Genetic Data Aren't So Special: Causes and Implications of Re-identification

T.J. Kasperbauer & Peter H. Schwartz
Indiana University Center for Bioethics
Indiana University School of Medicine

This is the pre-peer-reviewed version of the article published here:
https://onlinelibrary.wiley.com/doi/abs/10.1002/hast.1183?af=R

Full citation: Kasperbauer, T.J. & Schwartz, P.H. (2020). Genetic data aren't so special: Causes and implications of re-identification. *Hastings Center Report, 50,* 30-39.

Abstract: Genetic information is widely thought to pose unique risks of re-identifying individuals. Genetic data reveals a great deal about who we are, and, the standard view holds, should consequently be treated differently from other types of data. Contrary to this view, we argue that the dangers of re-identification for genetic and non-genetic data—including health, financial, and consumer information—are more similar than has been recognized. Before we impose different requirements on sharing genetic information, proponents of the standard view must show that they are in fact necessary. We further argue that the similarities between genetic and non-genetic information have important implications for communicating risks during consent for healthcare and research. While patients and research participants need to be more aware of pervasive data sharing practices, consent forms are the wrong place to provide this education. Instead, health systems should engage with patients throughout patient care to educate about data sharing practices.

Introduction

Genetic data and biological samples containing genetic material are widely thought to differ from other sorts of health data due to the impossibility of truly "de-identifying" genetic information. The essential idea is that even if the individual's name and other identifying information have been removed, genetic data are still linked back to the relevant individual due to the uniqueness of the genetic code. This idea stands behind numerous proposals regarding the ethics and regulation of using genetic samples and information in many settings, including research, healthcare, and direct-to-consumer testing.

For instance, under current regulations, research involving de-identified data or biospecimens can often be performed without the consent of the people who donated the sample or who the data is about. In 2015, the Department of Health and Human Services proposed changing this policy in its revisions to the Common Rule, to require consent when using de-identified genetic samples and genetic information.[1] More than 1500 public comments objected to this change,

[1] Federal Register, *Federal Policy for the Protection of Human Subjects* (January 19, 2017), https://www.federalregister.gov/documents/2017/01/19/2017-01058/federal-policy-for-the-protection-of-human-subjects#h-3

mostly emphasizing the large negative effect it would have on research.[2] The final version of the Common Rule did not include this change – genetic samples and information may still be used in certain circumstances without the consent of the individual. But the final version of the Common Rule did require that institutions revisit what counts as an "identifiable biospecimen" within a year of implementation and again at four-year intervals.

Concerns about re-identification of genetic information are also widely expressed in arguments for disclosing the risks of re-identification in research consent. For example, Schwab et al. recommend that "researchers should inform potential research participants that they should assume their identity will be discoverable by anyone who gains access to the data from their genetic sample."[3] Similarly, the National Institute of Health's policy on sharing genomic data recommends that research participants be informed that "it may be possible to re-identify de-identified genomic data," but does not make any such recommendation for health data generally.[4] Concerns about genetic re-identification have also made their way into biobank consent forms. For example, the Indiana Biobank states, "There is always a very small chance that someone outside of the Indiana Biobank could identify you based on your genetic information." All major international guidelines on biobanks and data sharing also recommend communicating the re-identification risks of genetic data.[5]

What is striking about this debate is the lack of careful examination of the assumption that the dangers of re-identifying genetic data differ in essential and important ways from re-identifying other sorts of data. Even recent rejections of genetic exceptionalism (the idea that genetic data are unique) have largely accepted that genetic data pose unique risks to re-identification. For example, Garrison et al. note that "concerns about privacy are not unique to genomic data," but nonetheless suggest that genetic information deserves special protection because "Many consider genetic information to be sensitive and private" and there is "a possibility that people could be reidentified with genomic data and some basic demographic information."[6] They briefly consider, but ultimately do not endorse, the idea that health and financial data have a similar potential to re-identify individuals.

In this paper, we reexamine the assumed differences between genetic and non-genetic data and conclude that they are not as significant as many think. Genetic data and other sorts of information are similar in three important ways: First, genetic and non-genetic data can both be

---

[2] Council on Governmental Relations, *Analysis of Public Comments on the Common Rule NPRM* (2016), https://www.cogr.edu/COGR/files/ccLibraryFiles/Filename/000000000346/Analysis%20of%20Common%20Rule%20Comments.pdf

[3] A. P. Schwab et al., "Genomic Privacy," *Clinical Chemistry* 64, no. 12 (2018): at 1702.

[4] National Institute of Health, *NIH Guidance on Consent for Future Research Use and Broad Sharing of Human Genomic and Phenotypic Data Subject to the NIH Genomic Data Sharing Policy* (2018), https://osp.od.nih.gov/wp-content/uploads/NIH_Guidance_on_Elements_of_Consent_under_the_GDS_Policy_07-13-2015.pdf

[5] T. J. Kasperbauer et al., "Communicating Identifiability Risks to Biobank Donors," *Cambridge Quarterly of Healthcare Ethics* 27, no. 1 (2018): 123-136.

[6] N. A. Garrison et al., "Genomic Contextualism: Shifting the Rhetoric of Genetic Exceptionalism," *American Journal of Bioethics* 19, no. 1 (2019): at 60; also see M. Sabatello and E. Juengst, "Genomic Essentialism: Its Provenance and Trajectory as an Anticipatory Ethical Concern," *Hastings Center Report* 49, no. S1 (2019): S10-S18.

unique to individuals. Just as we each have a unique genetic code, so we also have a unique history of surfing the internet, making purchases, and receiving healthcare. Second, both genetic and non-genetic information reveal highly sensitive personal details that we often want to keep private. Third, connecting any unique set of data to an individual requires a "key" linking the two, and such keys are easily obtainable for both genetic and non-genetic data. Commentators frequently fail to appreciate the uniqueness and harmfulness of non-genetic data as well as the availability of keys linking such data to individuals.

We begin by comparing the techniques for re-identifying genetic data to those used for re-identifying other sorts of de-identified information. We use two fictional cases to illustrate the similarities between genetic and non-genetic information, followed by a discussion of relevant empirical research to further flesh out the fictional accounts. We argue that while there are important differences between genetic data and other sorts of data, both the chances of re-identification and the dangers of misuse are more similar than has been recognized. Before we impose special requirements on sharing genetic information, in the context of research, healthcare, or commercial products, proponents of the standard view must show that they are in fact necessary.

We then discuss the implications of our argument for obtaining consent for research and healthcare. Contrary to the common view, we argue that there is no need for separate or special disclosure that highlights the danger of re-identification of genetic data, beyond what is disclosed when collecting other sorts of health information. In short, genetic and non-genetic information should be treated similarly: either warn subjects of the danger of re-identification for both, or don't warn them, for both. In many contexts, we believe a general disclosure of the risks of sharing data is adequate without any specific discussion of re-identification or the distinction between genetic and non-genetic data. Explanations of re-identification of genetic or other sorts of data can be provided to those who wish to learn more, but these explanations do not need to be part of standard disclosure. As we confront the burgeoning market for personal data that exists in our society, we must consider the risks posed by the entire information ecosystem, not just those from genetic information.

Two Cases of Re-identification

While a tremendous amount of research highlights the dangers of re-identification from de-identified genetic information, the process of re-identification may remain opaque to non-specialists. To evaluate the ethical implications of potential re-identification, it is helpful to spell out the techniques that may be used, as we do in the following (fictional) stories.

Genetic Identification

Imagine that one day you receive a letter in the mail from a company called Bed4U advertising their new mattresses. The letter explains that because you have a gene that puts you at risk for back injuries, you should buy one of their specialty mattresses. This is upsetting for many reasons, starting with that you never had genetic testing, as far as you know, and certainly not for anything involving your back. You remember that you provided a tube of saliva to a public ancestry database so they could explore your ancestry and find relatives. The data confirmed that you are mostly of Irish descent, and you learned about some distant relatives. But, as far as you know, you didn't agree to genetic testing or to have your genetic data shared with other

companies, and certainly not ones selling mattresses. You have a vague recollection that you agreed your data could be used for research, but you never thought that such research would involve revealing your identity to private companies.

You call Bed4U to inquire and eventually reach someone who is willing to speak to you about what happened. She explains that finding you, and your genetic risk for back injuries, took a few steps:

First, they purchased a de-identified dataset from the ancestry database you contributed to, which had removed any names and other identifying information. The consent you signed allowed the company to share de-identified information in exactly this way.

Second, they uploaded this data into a different ancestry database that *does* provide identifying information, including location details like city and state, which draws on information that people voluntarily share on 23andme, GEDmatch, and other sites. The company identified all the DNA sequences in the first database that matched a DNA sequence in the second database that had such personally identifiable data.

Finally, the helpful company representative explains, although you did not contribute to the second database, some of your relatives have. In fact, your first cousin who lives on the other side of town contributed his DNA as well as his name. The database also uses algorithms that can accurately infer the approximate locations of close relatives, as well as sex, date of birth, and much else. Using this information, the company inferred that you live in the same town and searched online residential information (e.g., whitepages.com) for the home address of the only other person in town who shares your cousin's last name: you.

That might seem like a lot of work to find you, but the company representative explains that they run these algorithms on all the genetic data they can acquire to identify a large number of people with associated genetic information. The final step, then, was to search these data sets for genes linked to sleep-related problems and products that could be marketed by Bed4U.

This story highlights the way that genetic data appears special, in particular how de-identification of genetic data can fail. But is it so different from other types of information? To answer that question, consider a similar story involving financial information.

Consumer Identification

Imagine now that the letter from Bed4U does not assert that you have a genetic risk for a back injury but instead just refers to your "recent back pain or injury." You had in fact tweaked your back in a pickup basketball game last month, but you didn't even see your doctor, much less provide information to an online retailer about your problem. Once again you call Bed4U, and someone is willing to explain how they identified you as a potential mattress customer. The representative explains that by examining your credit card information, they determined that you recently bought new pillows, marketed specifically to people who have back pain. They also know that you have been searching for a new mattress online, based on your internet browsing activity.

All of this is startling to hear, since you thought your credit card company and internet service providers would keep this information private. The helpful representative explains further: Bed4U got your information from professional data brokers. The data brokers had purchased data sets from your credit card company, internet service provider, insurance company, and dozens of other sources. Many of the datasets were not even legally required to be de-identified. This allowed the data broker to attach your name and contact information to other de-identified information. One data broker even had GPS information from apps you use on your smart phone, providing additional confirmation that you have recently been shopping at mattress stores.

Before you have the chance to protest, they assure you that all the information was obtained legally and that you are not being targeted. They have profiles on everyone, not just you. They also sympathize with your concerns and suggest that a good night's sleep on one of their mattresses will make you feel better.

What makes these cases different?
These two cases present very real possibilities and are based on numerous real-life cases of re-identification, reviewed below. What is different about the two cases with respect to identification? The first case might seem unique because you contributed only genetic information. But closer examination of the process of identification shows that in fact the risk of re-identification from genetic data is not unique.

People already face significant identification risks as a result of systematic data collection from data brokers. Every consumer transaction generates some form of data that, especially in the United States, is relatively easy to combine with other sources of information to build a profile of individuals filled with sensitive personal information. The main challenges are cost, inaccurate data, and technical acumen required to compare large amounts of data from multiple sources.

Contrary to popular belief, and much writing in ethics, the risk of re-identifying individuals from de-identified genetic information does not appear to exceed the risks posed by other types of de-identified data. While re-identification from genetic data is getting easier and easier, it is still on par with the risks of identification from many other types of data we share on a regular basis, both in terms of the likelihood of identification and the negative consequences that can result.

Routes to Identification

Genetic Data
The first case described above is based on recent cases of re-identification from publicly available genomic databases. The information in these databases can come from a variety of sources, including research participants, data provided by individual consumers, and sometimes patient care. The combination of genetic data with pervasive data sharing practices allows anyone with internet access to attach identifying information to public de-identified genetic data.

A study by Erlich, Shor, Pe'er, and Carmi illustrates how the above case might occur.[7] They utilized a technique that has also been used by law enforcement for unsolved crimes, where

[7] Y. Erlich et al., "Identity Inference of Genomic Data Using Long-Range Familial Searches," *Science* 362, no. 6415 (2018): 690-694.

individuals are identified based on genomic data their relatives have contributed to public databases. Erlich et al. took an individual's genome from the 1000 Genomes Project and uploaded it to GEDmatch, a genetic ancestry database. Genetic matches in GEDmatch are publicly available, so without any special access the researchers were able to find a number of the individual's relatives. GEDmatch includes basic demographic information, which allowed the researchers to learn the state of residence for the individual's relatives. With this information, they were able to use publicly available genealogical records to map out the entire pedigree structure of the individual's lineage, including birth dates and rough approximations of where everyone lives, including the original individual.

Erlich et al. conducted a similar analysis on 1.28 million individuals' genetic data held in MyHeritage, a direct-to-consumer genetic testing and ancestry database that until recently permitted access to de-identified data for research purposes. Their results indicated that 60% of Americans of European descent would have at least a third cousin match in GEDmatch. This means that, in principle, 60% of Americans could be re-identified through their relatives' genetic data. They further determined that eventually—perhaps within the next few years—these public databases would contain enough information to identify every single American of European ancestry.

Similar techniques can be used to obtain much more than just contact information. For example, mere membership in certain genomic databases allows inferences to be made about a person's health status. Shringarpure and Bustamente were able to use a person's genetic profile to determine their membership in public "beacon" websites.[8] These websites allow anyone to enter "yes or no" questions about specific information in genomic databases without viewing the information itself. Limiting the types of questions and answers is meant to protect genomic information while also granting access to researchers. For example, you can ask whether a particular nucleotide is present at a specific position on a specific chromosome in the database. But you cannot ask generally if a whole nucleotide sequence exists anywhere in the database. With just a small amount of a person's DNA, Shringarpure and Bustamente were able to determine whether they were a member of specific disease databases.

The fact that someone's DNA is stored in a certain database can then shed light on that person's health status. There are many disease-specific databases with some degree of public accessibility for "beacon" style queries. A company like 23andme could simply test their members' genetic profiles against all of these databases in order to learn more about the health status of particular individuals. Comparing their members' DNA to a genomic database for HIV, for instance, would allow them to learn which of their members might have HIV. Similar analyses could be performed for diabetes, cancer, autism, and so on.[9]

The studies suggest that, with the right resources, someone could find out a lot about you just based on DNA.[10] However, other types of information can be manipulated in similar ways, with

---

[8] S. S. Shringarpure and C. D. Bustamente, "Privacy Risks from Genomic Data-Sharing Beacons," *American Journal of Human Genetics* 97, no. 5 (2015): 631-646.

[9] See the Beacon Network and the Global Alliance for Genomics & Health for lists of such databases.

[10] For reviews of other genomic de-identification methods, see Y. Erlich and A. Narayanan, "Routes for Breaching and Protecting Genetic Privacy," *Nature Reviews Genetics* 15 (2014): 409-421; S. Wang et al.,

similar results, when combined with the right resources and expertise. While genomic information poses risks to re-identification, these risks appear to be similar both in terms of likelihood and magnitude of harm to those from other types of information.

Consumer Data

Bank statements and credit reports reveal a lot about us, indicating what we buy, where we make our purchases, how much we spend, and what we like. Because they are so revealing, there are federal regulations to make it hard for unauthorized entities to directly access this information. It is not so hard, however, for authorized entities to sell this information to third parties. Banks, credit card companies, retailers, and consumer credit reporting agencies (e.g., Equifax) make a lot of money selling the data they collect from consumers. The organizations that buy and collect this information, known as data brokers, then resell the information.[11] There are currently more than 2,500 distinct companies in the United States selling financial and consumer data in this way.[12] Obtaining significantly revealing data can cost less than $20, depending on the type and amount of data purchased.[13]

Data brokers also routinely combine consumer and financial information with numerous other types of reports, including measures of our health, political activities, and much else.[14] The World Privacy Forum reports that over 100 different variables go into such reports.[15] The New York Times recently found that 20 popular smartphone apps sent precise location information of users to more than 70 businesses, tracking an individual's movement thousands of times a day.[16] If those businesses have the data, it is likely that data brokers do as well.

The information contained in the reports are, like genomic data, often shared in aggregate form using indirect identifiers (e.g., demographic information). But as with genomic information, it is easy to compare information across multiple databases to uncover personally identifying information. For example, one study of de-identified credit information found that data and location information from four purchases were enough to identify individuals within the

---

"Genome Privacy: Challenges, Technical Approaches to Mitigate Risk, and Ethical Considerations in the United States," *Annals of the New York Academy of Sciences* 1387, no. 1 (2017): 73-83.

[11] Federal Trade Commission, *Data Brokers: A Call for Transparency and Accountability* (May 2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf; J. Barrett Glasgow, "Data Brokers: Should They Be Reviled or Revered?" in *The Cambridge Handbook of Consumer Privacy*, ed. E. Selinger, J. Polonetsky, and O. Tene (Cambridge University Press, 2018), 25-46.

[12] P. Boutin, "The Secretive World of Selling Data About You," *Newsweek*, May 30, 2016,

[13] J. Choi et al., "Cybercasing 2.0: You Get What You Pay For," (2018), https://arxiv.org/abs/1811.06584

[14] A. J. Schmitz, "Secret Consumer Scores and Segmentations: Separating Consumer "Haves" and "Have-Nots."" *Michigan State Law Review* 1411 (2015): 1411-1473.

[15] P. Dixon and R. Gellman, "The Scoring of America: How Secret Consumer Scores Threaten Your Privacy and Your Future," *World Privacy Forum* (2014), http://www.worldprivacyforum.org/wp-content/uploads/2014/04/WPF_Scoring_of_America_April2014_fs.pdf

[16] J. Valentino-DeVries et al., "Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret," *New York Times*, Dec. 10, 2018, https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html

database.[17] That is, our spending habits are sufficiently unique that only four purchases can distinguish us from everyone else. The four purchases that identify me in one database can be used to identify me in other databases. Enough of those other databases will include my mailing address, phone number, email address, and other personal information that very little effort is needed to attach such information to my spending behavior.

Health Data

Someone might object that genomic data are more akin to health data than consumer data. Health data is sensitive, which is why we require special consent before it is collected or shared. Consequently, one might argue, the vulnerability of consumer data is irrelevant to the question of how to deal with genomic data.

While health information is protected by HIPAA, HITECH, and other privacy policies, it is still widely collected and shared outside of health-care contexts, where it is unprotected by these regulations.[18] Data brokers like those mentioned above actively seek health information just as they do consumer data. Health care providers as well as insurance companies sell anonymized data from medical records.[19] Perhaps more problematically, people willingly share their own health data outside of health-care contexts, where there are fewer requirements to anonymize information.

For example, Huesch found that 13 popular health websites (e.g., WebMD) used tracking software, and 7 of those sites allowed people's search terms to be shared with third parties.[20] Facebook reportedly uses similar techniques to collect health information from "private" patient support groups (e.g., for breast cancer).[21] A recent study also found that 19 popular mobile health apps shared data with third parties like Amazon, Facebook, Google, and even the Department of Health and Human Services.[22] This sort of information sharing is not protected by

---

[17] Y-A. De Montjoye et al., "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata," *Science* 347, no. 6221 (2015): 536-539.

[18] I. G. Cohen and M. M. Mello, "Big Data, Big Tech, and Protecting Patient Privacy," *JAMA* 322, no. 12 (2019): 1141-1142; T. Glenn and S. Monteith, "Privacy in the Digital World: Medical and Health Data Outside of HIPAA Protections," *Current Psychiatry Reports* 16, no. 11 (2014), 494; K. C. O'Doherty et al., "If You Build It, They Will Come: Unintended Future Uses of Organised Health Data Collections," *BMC Medical Ethics* 17, no. 1 (2016): 54.

[19] M. Allen, "Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates," *Propublica*, July 2018, https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates; A. Tanner, "How Data Brokers Make Money Off Your Medical Records," *Scientific American*, Feb. 2016, https://www.scientificamerican.com/article/how-data-brokers-make-money-off-your-medical-records/

[20] M. D. Huesch, "Privacy Threats When Seeking Online Health Information," *JAMA Internal Medicine* 173, no. 19 (2013): 1838–1839.

[21] K. Ostherr, "Facebook Knows a Ton About Your Health. Now They Want to Make Money Off it," *Washington Post*, April 2018, https://www.washingtonpost.com/news/posteverything/wp/2018/04/18/facebook-knows-a-ton-about-your-health-now-they-want-to-make-money-off-it

[22] Q. Grundy et al., "Data Sharing Practices of Medicines Related Apps and the Mobile Ecosystem: Traffic, Content, and Network Analysis," *British Medical Journal* 364 (2019): I920.

HIPAA, nor is information gathered from other web browsing, nor is the information shared with cosmetic companies, fitness centers, advertisers, and so on. Data brokers can easily combine and compare this information to obtain personally identifying information.[23]

These studies again illustrate that health and consumer information present a similar risk as genomic data. Health and consumer databases are easily accessible, given the wide availability of keys linking data to individuals and direct access through data brokers. It also appears that the potential harms from accessing health and consumer information may be just as damaging as the harms from accessing genetic information. If data brokers have already collected and re-identified all the consumer and health data just mentioned, it's not clear what additional risk to re-identification genomic information would pose. Potential harm from knowing our purchasing or browsing behavior is much higher for most of us than anything that could be inferred from re-identified genetic data, especially given the current state of the science. Those claiming that genomic information is unique, or should be handled in a special way, need to explain more carefully why the probability of re-identification or the magnitude of potential harm is higher for genomic data than other sorts of data.

Potential Objections

Before discussing the implications of our argument for informed consent, we will briefly address three objections one might make in response to our position here.

Financial and health information is essential; genomic sharing is not

Sharing financial and some amount of health information is necessary for daily life. Even if there are third parties tracking my purchasing behavior and web browsing, I am not going to stop using a credit card or the internet. In contrast, people do not have to share their genomic information. Contributing to a biobank, for example, is optional for most people and likely will not improve their personal health care in the short-term. We might think that people are willing to accept the identification risks from sharing financial and health information, even if they don't explicitly consent to them, because they are so essential to everyday life. They do not have to accept such risks, however, with sharing genomic information.

Although sharing genomic data is currently non-essential, we would point out that it could very well become a standard part of health care in the near future. That is in fact one of the promises of precision health and personalized medicine. Advances in pharmacogenomics, for example, are enabling some degree of individualized treatment based on patients' genomes.[24] If sharing genomic data became a common part of life, our recommendations would remain the same. Genomic data would still be just one of many types of data that already make us vulnerable.

---

[23] There have also been notable challenges for achieving thorough de-identification of health data, as well as health-related data from wearable fitness trackers. N. Liangyuan et al., "Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning," *JAMA* 1, no. 8 (2018): e186040; L. Sweeney, "Only You, Your Doctor, and Many Others May Know," *Technology Science* (2015): 2015092903.

[24] J. A. Johnson and K. W. Weitzel, "Advancing Pharmacogenomics as a Component of Precision Medicine: How, Where and Who?" *Clinical Pharmacology & Therapeutics* 99, no. 2 (2016): 154-156.

Suppose, for example, what it would mean if every patient received whole genome sequencing as part of their care. Their genetic information would, as a result, become part of their health record, which would affect their recommended treatments, medications, and nutrition, and would furthermore be reflected in what people buy and search for online. Eventually all this information would make its way to data brokers. But nearly every American's personal information is already vulnerable through data brokers, even without genetic data. Routine genomic data sharing might add some sensitive information to these databases but would not change the nature of the underlying risk.

Law enforcement can easily find me with DNA

There are numerous recent cases of law enforcement using genetic databases to solve crimes, even long after the criminal has died. For this reason, DNA is arguably a persistent threat to identification in a way that consumer or other health information are not. Biobanks and genetic databases do not generally share their data with law enforcement, but that might change. And even if one thinks it is ok for law enforcement to use genetic data to solve crimes, who knows whether such databases will be used accurately and for legitimate law enforcement purposes. I don't want to donate my biospecimens if it exposes me to wrongful incrimination for the rest of my life.

Tracking down criminals through DNA is merely a new application of a technique that law enforcement has used on financial information for years. Financial information may not directly place you at the scene of a murder, but it does leave a trail of information for law enforcement to deduce that you were likely the murderer. Financial and consumer data are also persistent threats in that the information is stored indefinitely. Even after a person has died the information still says a lot about them. It can also be used to make inferences about family members (including where they live). And unlike data shared as part of genetic research, there is little oversight for how the information is used.

New privacy laws will put an end to data brokers

The EU's General Data Protection Regulation (GDPR) as well as the California Consumer Privacy Act (CCPA) give consumers significant control of their personal data (as do many other state laws currently under consideration). Though not yet fully implemented, perhaps these new laws will put data brokers out of business, and thereby decrease the risk of identification from consumer databases. The GDPR, for instance, requires companies to obtain personal consent before processing an individual's data. The CCPA requires companies to tell people when their data is collected, sold, and shared, and allow them to opt out. If other states adopt similar laws, there could be massive withdrawal from consumer databases.

While these laws do protect identifiable consumer information, they don't currently pose a significant obstacle to processing de-identified information. As Barocas and Nissenbaum argue, data processers do not need your consent to share identifiable information if they obtain enough of the right de-identified information.[25] Data brokers are adept at navigating de-identified datasets precisely for this reason. Your right to withdraw is relatively meaningless if data brokers

[25] S. Barocas and H. Nissenbaum, "Big Data's End Run Around Anonymity and Consent," in *Privacy, Big Data, and the Public Good*, ed. J. Lane, V. Stodden, S. Bender, and H. Nissenbaum (Cambridge University Press, 2014), at 45.

can plausibly maintain that they do not know which information is yours (even while making inferences about you and sharing information with others who would be able to identify you).

Data processors have also apparently adapted to new GDPR requirements.[26] Disclosure requirements have changed, but as has been well established empirically, people tend to automatically accept such disclosures without reading them.[27] Price et al. also argue that there's so much ambiguity in the guidance provided by the GDPR that member countries could choose to take no action on certain GDPR requirements, especially in protecting health data.[28] In short, it does not seem that the risk of re-identification from non-genetic data will decrease any time soon.

Communicating Genomic Identification Risks

The issue that sparked recent debates over re-identification was the proposal, in suggested changes to the Common Rule, that all genetic data and biospecimens containing genetic information could not ever be considered de-identified. This proposal was rejected, in part due to the negative effect it could have on medical research. We have provided another reason to reject the proposed change in the Common Rule, by arguing that 1) de-identified genetic data is not more likely to be re-identified than other sorts of de-identified data and 2) the resulting harms are not clearly worse for re-identified genetic data. This similarity between the risks of misuse of genetic and non-genetic data has implications extending to clinical and even consumer settings (such as direct-to-consumer testing). Here we draw out the implications of our argument for consent for research and care.

According to the Reasonable Person Standard, a widely used ethical and legal guideline, consent processes should communicate risks that a "reasonable person" would want to know, which is often understood to be related to what the "typical" person would want to know.[29] Applying this standard, we argue, has the following implications: consent forms should disclose, generally, the reasons for collecting and sharing data for healthcare or research, that steps are taken to protect privacy in both cases, and that privacy violations and potential harms can result from legal and illegal use of the data. Consent forms do not need to explain the risks of re-identification, and if they do, they should not draw a sharp line between the risks of re-identification from genetic vs. non-genetic data.

---

[26] G. Aridor et al., "The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR," January 29, 2020, https://ssrn.com/abstract=3522845

[27] J. Obar and A. Oeldorf-Hirsch, "The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services," *Information, Communication & Society*, 23, no. 1 (2018): 128-147; I. Van Ooijen and H. U. Vrabec, "Does the GDPR Enhance Consumers' Control over Personal Data? An Analysis from a Behavioural Perspective," *Journal of Consumer Policy*, 42 (2019): 91-107.

[28] W. N. Price et al., "Shadow Health Records Meet New Data Privacy Laws," *Science*, 363, no. 6426 (2019): 448-450.

[29] J. Greenblum and R. Hubbard, "The Common Rule's 'Reasonable Person' Standard for Informed Consent," *Bioethics* 33, no. 2 (2018): 274-277; L. M. Odwazny and B. E. Berkman, "The 'Reasonable Person' Standard for Research Informed Consent," *American Journal of Bioethics* 17, no. 7 (2017): 49-51.

Consent forms for patient care often describe how the health system will collect and share health data, but not that the information is inherently identifying. While we do not know of any systematic analyses of "consent to treat" forms, many health systems make their forms and associated privacy policies publicly available. We reviewed dozens of these forms from large health systems throughout the U.S. (examples available on request). They typically outline various third parties that will receive patient information, both identified and de-identified, and patients' rights in viewing or controlling access to this information. None mentioned the risks to re-identification from such sharing, either for genetic or non-genetic data.

We support this approach to disclosure, whether for healthcare or research, and for genetic or non-genetic data. Consent forms must disclose that there is a risk of data misuse, since the resulting harms are arguably ones that a person who is agreeing to healthcare or research would reasonably want to know. Data breaches for healthcare data are well-known, as is the risk of that information being used for identity theft.[30] But even when consent forms disclose this risk, they should do so consistent with our suggestions: without explaining that there is an inherent risk of re-identification, and without drawing a distinction in the risks from loss of privacy resulting from genetic and non-genetic data. While such information may be made available to those who wish to learn more, there is no need to make it part of standard disclosure.

Empirical research about attitudes toward data sharing and informed consent support our interpretation of the reasonable person standard in this context. Some studies suggest that people do want special notice about the risks of sharing genetic information. For example, in the context of research participation, Lee et al. found that people were concerned their genetic information could be used against them in various ways (e.g., they mentioned cloning) and were in fact identifiable in ways their medical data were not.[31]

However, many other studies find that people view genetic and non-genetic risks as roughly the same. In some cases, individuals appear to view non-genetic information as more risky than genetic data. For example, Kim et al. recently asked patients to indicate which of 59 types of health and genetic information, including biospecimens, they were willing to share for research at different types of institutions.[32] The results showed that people were equally willing to share both genetic and non-genetic information. Even when consent forms indicated that the information would be shared for commercial purposes, patients agreed to share blood, tissue, urine, and the results of genetic tests at roughly the same rate as information about past medical conditions and family health history (39-40% of patients). Patients were also much more likely to refuse to share contact information, like their home address or telephone number, with any type of researcher (15% of patients) than they were blood or other tissue (10% of patients).

---

[30] J. Jiang and G. Bai, "Types of Information Compromised in Breaches of Protected Health Information," *Annals of Internal Medicine* 172, no. 2 (2020):159-160.
[31] S. S.-J. Lee et al., "I Don't Want to be Henrietta Lacks: Diverse Patient Perspectives on Donating Biospecimens for Precision Medicine Research," *Genetics in Medicine* 21, no. 1 (2019): 107-113.
[32] J. Kim et al., "Patient Perspectives About Decisions to Share Medical Data and Biospecimens for Research," *JAMA Network Open* 2, no. 8 (2019):e199550.

Numerous studies have also found that people overwhelmingly support sharing both health and genetic information.[33] Support is especially strong among those who are familiar with sharing health data. For example, Peppercorn et al. recently found that 65% of cancer patients approved sharing their leftover specimens after surgery, even without their consent.[34]

Similarly, Mello et al. found that 93% of people involved with clinical trials were very or moderately willing to share both biospecimens and other health data as part of clinical trials.[35] Respondents furthermore thought the possibility of being re-identified based on their information was very low, with few participants saying that identification (6.6%) or discrimination based on identification (5%) were their biggest concerns. Importantly, in the qualitative component of their study, Mello et al. found that participants' expressed high confidence in the de-identification process.

This confidence in the de-identification process is reasonable, even for genetic data. A number of the database comparison methods for re-identifying genetic data, described above, have been addressed and protections against them increased.[36] Despite the theoretical dangers and widely expressed concerns, there are still no published reports (that we are aware of) of harm from re-identification of de-identified genomic data.

Genomic identification is also currently much more difficult and resource-intensive than identification through other means. Consider again Erlich et al.'s finding that 60% of Americans could currently be identified through their relatives in genomic databases. Although 60% may seem like a lot, the number for non-genetic data is even higher: Paying a data broker provides direct access to a plethora of personally identifying information about nearly every American based on their consumer and online behavior.

Thus, we conclude the reasonable person standard would not require special disclosure about the risks of identification from genomic data. The risks of re-identification of de-identified genetic data need not be identified as a unique risk or emphasized in informed consent for research or clinical care. Information about re-identification may be made available, but it need not be included in standard consent.

Education outside of the informed consent process

---

[33] E. W. Clayton et al., "A Systematic Literature Review of Individuals' Perspectives on Privacy and Genetic Information in the U.S." *PLoS One* 13, no. 10 (2018): e0204417.
[34] J. Peppercorn et al., "Patient Preferences for Use of Archived Biospecimens from Oncology Trials When Adequacy of Informed Consent Is Unclear," *The Oncologist* 25, no. 1 (2020): 1-9.
[35] M. M. Mello et al. "Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing," *New England Journal of Medicine* 378 (2018): 2202-2211.
[36] H. Cho, D. J. Wu, and B Burger, "Secure Genome-Wide Association Analysis Using Multiparty Computation," *Nature Biotechnology* 36 (2018): 547-551; C. M. O'Keefe and D. B. Rubin, "Individual Privacy Versus Public Good: Protecting Confidentiality in Health Research," *Statistics in Medicine* 34, no. 23 (2015): 3081–103; J. L. Raisaro, F. Trafer, Z. Li et al., "Addressing Beacon re-Identification Attacks: Quantification and Mitigation of Privacy Risks," *Journal of the American Medical Informatics Association* 24, no. 4 (2017): 799-805.

Ultimately, however, the broader lesson is that we must rethink the current consent-based approach to educating patients about collection, sharing, use, and potential misuse of information. Consent is not an appropriate outlet to educate people about the entire network of data sharing. It is well known that people routinely fail to understand privacy issues in consent forms. Garret et al. found that only 19% of patients being enrolled in a biobank understood disclosed information that pharmaceutical and biotechnological companies could receive their information, and only about half understood that government agencies and other researchers might receive their information.[37] Similarly, Kasperbauer and Schwartz found that only 2% of biobank participants enrolled in a healthcare setting correctly recalled that information would be shared outside the biobank, despite clear statements that it would be.[38] Studies also find that patients often find discussions of identification, de-identification, and re-identification in consent forms to be confusing.[39] In short, trying to educate through consent is unlikely to be effective.

Instead, health systems should look for other opportunities to educate patients as part of their care. Discussions of data sharing within learning health systems have outlined how this might work without jeopardizing patient care.[40] For example, patient health portals could be used to explain the health system's data sharing practices in more detail. Reception and registration procedures could also be modified to provide a more direct, personal conversation about data sharing, separate from consent for treatment. To be effective, the explanation needs to be linked to other information patients care about (like their own health) rather than included within a long list of disclosures that must be acknowledged to receive care. Additional empirical work in this area is needed to determine the best ways of communicating such information.

It is also important that health systems take on the responsibility of educating patients. Other groups have little incentive to explain the benefits and risks of data sharing and use. Health technology and direct-to-consumer testing companies often underdescribe privacy risks, sometimes have no privacy policy at all, and routinely change their terms of service without informing customers.[41] While we should ask more from these companies, it is unlikely that they will proactively communicate the risks of pervasive data sharing to their customers. Instead, the responsibility must lie with those who have the opportunity to educate patients, are trusted, and already have obligations to protect sensitive information.

Conclusion

---

[37] S. B. Garrett et al, "Standard Versus Simplified Consent Materials for Biobank Participation: Differences in Patient Knowledge and Trial Accrual," *Journal of Empirical Research on Human Research Ethics* 12, no. 5 (2017): 326–334.

[38] T. J. Kasperbauer and P. H. Schwartz, "Measuring Understanding and Respecting Trust in Biobank Consent," *American Journal of Bioethics* 19, no. 5 (2019): 29-31.

[39] E. Rothwell et al., "An Assessment of a Shortened Consent Form for the Storage and Research Use of Residual Newborn Screening Blood Spots," *Journal of Empirical Research on Human Research Ethics* 12, no. 5 (2017): 335-342.

[40] J. R. Botkin, "Transparency and Choice in Learning Healthcare Systems," *Learning Health Systems* 2, no. 1 (2017): e10049.

[41] J. W. Hazel and C. Slobogin, "Who Knows What, and When?: A Survey of the Privacy Policies Proffered by U.S. Direct-to-Consumer Testing Companies," *Cornell Journal of Law and Public Policy* 28 (2018): 35-66; J. L. Roberts and J. Hawkins, "When Health Tech Companies Change Their Terms of Service," *Science* 367, no. 6479 (2020): 745-746.

Genomic data contains identifying information that can pose risks to patients and research participants. However, as we have argued, the re-identification and privacy risks are not obviously greater than similar risks from other types of de-identified data. Appreciating these similarities undermines the assumption that patients and research participants need a special type of disclosure when sharing genetic data or samples containing genetic material.

More generally, action is needed to address the risks of pervasive data sharing. The network of data sharing that exists in society, driven by data brokers, is a threat to privacy. The problem is not access to genetic data, nor research data, but the aggregation of many types of data from multiple sources. As we argued, moving education about these practices outside of consent forms may provide a new avenue for increasing general knowledge about the entire network of data sharing.

While greater societal awareness of data sharing practices would help, it is also necessary to enact broader privacy legislation than currently exists. As others have recently argued, increased transparency and accountability for data brokers are essential first steps in protecting personal data.[42] Existing laws focus on protecting identifiable individuals in specific contexts (like HIPAA). But data brokers are adept at overcoming anonymization techniques and exploiting consent policies that limit data sharing in one context but not others. Comprehensive federal privacy legislation could regulate data processing across contexts.[43]

[42] T. J. Kasperbauer, "Protecting Health Privacy Even When Privacy is Lost," *Journal of Medical Ethics*, Epub ahead of print, doi:10.1136/medethics-2019-105880; W. N. Price et al., "Shadow Health Records Meet New Data Privacy Laws," *Science* 363, no. 6426 (2019): 448-450.
[43] A. Chander et al., "Catalyzing Privacy Law," *Minnesota Law Review* (forthcoming), https://scholarship.law.georgetown.edu/facpub/2190.