ETHICAL CONSIDERATIONS FOR HOME DNA TESTING

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Zachary Thomas
Spring, 2021

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature: Zachary Thomas

Approved: Sharon Tsai-hsuan Ku, Department of Engineering and Society

Data science approaches to any topic rely on data availability, and a lot of it. This so called “Big Data” can be used to make decisions based on statistics and trends, it can also provide insight into the cause of these trends. Genomic data is no different, 5 providing statistical likelihoods of low frequency events, like cancers, drug treatment side effects, and drug-drug interactions on a per patient level (Francis, 2014). The availability of genomic data also proves crucial for the development, maintenance, and progression of bioinformatics tools like BART (Wang et al., 2018). For this reason, a number of international policies demand immediate release of all sequencing data without explicit patient consent (Johnson, Slade, Giubilini, Graham, 2020, p. 150). However, this public release of genome data introduces a number of concerns including patient privacy and questions of data ownership; whether or not genomic data sets belong to the patient or the researcher. For example, in 1989, a member of the Havasupai Tribe of Arizona had donated DNA samples for genetic research on type II diabetes. Yet, in 2003, she found that her DNA samples were being used for other, nondiabetic research studies. After a six-year trial, a settlement was reached that included monetary compensation but no legal ramifications for the researchers that misused the DNA samples (Garrison, 2013, p. 202).

Many of the issues regarding genomic data sharing have been addressed by the Office of Human Research Protections in the form of alterations to informed consent policies and the introduction of Broad Consent (Fisher & Layman, 2018). However, these protections and regulations only apply to research institutions. With the recent surge in at-home DNA testing kits, like the ones provided by 23andMe, new ethical challenges are coming to light.

This thesis is about the new ethical challenges at-home DNA testing kits bring. These include a shift from patients to consumers, data shifts that render current patient centered legislation useless, new data privacy challenges, and the introduction of a number of third

parties: these include genetic researchers, bioinformaticians, biomedical data scientists, insurance companies, medical companies and many others.

Literature Review

With genomic data as the technology of interest, there have been a number of studies and papers written focusing on the ethical and legal issues of storing and sharing that kind of information. For example, the issue concerning the Havasupai Tribe has been studied as an example of poor legislation regarding data sharing. In his paper, Garrison discusses the result of the lawsuit and the institutional review board's practices through carefully examination of interviews with members of the board (Garrison, 2013). The author concludes the piece by demanding the institutional review board alter their practices as to foster inclusivity and potential benefit from genomic data for all people. This article puts the ethical issues of data sharing and publicly available data in context for further review of data collecting and distributing practices.

Other research has looked at the positive impacts of looser rules for genomic data sharing. Stephanie Johnson is a researcher at the Wellcome Center for Ethics and Humanities at the University of Oxford. In her article, "Rethinking the Ethical Principles of Genomic Medicine Services", she and her colleagues from the University of Oxford argue that any and all genomic datasets should be available to all researchers (Johnson, Slade, Giubilini, Graham, 2020). They argue that these types of data sets carry substantial weight and those involved with administering a genomic data set have an ethical obligation to share that data.

These are just some examples of current research regarding genomic data. However, they only focus on genomic data generated by studies and fail to address genomic data generated by at-home DNA testing.

At-Home DNA Technology

23andMe

As one of the first at-home DNA testing companies, 23andMe has sequenced the DNA of about ten million individuals as of 2019 (Carson, 2019). The genetic testing company offers sequencing kits ranging from ninety-nine to one-hundred and ninety-nine dollars for individuals to gain insights into their own ancestry or health reports. While the health reports now meet the Food and Drug Administration (FDA) requirements, this was not always the case. At the start of the company in 2007, they were marketing their product as a health risk assessment system. In 2013, they were ordered by the FDA to stop any and all advertising related to health risk assessment as it wasn't an FDA approved medical testing device (Carson, 2019). In 2015, the kits included a revised health component to meet the FDA requirements.

While 23andMe works to sequence as much DNA as fast as possible and return it to their consumers, they continuously work to support legislative efforts intended to prevent genetic discrimination. The company was active in the development of both the Genetic Information Non-discrimination Act and CalGINA (2011) which includes a longer list of organizations that are not allowed to access or request genetic data in the state of California.

Another important facet of 23andMe is their research program. When a customer is checking out on the 23andMe website, they have the option of whether or not to have their results included in research studies. Their website provides the customer with a plethora of reasons to participate in their research program from "contributing to over 230 studies" to quotes from previous customers who feel inspired from their participation (23andme.com/research/). Over its lifetime, the company has partnered with the National Institutes of Health, Genentech, the Broad Institute of MIT and Harvard, Stanford University, and, most recently,

GlaxoSmithKline. When participating in these studies, 23andMe works to protect their customers privacy by de-identifying their data. This means that any kind of personal data or registration information that was supplied at checkout will be stripped from the data before it is put into the research pool.

Public Use of Data

While it would seem the research community and government are well-informed of the weight of genomic data sets, the public is not so well educated. There have been substantial reports of parents freely sharing their children's genomic information obtained from at-home DNA testing kits like the ones provided by 23andMe (Bala, 2020). This lack of awareness of the ramifications for sharing genetic information highlights the issue with the public's perception of genetic data. In terms of at-home DNA testing, there are a number of issues that skew the public's perception of the genome. First, the testing kits are returned without mediation by a doctor or genetic counselor. This limits the ability of the user to extrapolate key information that might reveal privacy concerns from the results. Second, the kits allow the user to freely post their results in the form of whole-genome sequences. This allows for potential identification even if they posted anonymously. Finally, the data contains information about the user who took the test, but also their closely related family members. This allows for interpolation of genetic information between family members should the data be re-identified.

Social Construction of Technology Overview

When examining such intricate problems like genomics and privacy in a data driven world, it is helpful to use an analytical framework. For this particular problem, a social construction of technology (SCOT) model as developed by Trevor Pinch and Wiebe Bijker will be used to analyze the parties at play (Bijker & Pinch, 1984). As a model, SCOT is made up of

four major parts: relevant social groups, interpretive flexibility, closure, and stabilization. Figure 1 illustrates social groups relevant to genomic data and privacy.

With genomic data and privacy as the technology of interest, there are several relevant groups that perceive value from it. In this model, each group applies different values to the technology and have either competitive or collaborative problems with the technology. For example, the researchers using the data to build data science driven tools might need to know the age and sex of the patient whose data they are using in order to maintain a robust pipeline. However, the patient might not want that level of detail disclosed. Likewise, commercial DNA testing companies might want to protect their customers data so they can be trusted and make more money where as insurance companies would like to access that data to adjust prices to similarly make more money.

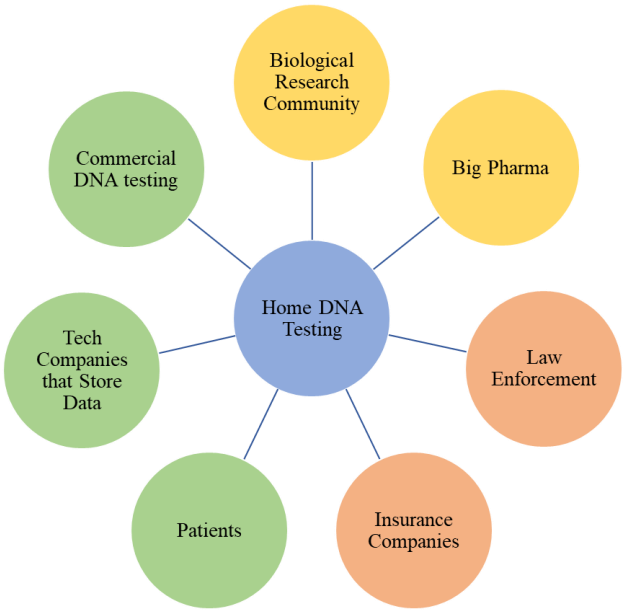

Figure 1: Social Construction of Genomic Data Privacy. This figure illustrates the relevant social groups to genomic data and privacy. The social groups are colored based on how useful increasing strict privacy regulations would be. Orange circles would see the most detriment to increasing privacy, yellow sees both positive and negative outcomes, and green sees only positive outcomes to increasing privacy regulations (Adapted by Zachary Thomas (2021) from Bijker, Bönig, & van Oost, 1984).

SCOT Application

Commercial DNA testing and Patients

Although they have not been completely transparent with how they handle data, commercial DNA testing companies generally seem to favor privacy above all else. This evidence can be seen in their continued support of legislation like GINA and CalGINA as well as clear privacy policies and options to be included or excluded from research programs. Here, commercial testing companies and patients apply values of privacy and anonymity to genetic data. Commercial testing companies take this a step further and apply values of research and scientific discovery to these data sets. Similarly, the companies that design and sell databases for storing genetic data would maintain privacy as a top value for to make it more attractive to sell and implement for these commercial DNA testing companies. For this reason, they are colored green in Figure 1 to demonstrate that these groups see no detrimental effects of increasing genetic privacy.

Research and Big Pharma

Research institutions and pharmaceutical companies perceive both positive and negative values from increasing genetic privacy. On one hand, increasing privacy and anonymity might increase the number of data sets available as people will feel more comfortable posting their genetic testing results on public databases or will be more willing to participate in studies. Additionally, continuing to de-identify data and make it more secure means that for-profit researchers have no obligation to compensate the data-owners as they do not have access to their identities. Conversely, increasing privacy protections and anonymity limits the amount of data these groups can obtain from the sequencing results. This limitation can impact how useful the data is as no demographic information directly related to the genetic data is easily accessible.

These demographics may be critical to ensuring diversity in a general population type study or demonstrating a particular drug's effectiveness.

Insurance and Law Enforcement

Finally, insurance companies and law enforcement derive value from genetic information as well. As previously mentioned, various legislation exists that limits the ability for health insurance companies to require or access genetic data of their customers. However, not all types of insurance are covered in this legislation. For this reason, increasing privacy policies regarding genetic data has detrimental effects on insurance companies that are not health insurance. Additionally, there have been instances of law enforcement leveraging genetic data testing and its public availability to solve cases such as the Golden State Killer (Chamary, 2020). Should genetic data privacy become stricter, law enforcement would have a harder time accessing this data and using it to directly close cases.

New Ethical Challenges

Patient to Customer

Prior to the surge in at-home DNA testing, any genetic data that was generated was part of a research study or collected by health care professionals. These genetic data sets can be classified as patient data and is protected by specific legislature that will be discussed later in this section. However, when a customer orders and implements a DNA testing kit, there is no highly protected party to oversee its use. This specifically brings up concerns regarding research ethics. For 23andMe, research almost always involves a partnership with a research institution or big pharmaceutical company. Even though they list pharmaceutical companies on their website, customers were still disappointed to learn that their data was being used in for-profit research through GlaxoSmithKline when the news of their partnership was announced (Ducharme, 2018).

Others, like the president of the Center for Medicine in the Public Interest, took it a step further and suggested that customers whose data are being used in for-profit research like that being conducted by GSK should be compensated (Ducharme, 2018). Additionally, with all 23andMe partnerships, the data that is being shared must be moved to the group it is being shared with. Genetic information is already a high-risk data type and this risk increases when moving data from one location to another. The highest risk being interception.

Other ethical concerns arise when the distribution of 23andMe data is analyzed. First, the data set is limited to those that can afford at least the ninety-nine-dollar test kit. This already biases any kind of "general population" level insights from 23andMe studies and limits widespread relevance of results. This is very different from NIH type research. The NIH follows informed consent policies, Broad Consent, and other policies regarding patients and their data. Additionally, the NIH actively reaches out to specific people in order to secure genetic diversity.

This issue of demographics is not limited to 23andMe and research that uses their data, this plagues the biomedical engineering data science and bioinformatics community as well. As of 2016, nearly eighty percent of genetic data in genome wide association study databases belongs to individuals of European descent (Popejoy & Fullerton, 2016). As many artificial intelligence and machine learning based bioinformatics studies rely on these data, there are many questions as to how this will bias these studies and if they should or should not be published (Bentley et al., 2017).

Current Policy and its Limitations

The concept of privacy as it relates to genomic data has undergone a number of revisions in recent years, especially in the United States with the introduction of the Genetic Information Nondiscrimination Act (GINA) of 2008. Under this act, it is declared unlawful for health

insurance or employers to discriminate based on genetic information. More specifically, health insurers are not allowed to "make coverage, underwriting, or premium-setting decisions" based on genetic data of individuals. Similarly, GINA prevents employers from making hiring, firing, promotional, pay, and job assignment decisions using genetic information of their employees. Surprisingly, GINA also protects against using genetic testing results of family members to make insurance or employment decisions.

In the United States, defining genetic privacy is much more flexible than other parts of the world because genetic information protections are being developed separately instead of being covered in more general laws (Molteni, 2019). In the European Union, for example, DNA has been designated as personal data and is therefore covered by all privacy laws relating to personal data (Molnár-Gábor & Korbel, 2020). The Genetic Information Nondiscrimination Act of 2008 is not the only form of legislation that attempts to protect an individual's genetic privacy. Other forms of privacy legislation include the US Common Rule (1981), the 21st Century Cures Act (2016), the Health Insurance Portability and Accountability Act (HIPAA) (1996), and the Affordable Care Act (2010). Table 1 summarizes which piece of legislation applies depending on where the genetic data was generated or posted. Most importantly, the US Common Rule, 21st CCA, and HIPAA all maintain that privacy is protected when the data is stripped of any kind of personally identifiable information.

Table 1: Who protects genomic data?

Where the data came from/is	**Who can access it**	**Who protects it?**
Research Study/Clinical Trial	Researchers Those disclosed prior to the study (other researchers)	Us Common Rule 21st Century Cures Act
Electronic Health Record	Doctor Law Enforcement Insurance Provider	HIPAA GINA ACA
Public Database	Virtually anyone	GINA

The Genomic Information Nondiscrimination Act protects against health insurers requiring or accessing any kind of genetic data about their customers to manipulate pricing, it falls short on other fronts. Most notably, GINA does not protect against discrimination in other forms of insurance. For example, life, disability, and long-term care insurance do not have any equality obligations like health insurance does under GINA. CalGINA picks up some of this slack by identifying other areas like medical services and housing as areas where genetic information could be used to manipulate prices, but CalGINA doesn't cover them all and only applies in the state of California. This issue stems from the fact that the United States refuses to group DNA with other personal data.

Another seemingly roundabout way to maintain some form of privacy is to "de-identify" the data prior to use in research, trial, or public release. A perfect example of this can be seen through the 23andMe research program. According to their privacy policy, 23andMe strips any data used for research of personal details such as name, contact information, address, etc. (23andMe, 2021). This is very similar to HIPAA policies to protect health information. To HIPAA (1996), de-identification means mitigating privacy risks and "thereby supporting the secondary use of data for comparative effectiveness studies, policy assessment, life sciences research, and other endeavors". The act of deidentifying data is similar to that of 23andMe where all types of identifies, of which there are eighteen, are removed from the data.

For health care records, the HIPAA policies work wonderfully. Generally, the risk of re-identifying HIPAA protected data is on the order of 0.01% (Benitez & Malin, 2010). Prior to the explosion of at-home DNA testing kits around 2010, it was nearly impossible to re-identify genetic data, much less find pure sequencing data. However, in 2013, Yaniv Erlich of the Whitehead Institute for Biomedical Research demonstrated it was possible to re-identify people

listed on anonymous genetic databases. His team did this only relying on publicly accessible internet resources (Gymrek et al., 2013). This feat was accomplished in large part due to the vast number of people of European descent that have had their DNA tested and posted online (Molteni, 2018).

For government regulated groups or research institutions and their patients, this is much less of issue. This is because groups like the National Institutes of Health restrict access to their genetic databases (Pereira et al., 2014). The main concern here is when people obtain their own genetic information. With no professional intermediary, people are unaware of the risks that posting their genetic information warrant to not only themselves, but their family members and future generations as well.

Privacy and Third Parties

As we saw with the 23andMe and GSK partnership, there have already been concerns raised about third-party interference in research and data collection. However, this is not just limited to what could be considered unethical research. By commercializing testing kits, outside parties that current policy has not had a need to address are introduced to the equation. For example, HIPAA only works with patients and the health care professional. Now, with DNA testing kits being widely available new actors come in to play such as GSK and other pharmaceutical companies, companies that design and sell databases to store data for these companies, malicious actors that now have access to a wealth of genetic data posted on public forums, law enforcement, and many others. Current policy really only addresses patients and consumers, health care professionals, and insurance companies. This is a major gap that needs to be filled if users are expected to trust DNA testing as a form or education and entertainment.

No Simple Solution

From the perspective of the patient or customer, there are many reasons one should be concerned about the state of genetic privacy as it exists in the United States today. However, until there is proper legislature to address it, genomic data privacy is still a gray area. The largest barrier to any kind of change in genomic privacy is the fact that so many, including 23andMe and HIPAA, believe that privacy is maintain if the data is de-identified. This stems from the belief that if the data is de-identified, it is impossible to trace back to an individual. This is simply not true and has been shown to be an attainable feat. Yet, changing this belief that de-identification equals privacy is not so simple. Privacy as it relates to genetics is complex. The genome of an individual is static over time, meaning once it has been sequenced, the data is relevant and accurate for the remainder of the individual's lifetime. Genomic data is also very personal, but simultaneously revealing about family members. Finally, genomic data could contain information that a patient would like to know about themselves but is impossible to relate back to them should privacy regulations increase. All of these reasons, paired with the fact that any genomic data set is identifiable, make genetic privacy in a data driven world that much more complex.

References

21st Century Cures Act. Pub. L. No. 114-255.

23andMe. (n.d.). *Privacy and Data Protection—23andMe*. Retrieved March 25, 2021, from https://www.23andme.com/privacy/

Affordable Care Act. Pub. L. No. 111-148.

Bala, N. (2020, January 2). Why are you publicly sharing your child's DNA information? The New York Times. Retrieved from https://www.nytimes.com/

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association : JAMIA*, *17*(2), 169–177. https://doi.org/10.1136/jamia.2009.000026

Bentley, A. R., Callier, S., & Rotimi, C. N. (2017). Diversity and inclusion in genomic research: Why the uneven progress? *Journal of Community Genetics*, *8*(4), 255–266. https://doi.org/10.1007/s12687-017-0316-6

Bijker, W., Bönig, J., van Oost, E. (1984). The social construction of technological artefacts. Zeitschrift für Wissenschaftsforschung, 2, 39-52.

Bijker, W., & Pinch, T. (1984). The social construction of facts and artifacts: or how the sociology of science and the sociology of technology might benefit each other. Social Studies of Science, 14, 399-441. doi:10.1177/030631284014003004

Carson, B. (2019, June 6). *Live Long And Prosper: How Anne Wojcicki's 23andMe Will Mine Its Giant DNA Database For Health And Wealth*. Forbes. https://www.forbes.com/sites/bizcarson/2019/06/06/23andme-dna-test-anne-wojcicki-prevention-plans-drug-development/

Ducharme, J. (2018, July 26). *A Major Drug Company Now Has Access to 23andMe's Genetic Data. Should You Be Concerned?* Time. https://time.com/5349896/23andme-glaxo-smith-kline/

Fisher, C. B., & Layman, D. M. (2018). Genomics, Big Data, and Broad Consent: A New Ethics Frontier for Prevention Science. *Prevention Science*, *19*(7), 871–879. https://doi.org/10.1007/s11121-018-0944-z

Francis, L. P. (2014). Genomic knowledge sharing: A review of the ethical and legal issues. *Applied & Translational Genomics*, *3*(4), 111–115. https://doi.org/10.1016/j.atg.2014.09.003

Garrison, N. A. (2013). Genomic Justice for Native Americans: Impact of the Havasupai Case on Genetic Research. *Science, Technology & Human Values*, *38*(2), 201–223. https://doi.org/10.1177/0162243912470009

Genetic Information Nondiscrimination Act. Pub. L. No. 110-233.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying Personal Genomes by Surname Inference. *Science*, *339*(6117), 321–324. https://doi.org/10.1126/science.1229566

Health Insurance Portability and Accountability Act. Pub. L. No. 104-191.

Johnson, S. B., Slade, I., Giubilini, A., & Graham, M. (2020). Rethinking the ethical principles of genomic medicine services. *European Journal of Human Genetics*, *28*(2), 147–154. https://doi.org/10.1038/s41431-019-0507-1

Molnár-Gábor, F., & Korbel, J. O. (2020). Genomic data sharing in Europe is stumbling—Could a code of conduct prevent its fall? *EMBO Molecular Medicine*, *12*(3). https://doi.org/10.15252/emmm.201911421

Molteni, M. (2018, October 11). Genome Hackers Show No One's DNA Is Anonymous Anymore. *Wired*. https://www.wired.com/story/genome-hackers-show-no-ones-dna-is-anonymous-anymore/

Molteni, M. (2019, May 1). The US Urgently Needs New Genetic Privacy Laws. *Wired*. https://www.wired.com/story/the-us-urgently-needs-new-genetic-privacy-laws/

Pereira, S., Gibbs, R. A., & McGuire, A. L. (2014). Open Access Data Sharing in Genomic Research. *Genes*, *5*(3), 739–747. https://doi.org/10.3390/genes5030739

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. https://doi.org/10.1038/538161a

Revised Common Rule. (2017, January 17). [Text]. HHS.Gov. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html

Wang, Z., Civelek, M., Miller, C. L., Sheffield, N. C., Guertin, M. J., & Zang, C. (2018). BART: A transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics (Oxford, England)*, *34*(16), 2867–2869. https://doi.org/10.1093/bioinformatics/bty194