Genetic Data Governance in Crisis: Policy Recommendations for AI Applications

Vivek Ramanan
Brown University
Providence, USA

Ria Vinod
Brown University
Providence, USA

Cole Williams
Brown University
Providence, USA

Sohini Ramachandran
Brown University
Providence, USA

Suresh Venkatasubramanian
Brown University
Providence, USA

ACM Reference Format:
Vivek Ramanan, Ria Vinod, Cole Williams, Sohini Ramachandran, and Suresh Venkatasubramanian. 2025. Genetic Data Governance in Crisis: Policy Recommendations for AI Applications. In *Sociotechnical AI Governance: Opportunities and Challenges for HCI, April 2025, Yokohama, Japan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The scale of genetic data generation has increased at a staggering rate since 2001 [4, 5, 13]. A primary driver of this is decreased DNA sequencing costs [24], which have enabled the curation of large genetic datasets for personal and public health, ancestry inference, relative identification and advanced forensics. For example, there are whole genome sequences combined with medical records from approximately 400,000 US residents in the National Institutes of Health (NIH) All of Us Research Study [16]. Direct-to-consumer (DTC) genetic testing companies have much more data: 23andMe has 14 million customer sequences [1] and Ancestry has 25 million [3]. The increase in genetic data and improvements in technologies has resulted in the following aspirational “bold prediction” for 2030 by the NIH: “a person’s complete genome sequence along with informative annotations can be securely and readily accessible on their smartphone” [9]. But current genetic data use is commonly siloed with their own regulatory environments.

In this paper, we organize the uses of genetic data along four distinct ‘pillars’: clinical practice, research, forensic and government use, and recreational use [23] and make the argument that inconsistent and *leaky* regulation across these pillars introduces opportunity for genetic discrimination and privacy violations. Since 2020, the U.S. Department of Homeland Security has grown its genetic database by more than 1.5 million people, the majority of whom are people of color [8]. 23andMe suffered a data breach after a hacker accessed 14,000 customer profiles[1], primarily targeting individuals of Chinese and Ashkenazi Jewish descent [6]. These are serious concerns that must be addressed today given the number of recent proposals for genetic data applications in public infrastructure, (e.g., to assess intelligence proxies, like IQ [19]; insurance premiums [11]; or educational outcomes [10]). The only federal-level legislation currently in effect is the Genetic Information Non-Discrimination Act (GINA) of 2008, which prohibits the use of genetic information in employment decisions and discrimination based on genetic information in health insurance coverage but does not cover long-term care, disability, or life insurance. The question of how genetic data is being analyzed is critical when understanding what protections are necessary for individuals who submit their genetic data to these ‘pillars’.

From a scientific and modeling perspective, the analysis of genetic data is also critically limited to simple linear models correlated with trait data. We outline several concerns of this status quo in a case study in Section 3. As of now, the deployment of **AI technologies** for genetic inference is nascent. AI will, we believe, exacerbate the risks we outline as they will likely result in seemingly better trait prediction by modeling nonlinear interactions and taking advantage of gene-environment correlations [25]. However, the uncertainties associated with the basic scientific inference will be compounded because of the lack of interpretability and complexity of AI models. With the vast increase of genetic data collection, its increasing use (and misuse), and the complexities involved in its analysis and interpretation, comprehensive genetic data governance is urgently needed.

2 Risk Assessment Framework

Our goal in this section is to develop a risk assessment framework for genetic data governance. Similar to the aims of the Blueprint for the AI Bill of Rights [18], we motivate our framework with three central questions:

I. What values (i.e. moral principles and civil liberties) should be preserved?

(1) **Right to action**: The individual, only, has the choice to submit (and the freedom to not submit) their genome[2];
(2) **Ownership of the genome**: The individual owns their genome and controls the usage of their genome;
(3) **Right to privacy**: the individual has a right to privacy to their genome and inferences made from their genome;
(4) **Right to knowledge**: The individual has a right to know or *not* know inferences made from their genome;

[1] While genetic data was not accessed, ancestry information and relative names were[2]

[2] We use “genome” to refer to a physical DNA samples of an individual and any sequencing/genotyping data.

Staig@CHI, April 2025, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(5) **Opportunities for advancement**: Genetic data should not be used to deprive the individual of opportunities, including education, access to financial tools, insurance, housing, social services, and reproductive choices;
(6) **Benefits of inclusion**: The Belmont principle states "those who bear the burdens of research...should receive the benefits in equal measure to the burdens"[22], which applies to individuals and their genetic data.[3]

II. What are the vulnerabilities in the current system that can compromise these values?

(1) **Unsettled science:** The role of genetics in shaping complex traits is poorly understood, leading to an overstated role of genetics (genetic determinism), especially for behavioral and cognitive traits.
(2) **Rapid evolution of genetic data/methods:** Genetic data collection and analysis are advancing faster than legal protections, leaving gaps such as the exclusion of education discrimination from GINA.
(3) **Guilt by association:** DNA databases used in criminal investigations can implicate biological relatives, even if they have never submitted their DNA.
(4) **Geographical legislative patchwork:** Genetic privacy laws vary widely by state, creating legal uncertainties for individuals who move or share data across state lines.

III. What are the harms that arise as a result from the vulnerabilities?

(1) **Leakage to the family:** Genetic data can reveal information about relatives who never consented to sharing their DNA, affecting insurance, medical decisions, and identity security.
(2) **Loss of anonymity:** Genetic information can expose sensitive traits like race, gender, or disease markers, compromising privacy for individuals and their genetic relatives.
(3) **Loss of data control:** Private genetic testing companies control how genetic data is collected, stored, and monetized, often without transparency or clear user consent.
(4) **Misinformed actions:** Individuals may make critical medical, financial, or lifestyle decisions based on incomplete or evolving genetic interpretations, sometimes indirectly through relatives' test results.
(5) **Financial impact:** People may face financial burdens due to denied insurance coverage, legal defense costs, or extortion schemes related to genetic data leaks.

3 Case Study: Genetics and Educational Attainment

Researchers typically conduct genome-wide association studies (GWAS), which identify genetic variants that are associated with a trait and quantify the *effect* of each variant on the trait. The learned effect sizes of variants (also called *weights*) are directly interpretable[4] and can be used for trait prediction. The **polygenic score (PGS)**, the gold standard for trait prediction, is calculated as a weighted sum of genetic variants, with each variant scaled by its corresponding GWAS effect size. Any GWAS is heavily influenced by the choice of a reference dataset — analogous to training data in Machine Learning — which contains genetic data and their "ground truth" annotations of trait values, whether it be ancestry, disease status, height, etc. Critically, non-genetic factors of a cohort also influence a GWAS (e.g., environmental exposures and social determinants of health can introduce bias in a genetic study).

PGS are popular in the social/behavioral sciences for predicting social outcome traits, such as educational attainment (EA; number of schooling years completed by an adult). There is also interest in predicting standardized testing scores, performance in mathematics, and other traits with substantial environmental influences. A common metric for assessing PGS accuracy is the percentage of trait variance it explains: higher percentages indicate better predictive performance. A recent 23andMe EA PGS, based on data from over 3 million customers of European descent, explains 12-16% of the variance [5] in educational attainment (EA) [17]. In their test dataset, 50-70% of individuals with PGS scores in the top 10% for EA did, as predicted, graduate college. However, the PGS accuracy significantly decreases when applied to African American customers. This is an example of the commonly observed "portability problem" [14], where a PGS derived from GWAS in one population predicts poorly in another due to confounding[6].

There have been several calls to use EA PGS to inform education policy. Harden et al. propose the use of math-performance PGS to identify "leaks" in the education system: for example, by identifying high math PGS students who perform poorly, they claim educators could pinpoint *why* and *how* students are failing to reach their potential[7] [10]. Plomin & von Stumm take it further: they use the term "precision education" (akin to precision medicine) to propose a tailor-made, individualized education that is genetics-informed [19]. Statements like this, combined with statements such as "students with higher polygenic scores for years of education have, on average, higher cognitive ability, better grades and come from families with higher SES [socioeconomic status]" [20] are cause for concern because they invoke a sense of genetic determinism. However, other predictors (parents' educational status, socioeconomic status) explain similar amounts of variance in EA [12, 15] and—unlike DNA—are mutable through social policy changes.

Several of our values would be violated if children were required to submit their DNA (Right to Action) or educational opportunities were denied to children based on their genetic potential (Opportunities for Advancement). Through the vulnerabilities of unsettled science, the rapid evolution of genetic methods, and limited scope of current state genetic anti-discrimination and data privacy laws – which often differ in protections and rights from state to state – harms such as leakage to the family can occur and affect not only children but their families and future. Federal protections are unclear in this context since EA data is neither explicitly classified as

[3] Variant Bio is a biotech company that collects genetic data from Indigenous groups from around the world and participates in revenue sharing with the communities they collect data from.

[4] For example, in a GWAS for height, an effect size of 0.01 would mean that the variant increases height by an average of 0.01 cm.

[5] As a useful comparison, mother's education explains 15% of the variance in EA [12]

[6] Confounding in GWAS can be genetic (non-causal variants correlated with causal ones) or environmental (non-causal variants correlated with causal environmental factors), leading to potential statistical artifacts in effect sizes.

[7] An example they use: 31% of high PGS students in good schools take calculus, compared to 24% of the same-scoring students in poor-performing schools.

protected health information under HIPAA nor related to employment, housing, or insurance, which would fall under GINA. Other potential safeguards, such as institutional review boards (IRBs), the Food and Drug Administration (FDA), and the Federal Trade Commission (FTC), are also inapplicable. The lack of centralized regulation poses serious risks to individuals in similar future scenarios, where the issue of poor data and model governance leads to significant oversight and insufficient protections.

4 Recommendations

We propose three amendments to existing policy to ensure robust and future-proofed data governance. These address open privacy concerns, legislative scoping for policy changes, and best practices for bodies handling genetic data.

4.1 Recommendation 1: Redefining Genetic Data

Issue: Legal policy surrounding genetic privacy notably excludes deidentified or anonymized data from protection.

Recommendation: Given that we argue that genetic data is unique compared to any other identifying data, we suggest genetic data be defined using the following language: “Genetic data includes any information on an individual’s genetic traits, such as DNA/RNA sequences, gene expression, or data from biological samples—including relatives—regardless of format. **This data is inherently identifiable** and considered PII, as it pertains to unique biological attributes linkable to individuals or groups. De-identified, pseudonymized, or anonymized genetic data remains genetic data, recognizing the potential for re-identification through advanced methods.”

4.2 Recommendation 2: Extending Protections for Genetic Discrimination

Issue: GINA is a vital piece of federal legislation that protects against genetic discrimination in employment and health insurance domains. However, other domains in which there exists potential for genetic discrimination are *not* protected by GINA: other insurance domains (life, long-term care, disability), housing, and education.

Recommendation: We recommend expanding GINA’s protections beyond employment and health insurance. While California’s CalGINA covers housing, mortgage brokerage, and education, it still excludes life, disability, and long-term care insurance. Any and all GINA extension bills should explicitly protect these areas, along with education and any opportunities for advancement, ensuring **no** barriers to opportunity. Additionally, legislation should prohibit considering genetic risk for *complex traits* with significant environmental influences (e.g., cardiovascular disease) as preexisting conditions.

4.3 Recommendation 3: A Genetic Data Regulation Framework

Issue: Current regulations were designed to govern one application of genetic data—or Pillar—at a time. This has led to “leaky protections”, where the use of genetic data in one Pillar can affect opportunities and decisions in other Pillars (e.g., clinical tests being used in life insurance).

Recommendation: To prevent *leaky protections*, we propose a uniform regulatory framework covering all genetic data use, from collection to inference. We emphasize privacy rights to hold data-holding organizations accountable, not the individuals from whom data is collected [21]. We suggest that any entity must have prior regulatory approval to *collect* or *store* genetic material or data. Approval requires a clear commitment to individuals’ basic rights over their genetic data. After *entity approval*, we examine genetic tests and propose requiring entities to publish white papers detailing test procedures, quality control, inferential models, result presentation, and reproducibility. Next, individuals should be able to (1) request data removal, (2) opt-in/out of third party data transfers, (3) opt-in/out of being informed of incidental genetic discovery. Finally, genetic data ownership should extend only to approved companies acquiring it through mergers or bankruptcy. If no approved entity can manage the data, we propose a protocol akin to nuclear waste disposal[7], ensuring all samples, data, and models are irreversibly destroyed. This framework not only addresses leaky protections but also ensures ethical, safe science in the public interest while safeguarding the rights of individuals contributing to or analyzing genetic data.

References

[1] 2023. 23andMe Reports FY2023 Fourth Quarter and Full Year Financial Results. *23andMe Investor Relations* (25 5 2023). https://investors.23andme.com/news-releases/news-release-details/23andme-reports-fy2023-fourth-quarter-and-full-year-financial Press Release. Accessed January 16, 2025.

[2] 2023. Addressing Data Security Concerns – Action Plan. *23andMe Blog* (5 12 2023). https://blog.23andme.com/articles/addressing-data-security-concerns Originally published October 6, 2023. Updated December 5, 2023. Accessed January 16, 2025.

[3] Ancestry. 2025. Company Facts: The Many Ways Our Family Has Grown. *Ancestry Corporate* (2025). https://www.ancestry.com/corporate/about-ancestry/company-facts Accessed January 22, 2025.

[4] Alexander G. Bick, Ginger A. Metcalf, Mayo, et al. 2024. Genomic data in the All of Us Research Program. *Nature* 627, 8003 (March 2024), 340–346. https://doi.org/10.1038/s41586-023-06957-x

[5] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 7726 (Oct. 2018), 203–209. https://doi.org/10.1038/s41586-018-0579-z

[6] Rebecca Carballo, Emily Schmall, and Remy Tumin. 2024. 23andMe Breach Targeted Jewish and Chinese Customers, Lawsuit Says. *The New York Times* (26 1 2024). https://www.nytimes.com/2024/01/26/business/23andme-hack-data.html Accessed January 16, 2025.

[7] Cory Doctorow. 2008. Personal data is as hot as nuclear waste. *The Guardian* (January 2008). https://www.theguardian.com/technology/2008/jan/15/data.security Published January 15, 2008.

[8] Stevie Glaberson, Emerald Tse, and Emily Tucker. 2024. Raiding the Genome: How the United States Government is Abusing Its Immigration Powers to Amass DNA for Future policing. *Center On Privacy & Technology at Georgetown Law* (2024). https://www.law.georgetown.edu/privacy-technology-center/publications/raiding-the-genome/

[9] Eric D. Green, Chris Gunter, Leslie G. Biesecker, et al. 2020. Strategic vision for improving human health at The Forefront of Genomics. *Nature* 586, 7831 (Oct. 2020), 683–692. https://doi.org/10.1038/s41586-020-2817-4

[10] K. Paige Harden, Benjamin W. Domingue, Daniel W. Belsky, Jason D. Boardman, Robert Crosnoe, Margherita Malanchini, Michel Nivard, Elliot M. Tucker-Drob, and Kathleen Mullan Harris. 2020. Genetic associations with mathematics tracking and persistence in secondary school. *npj Science of Learning* 5, 1 (Feb. 2020), 1. https://doi.org/10.1038/s41539-020-0060-2

[11] Richard Karlsson Linnér and Philipp D. Koellinger. 2022. Genetic risk scores in life insurance underwriting. *Journal of Health Economics* 81 (Jan. 2022), 102556. https://doi.org/10.1016/j.jhealeco.2021.102556

[12] James J. Lee, Robbee Wedow, Aysu Okbay, et al. 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics* 50, 8 (Aug. 2018), 1112–1121. https://doi.org/10.1038/s41588-018-0147-3

[13] Ruth J. F. Loos. 2020. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications* 11, 1 (Nov. 2020), 5900. https://doi.org/10.1038/s41467-020-19653-5

[14] Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. 2017. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics* 100, 4 (April 2017), 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004

[15] Tim T Morris, Neil M Davies, and George Davey Smith. 2020. Can education be personalised using pupils' genetic data? *eLife* 9 (March 2020), e49962. https://doi.org/10.7554/eLife.49962

[16] National Institutes of Health. 2024. All of Us Research Program Strategic Goals. *All of Us Research Program* (2024). https://allofus.nih.gov/about/program-goals Accessed January 2, 2025.

[17] Aysu Okbay, Yeda Wu, Nancy Wang, Hariharan Jayashankar, Bennett, et al. 2022. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics* 54, 4 (April 2022), 437–449. https://doi.org/10.1038/s41588-022-01016-z

[18] Eunice Park. 2023. The AI Bill of Rights: a step in the right direction. *Orange County Lawyer Magazine* 65, 2 (2023).

[19] Plomin R and von Stumm S. 2018. The new genetics of intelligence. *Nature reviews. Genetics* 19, 3 (March 2018). https://doi.org/10.1038/nrg.2017.104

[20] Emily Smith-Woolley, Jean-Baptiste Pingault, Saskia Selzam, Kaili Rimfeld, Eva Krapohl, Sophie von Stumm, Kathryn Asbury, Philip S. Dale, Toby Young, Rebecca Allen, Yulia Kovas, and Robert Plomin. 2018. Differences in exam performance between pupils attending selective and non-selective schools mirror the genetic differences between them. *npj Science of Learning* 3, 1 (March 2018), 1–7. https://doi.org/10.1038/s41539-018-0019-8

[21] Daniel J. Solove. 2023. The Limitations of Privacy Rights. *Notre Dame Law Review* 98 (2023), 975. http://dx.doi.org/10.2139/ssrn.4024790

[22] U.S. Department of Health & Human Services (HHS). 1979. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf. Accessed: 2025-01-22.

[23] Zhiyu Wan, James W. Hazel, Ellen Wright Clayton, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley A. Malin. 2022. Sociotechnical safeguards for genomic data privacy. *Nature Reviews Genetics* 23, 7 (July 2022), 429–445. https://doi.org/10.1038/s41576-022-00455-y Publisher: Nature Publishing Group.

[24] Kris A. Wetterstrand. 2023. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *National Human Genome Research Institute* (2023). https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data Accessed January 2, 2025.

[25] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, et al. 2024. Advancing Multimodal Medical Capabilities of Gemini. https://doi.org/10.48550/ARXIV.2405.03162 Version Number: 1.