Once approved for use of controlled access data, investigators and their institutions are required to enter into a data use agreement prior to data access being granted to ensure that data is accessed only for legitimate purposes and to help mitigate future re-identification risks. These

Trust, 'Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility' <https://cancergenome.nih.gov/abouttcga/policies/wellcome-trust-fort-lauderdale-principles> accessed 9 November 2017.; The Wellcome Trust Sanger Institute (WTSI), dedicated specifically to genomic research, has produced guidelines pertaining to the sharing of genomic data generated in their research projects. Access, ethical considerations, rights of data providers, and optimizing translation are the key principles that underlie the guidelines. See: Wellcome Trust Sanger Institute, 'Data Sharing Policy and Guidelines' <http://www.sanger.ac.uk/sites/default/files/Jul2017/Data_Sharing_Policy_and_Guidelines_July_2017_0.pdf> accessed 20 November 2017.

[100] UK Biobank is a project, funded in part by the Wellcome Trust, which recruited 500,000 people in the UK to provide samples and detailed health information, and agree to have their health followed over time. The Biobank is designed to be a resource for researchers to study disease etiology. See: 'About UK Biobank' (*UK Biobank*) <http://www.ukbiobank.ac.uk/about-biobank-uk/> accessed 13 December 2017.

[101] The Cancer Genome Atlas is a joint effort of the US National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to map critical genomic changes found in different types of cancer. The TCGA is committed to open sharing of their data with researchers with an aim to improve diagnosis and treatment of the disease. See: 'About TCGA' (*The Cancer Genome Atlas - National Cancer Institute*) <https://cancergenome.nih.gov/abouttcga> accessed 13 December 2017.

[102] 'NIH GDS Policy Oversight' (*Office of Science Policy*) <https://osp.od.nih.gov/scientific-sharing/policy-oversight/> accessed 16 January 2019.

agreements outline the appropriate uses of the data and the obligations of investigators, including the following provisions: restrictions on the re-distribution of the data; regular reviews and/or renewal of access authorizations; prohibition of attempts to re-identify the data and from contacting data subjects, and requirements to destroy data upon study completion. Data use agreements may create additional obligations in accordance with the terms of consent given by research participants, the characteristics of the particular data being accessed, and/or other institutional policies and procedures.

3. **Security Protocols**: Investigators using data in open access repositories and approved users of controlled-access data adhere to the security protocols specified by the respective organizations sharing genetic data. All organizations note that local storage is more secure than cloud storage of sensitive data, although there are risks associated with data stored on local management systems. The organizations variously address these risks by requiring the implementation of both physical and electronic security methods on devices, such as providing user training, restricting access to authenticated users, and encrypting data.

The trend to provide additional safeguards for genetic data also is increasingly recognized by entities that are not subject to the HIPAA Privacy Rule, such as consumer genetic testing companies (commonly referred to as direct-to-consumer genetic testing companies), which are recognizing de-identification challenges and taking proactive steps to reduce re-identification risk. Although consumer genetic testing companies are not covered entities or business associates, many nevertheless have committed to requiring safeguards in addition to de-identification of accompanying personal data, including limits on sharing without consent unless there are strong assurances that the data is not identifiable and protected from re-identification.[103] Further, rather than relying exclusively on technical solutions that may significantly reduce a dataset's utility for research or other secondary analysis, complementing less severe technical controls with legal and administrative safeguards, such as an ethical review process,[104] also can help

[103] See: Future of Privacy Forum, 'Privacy Best Practices for Consumer Genetic Testing Services' <https://fpf.org/wp-content/uploads/2018/07/Privacy-Best-Practices-for-Consumer-Genetic-Testing-Services-FINAL.pdf>.; The Best Practices recognize that due to the unique nature and sensitivity of genetic data, today's current de-identification techniques for genetic data alone cannot be represented as strongly protecting individuals from re-identification. It also recognizes that for genetic data, de-identification requires particular care given challenges of de-id this type of data. In light of this, the Best Practices requires that the de-identification measures taken establish strong assurance that the data is not identifiable and protected from re-identification. The Best Practices recognizes that this can be accomplished through de-identification in addition to the use of strong security protocols, encryption, contractual restrictions on sharing and use, and retention separate from or without matching datasets.

[104] An ethical review process (also referred to as a corporate IRB, Consumer subject review board, or corporate ethics boards) has been proposed for the ethical review for data that is not typically covered by the Common Rule. The purpose of these reviews is to identify both the risks and the benefits of the research and to balance the prospective risks to the Consumer, prospective benefits to Consumers or to the public, the rights and interests of the Consumer, and the legitimate interests of the company. See: Ryan Calo, 'Consumer Subject Review Boards: A Thought Experiment' (2013) 66 Stanford Law Review Online 97.;

maintain a balance between preserving the data's value and protecting individual privacy.

CONCLUSION

The field of human genetics is evolving rapidly. While this evolution is advancing biomedical science and informatics, it is also raising serious privacy concerns. As entities such as hospitals, research organizations, corporations, government agencies, and biobanks continue to publish "de-identified" genetic data (i.e. genetic data where the accompanying personal data has been de-identified), balancing genetic data sharing with privacy will become a challenge that we cannot ignore. While de-identification of the personal data that accompanies genetic data has been the most common practice for protecting genetic data, de-identification is a moving-target—data that could not be linked back to an individual at the time of its release could become identifiable over time, as new datasets and new re-identification techniques become available.

The combination of genetic data proliferation, its inherent identifiability, and advancements in the statistical and technological methods for re-identification indicate that methods like the HIPAA Safe Harbor Method may not be sufficient to protect privacy. Re-identification of genetic data may be more readily available than we think, and risk of re-identification may be larger than we currently consider. In light of recent re-identification studies, it is becoming increasingly difficult to maintain the stance that genetic data by itself and without accompanying personal data is not readily identifiable and that de-identification of that data alone is sufficient to ensure that genetic data won't be re-identified. Further, de-identification of accompanying personal data is only one of many tools to safeguard genetic data— and while it is an important piece of an overall privacy program, de-identification alone may not be a sufficient answer to protect the privacy of genetic data.

Given the rapidly evolving re-identification landscape for genetic data and the unique privacy implications that it creates, we must continue to research and develop new technical solutions. Differential privacy and secure computation promise novel solutions to the problem of how to safeguard genetic information and reduce the re-identification risk it poses to individuals. However, as it stands, neither of these solutions have been used much, if at all, in practice. As privacy norms in the space of genetic data sharing evolve, along with advances in genetics and the availability of tools to interpret and re-identify genetic data, technical privacy protections will need to advance beyond the lab and into practice.

While we await those technical developments, strong governance mechanisms are needed to lay the path forward for privacy-protective genetic data sharing. As leading genetic-data-sharing organizations' privacy and data protection policies demonstrate,

'"Beyond IRBs: Ethical Guidelines for Data Research" by Omer Tene and Jules Polonetsky' <https://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/7/> accessed 17 October 2018.

there appears to be a common consensus - non-public access to genetic sequences is unlikely to lead to the re-identification of individuals when: (1) the data is protected by appropriate access controls, data use agreements, and strong security protocols; and (2) accompanying personal data is de-identified. Through rigorous technical, legal, and organizational controls in addition to de-identification of personally identifying information accompanying genetic data, re-identification risk can be lowered to a degree that permits reasonable sharing of genomic data.

Appendix

Chart 1: Comparison of Genetic Data Sharing Policies from Leading Organizations

Organization / Policy	De-identification[105]	Access Controls	Data Use Agreements	Security Protocol
The National Institutes of Health (NIH) ***Genomic Data Sharing Policy***[106]	Investigators should de-identify human genomic data prior to submitting it to NIH-designated data repositories, in accordance with both the HHS Regulations for Protection of Human Subjects and the HIPAA Privacy Rule. The term de-identify, as it is being used in the NIH GDS policy, refers to removing information that could be used to associate a dataset or record with a human individual.[107] The GDS policy refers to removing "identifiers" that could lead to disclosure, suggesting that the NIH currently considers genomic	Data submitted to NIH-designated data repositories is available for secondary research use through either unrestricted (available to the public at large) or controlled access (available only for particular projects by investigators approved and monitored by Data Access Committees (DACs)). Data is generally unrestricted only if collected under informed consent for future research use and broad data sharing. However, for "compelling scientific reasons," the NIH may allow unrestricted use of genomic data collected without such consent. Controlled-access datasets are defined by the data use limitations established by the	Approved users of controlled-access data are encouraged to obtain a CoC as a precaution against the re-identification of de-identified data. CoCs prohibit the disclosure of identifiable, sensitive information about subjects to those not connected to the research.[108] For unrestricted data, investigators should not attempt to identify individual participants from whom the data was obtained. In addition, all investigators must sign a Data Use Certification, which stipulates additional limitations on the use of data, including: (1) Using the data only for the approved research	Approved users of controlled-access data should adhere to best practices for security. Such recommended practices for investigators include consulting with their institutions' respective IT officers to develop a security plan prior to receiving the restricted data, implementing physical and electronic security measures on devices, training users, and taking additional measures for the use of cloud computing.[109]

[105] Note that the terms "de-identification" and "anonymization" are used interchangeably in this discussion as the North American organizations included favor the former term and the European organizations the latter.

[106] National Institutes of Health, 'NIH Genomic Data Sharing (GDS) Policy' (n 88).

[107] We assume that the term "de-identify," as used in policies other than NIH GDS policy, is meant to be applied to any clinical or identifying personal data accompanying the genomic data (such as names, identification numbers, and other personal information) and not the genomic data itself as no standard de-identification methods exist for genomic data at this time.

[108] National Institute of Health (n 42).

[109] National Institutes of Health, 'NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy' <https://www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf>.

	data not to be identifying. However, the NIH has obtained a Certificate of Confidentiality (CoC) for its database of genotypes and phenotypes (dbGaP) and encourages investigators submitting large genomic datasets to do the same because "genomic data can be re-identified."	institution submitting the data. A DAC's decision to approve access is based on whether the proposed use of data conforms to the data use agreement of the submitting institution.	(2) Not selling any data or sharing it with individuals not listed in the data access request (3) Agreeing to report any violation of GDS policy to the DAC once discovered	
International Cancer Genome Consortium (ICGC) ***ICGC Goals, Structure, Policies & Guidelines***[110]	ICGC policy recognizes the possibility that de-identified genomic data can be indirectly re-identified, if similar data from the individual were obtained from a third-party database containing sufficient demographic or healthcare information. Open access data is said to be "permanently de-identified," although the policy does not specify exactly what data is to be de-identified or how. According to the ICGC access policy, the open-access data has been carefully considered and will be monitored to ensure it cannot presently be used to	Like the NIH, the ICGC distinguishes between open and controlled-access datasets. Open access datasets are accessible to the public and monitored to ensure that they cannot be aggregated to create a dataset unique to an individual without "reasonable efforts." They contain demographic information, clinical data such as vitals and disease status, and normalized data on gene expression. Controlled-access datasets contain data unique to individuals not directly identified, and may include raw genetic data, including probe-level data on gene expression and whole genome sequence files.	Investigators seeking access to open datasets must submit Assurance Agreement forms that include: (1) a written description of the purpose of the research to be done; and (2) an agreement to not try to identify or contact the donor subjects. In addition to the provisions above, investigators seeking access to controlled datasets must submit Assurance Agreement forms agreeing: (1) not to redistribute the datasets; (2) to destroy the datasets upon termination of their use; and (3) to protect patient confidentiality. However, individual data producers will be responsible for the protection of	The ICGC requires the establishment and use of a Data Coordination Center (DCC), which manages the integration and distribution of data submitted by ICGC participants and performs high-level quality control checks on the data. The system restricts access to protected data to authenticated and authorized investigators. The Center utilizes local franchise databases, which contain locally-stored information relating to a specific research project. The information is periodically stored in a public repository, a coordination backend that hosts all of the ICGC data and creates a uniform model to be used by researchers with access to the DCC.

[110] International Cancer Genome Consortium (n 89).

	generate a dataset unique to an individual without reasonable efforts.	The Data Access Compliance Office (DACO) processes requests for the use of controlled-access data in particular research projects. Projects must comply with ICGC policies developed by the International Data Access Committee (IDAC).	confidential information they submit.	
Wellcome Trust Sanger Institute (WTSI) ***Data Sharing Policy and Guidelines; Human Data Security Policy***[111]	Researchers applying for funding should incorporate anonymization procedures into the design of a study in order to ensure the privacy of subjects is properly safeguarded. WTSI stipulates that research data sets should be pseudonymized or fully anonymized in all but exceptional circumstances. Anonymization in this instance refers to the removal of "information which allows identification of an individual" from clinical or other accompanying data. WTSI provides some guidelines for anonymizing data, such as removing all but the first three digits of postal codes, generalizing	Projects involving genetic data must be assigned one of four data security levels. The requirements of each level are cumulative. Level 1 ("Open") – participants have given informed consent to make their data publicly available without restriction, OR data has already been publicly disclosed or could be. Level 2 ("Standard") – Data should typically be linked only to general demographic or phenotypic data and should be drawn from a relatively large population. Data may be transmitted between members of the access control group through a private channel (e.g., USB, not e-mail). Level 3 ("Strong") – This includes genetic data that is particularly sensitive or poses a	There are no restrictions on use of Level 1 genetic data. Level 2 – Investigators must make no attempts to re-identify participants. They must also destroy data after it is no longer needed. Level 3 – In addition to the restrictions on the use of Level 2 data, each member of the access control group must be expressly authorized to access the data by a principal investigator (PI), and changes to group membership must be approved and reviewed at least once every 6 months by a PI. Level 4 - In addition to the restrictions on the use of Level 2 and 3 data, transmission of unencrypted data between systems, even between members of	There are no security requirements for Level 1 genetic data. Level 2 – Investigators must limit access to systems used to process or store unencrypted data to authorized individuals and WTSI systems administrators. Data stored on systems without such access control restrictions must be encrypted, and decryption keys must be handled securely. Level 3 – Encryption systems should be "as secure as practically possible." Level 4 – Systems used to process or store unencrypted data must either isolate all projects or limit access to the entire system only to members of the access control group for a single project. Portable systems may not be used to store unencrypted genetic data

[111] Wellcome Trust Sanger Institute, 'Data Sharing Policy and Guidelines' <http://www.sanger.ac.uk/sites/default/files/Jul2017/Data_Sharing_Policy_and_Guidelines_July_2017_0.pdf> accessed 20 November 2017; The Wellcome Trust, 'Developing an Outputs Management Plan' (*Wellcome*, 2017) <https://wellcome.ac.uk/funding/managing-grant/developing-outputs-management-plan> accessed 12 September 2017; Wellcome Trust Sanger Institute, 'WTSI Human Data Security Policy' <https://www.sanger.ac.uk/legal/assets/wtsi-hgdsp-201510hpfinal.pdf>; The Wellcome Trust (n 90).

	dates to year only, and removing unique identifiers such as health insurance numbers; however, WTSI also stipulates that anonymization is context-specific and thus these steps may not sufficiently protect every dataset.[112] As a result, WTSI requires that before research data (such as summary statistics of genomic data) is made openly accessible, the risk of re-identification be assessed by researchers seeking to publish the results of a study using such anonymized data.	greater risk of re-identification. It also includes data that has had consent given by a participant explicitly requiring stronger security. Each member of the access control group must be expressly authorized to access the data. Authorization must be granted by a principal investigator and membership reviewed at least once every 6 months. Level 4 ("Personal") – This includes genetic data that contains "personal data" including name, address, email, and other information that could lead to re-identification if coupled with other information "reasonably available" to those accessing the data.	the control access group, is not permitted	unless they are physically locked and secured. The system must also use a higher-level authentication system, such as two-factor authentication.
UK BioBank ***Summary De-Identification Protocol; Data Management & Sharing Plan***[113]	Data must be de-identified in-house by UK BioBank before it is released it to researchers. The UK Biobank takes the stance that genomic data, such as "genetic	Like the NIH, ICGC, and WTSI, UK Biobank distinguishes between publicly accessible and controlled-access datasets. The Biobank Data Showcase makes publicly available genetic summary data, but individual-level anonymized data is only	Publications by researchers must comply with BioBank's de-identification protocol and cannot contain information that could lead to re-identification of an individual.	A limited number of Biobank staff have access to systems that could be used by an adversary for reverse de-identification and subsequent re-identification of a participant, and such systems are regularly audited.

[112] Data is considered to be either low risk (statistical figures like p-value, z-score, confidence intervals), moderate risk (allele frequency), or high risk (genome-wide linkage disequilibrium measures). Genomic summary statistics contains at least low-risk data. Decisions about publishing moderate and high-risk data are context-specific determinations. For example, re-identification risk is high when the dataset contains rare alleles.

[113] UK Biobank, 'Summary De-Identification Protocol' <http://www.ukbiobank.ac.uk/wp-content/uploads/2013/10/ukbiobank-summary-de-identification-protocol.pdf> accessed 19 September 2017; 'About UK Biobank' (n 91); UK Biobank, 'Access Procedures: Application and Review Procedures for Access to the UK Biobank Resource' <http://www.ukbiobank.ac.uk/wp-content/uploads/2012/09/Access-Procedures-2011.pdf> accessed 18 September 2017.

	sequence data," is not readily identifiable and therefore poses little privacy risk. Specifically, Biobank states that the practical risk of re-identification posed by genetic sequence data is small because of its "virtual obscurity", i.e., an adversary would need a reliable reference sample that identifies that participant. However, Biobank notes that their policies are subject to change as technology improves and genetic data coupled to demographic information becomes increasingly publicly accessible. In terms of clinical data, the UK Biobank only holds participant data in what it calls "reverse anonymized form." This is essentially pseudonymous data from which names have been removed and health numbers encrypted in a reversible manner. Specifically, the anonymization is only reversible by Biobank through its internal database, and not by the	available to approved researchers. The Access Sub-Committee (ASC) of the Board of UK Biobank is responsible for reviewing and granting access to protected data. Researchers must demonstrate that the data will be used for "public good" and otherwise comply with the requirements described in Access Procedures.	However, if researchers request more detailed information pertaining to their study (such as participants' geographical locations) which may increase the risk of re-identification, Biobank may elect to link the anonymized date back to the participant and return the requested data to the researcher. Unlike the genomic data sharing policies of the other leading organizations, there is no time limit on the retention of protected data. Upon termination of use, data need not even be "destroyed" – only rendered inaccessible for further use.	Researchers granted access to data must provide information about their institution's data security systems and protocol, and must agree that Biobank has the right to audit their security systems.

	accessing researcher. Data to which researchers are given access is further de-identified according to the UK Biobank's summary de-identification protocol. Generally, individuals' dates of birth and locations are generalized, and the UK Biobank does not release unedited free text, general practitioner details and healthcare location information.			
The Cancer Genome Atlas (TCGA) ***Data Sharing and Data Management; Data Access Policies***[114]	TCGA receives data from research sites in HIPAA Limited Data Set (LDS) format, which has some identifiers removed but is still considered to be protected health information. However, all data to which researchers are given access is de-identified according to the HIPAA Safe Harbor standard. Data collected from non-US sites is also treated in accordance with HIPAA, though not subject to its requirements Although TCGA considers research	Like the NIH, ICGC, and Biobank, TCGA has also implemented a two-tiered data access system, which will be discussed in more detail later. The open-access tier is accessible in public databases and "contain[s] only data that cannot be analyzed to generate a dataset unique to an individual." The controlled-access tier has been implemented to safeguard "data that are associated to a unique, but not directly identified, person," including individual-level SNP variants and whole genome sequence data.	TCGA's data access policy places limitations on all researchers' access to and use of data beyond what is required by OHRP. Controlled-access data is available to qualified researchers who submit Data Access Requests (DAR) and, in conjunction with their respective institutions, certify agreement to the TCGA Data Use Certification (DUC). Under the terms of the certification, approved users agree not to attempt to contact participants or redistribute data to third parties.	As of 2016, TCGA utilizes the National Cancer Institute Center for Cancer Genomics' Genomic Data Commons (GDC) to provide access to aggregated genomic data stored in a cloud computing environment. Genomic data from various projects are pooled into one database, making sharing of data easier, whereas local management systems, though more cumbersome, prevented unauthorized redistribution of data. In addition, the GDC harmonizes data, allowing comparison of different datasets (increasing their utility, but additionally increasing the risk of re-identification).

[114] 'About TCGA' (n 92); The Cancer Genome Atlas, 'Data Use Certification Agreement' <https://cancergenome.nih.gov/pdfs/Data_Use_Certv082014> accessed 9 November 2017; National Institutes of Health, 'NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy' (n 100).

	with de-identified data to fall outside of the scope of human subjects research as outlined in the NIH Office for Human Research Protections (OHRP) “Guidance on Research Involving Coded Private Information or Biological Specimens,” TCGA allows an investigator’s IRB/research review body to determine whether the project constitutes human subjects research.[115]			

[115] The Cancer Genome Atlas Program, ‘Human Subjects Protection and Data Access Policies’ <https://cancergenome.nih.gov/abouttcga/policies/tcga-human-subjects-data-policies>.