Journal of Law and the Biosciences, 1–45
https://doi.org/10.1093/jlb/lsad029
Original Article

Genetic data are not always personal—disaggregating the identifiability and sensitivity of genetic data

Johanna Rahnasto[1,2,3,*]

[1] Harvard University, Harvard Law School, Cambridge, MA 02138, USA
[2] University of Helsinki, Faculty of Law, 00101 Helsinki, Finland
[3] Roschier, Attorneys Ltd., OT IP & Technology, 00130 Helsinki, Finland

*Corresponding author. E-mail: johanna.rahnasto@gmail.com

ABSTRACT

In both the EU and USA, genetic data are recognized as a special category of data that requires heightened privacy protection. Identifiability and sensitivity are central pillars of the regulatory framework in both jurisdictions: the privacy concerns stem from the assumption that genetic data are capable of identifying the individual and reveals sensitive information about them. But not all genetic data are identifiable and sensitive, nor are genetic data necessarily different from other types of big data in terms of these issues. This article argues that a more nuanced approach is needed to assess the threat to privacy interests posed by uses of genetic data. The privacy interests involved should be distinguished in terms of proposed use, the amount of data in question, and its uniqueness and informational content. When these factors are disaggregated, it is clear that both regulatory schemes could better achieve their goals by focusing more on the ways genetic data can be used rather than on their status as a special category of data.

KEYWORDS: genetics, privacy, identification, sensitive data, data protection, genomic research

I. INTRODUCTION

Genetic data are everywhere—or soon will be. Who gets to use it for what and what rights does the individual have? These are important questions that have been debated ever since genetic data started to be available and have only become more topical with the rapid proliferation of big data genomics and elaborate data processing tools. However, the legal debate around these topics has given inadequate attention to the foundational question of what exactly constitutes genetic data and which parts of it should be protected. This article argues for more elaborate differentiation between different types of genetic data. Appropriate account for the properties of genetic data empowers the society to both advance more purposeful protection of privacy as well as enables productive uses of genetic data.

Balancing between privacy protection and encouraging useful research and applications has been a major point of debate ever since large amounts of genetic data started to become available in the 2000s.[1] Scientifically, genetic data are typically generated by sequencing the DNA or RNA in biological samples. These data are then analyzed by various statistical and computational methods, usually involving reference genomes and databases documenting the functions of different genetic regions.[2] Genetic sequences obtained from different individuals will have variations in their sequence, and these variations, together with other information about the individual, allow researchers to make conclusions regarding the functions of the genetic regions involved.[3] This information is central for basic research, but it can also be applied in, for example, drug development or marketing.[4] Furthermore, the information can be used to identify a person or to make conclusions or predictions regarding them, since genomic data has some predictive value for almost any human characteristic.[5] These include both diagnostic and predictive health data, information on ethnicity and family relations, and predictions on personality traits and demographic factors.[6] This kind of individual data can have value for the person's healthcare, lifestyle and reproductive choices, self-awareness, and curiosity, but it can potentially also be used to discriminate, stigmatize, or to grant or deny benefits.[7]

It is established that the privacy interests of individuals must be balanced with needs for access and use,[8] but what the optimal balance would be remains unresolved. Several recent developments have made these questions more complicated than before. As a result of biobanking, data sharing, and large-scale databases, an increasing amount of genetic data being processed is not obtained directly from a donor for the purposes

[1] William Lowrance & Francis S. Collins, *Identifiability in Genomic Research*, 317 Science 600, 602 (2007).

[2] *Genomic Data Science,* Fact Sheet, NIH Nat'l Hum. Genome Rsch. Inst. (Apr. 5, 2022), https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science [https://perma.cc/28B2-2MVB].

[3] *Id.*

[4] David Spiegel, *One of Google's Earliest Genetic Experiments, 23andMe, Paid Off — Here's What Will Make or Break Its Future*, CNBC (Jan. 25, 2022), https://www.cnbc.com/2022/01/25/how-one-of-googles-earliest-genetic-experiments-23andme-paid-off.html [https://perma.cc/K57P-M2QV]; Remi Daviet, Gideon Nave & Jerry Wind, *Genetic Data: Potential Uses and Misuses in Marketing*, 86 J. Mktg. 7 (2020).

[5] Daviet, *supra* note 4, at 13.

[6] *See,* eg *id.* at 11. *See also* Ellen Wright Clayton et al., *The Law of Genetic Privacy: Applications, Implications, and Limitations*, 6 J. L. & Biosci. 1, 20–26 (2019) (outlining various lawful uses of genetic data, including uses related to criminal justice, forensics, education, employment, family law, government benefits, immigration, insurance, occupational and environmental health, personal injury litigation and real property and commercial transactions).

[7] *See generally,* eg Gamze Gürsoy, *Genome Privacy and Trust*, 5 Ann. Rev. Biomed. Data Sci. 163 (2022) (outlining different uses of genetic data).

[8] *See,* eg Laura L. Rodriguez et al., *The Complexities of Genomic Identifiability*, 339 Science 275, 275 (2013).

of one project.[9] Direct-to-consumer genetic tests have brought down the costs of sequencing and made DNA tests accessible to the public at large.[10] The providers of these tests are increasingly seeking to use and share the sequencing data for commercial purposes, including research, marketing, and drug development.[11] Furthermore, progress of science has shifted focus from small genetic samples (eg one locus or one gene) to genome-wide sampling and whole genome sequencing.[12] Combined with the development of computer science tools, the trend has been toward so-called big data genomics.[13]

The need to balance privacy with utility is central to all debates over genetic data, from questions of genetic discrimination, data ownership, and data sharing to issues over commercial use, forensic use, data access, and informed consent.[14] Finding the right balance requires accurate understanding of the content known as genetic data or genetic information. In this article, I argue that genetic data should not be treated as a uniform data category, but instead different policies should be applied to different kinds of genetic data based on the identifiability, sensitivity, and intended use of the data. Identifiability relates to the question of whether the genetic data or the information derived from it can be connected to a particular individual in the absence of directly identifying information. Sensitivity relates to the type of information encompassed in genetic data and the question of what level of sensitivity it should be assigned to it by default or under specific circumstances.

In view of our current understanding of genetic data and the development of computer science, it is appropriate to revisit the question of whether and when genetic data warrants special treatment and special protections. This issue has been around ever since the start of DNA sequencing and genetic testing, and it has been noted that trying to define genetic data and restrict its use runs into difficult issues of scope and raises the question whether there is a rational basis for different treatment of genetic data compared with other types of personal data.[15] This article provides a novel mapping of the characteristics of genetic data in view of the statutory protections awarded to it in the USA and the EU and the complexities brought about by genomics and big data.

[9] Taner Kuru & Iñigo de Miguel Beriain, *Your Genetic Data Is My Genetic Data: Unveiling Another Enforcement Issue of the GDPR*, 47 Comput. L. & Sec. Rev. 105,752, 3 (2022), https://doi.org/10.1016/j.clsr.2022.105752.

[10] Spiegel, *supra* note 4. These developments can also be viewed as part of the larger trends of mobile health, biometrics, and personal health optimization. *See* Samuel Becher & Andelka M. Phillips, Data Rights and Consumer Contracts: The Case of Personal Genomic Services 19 (Oct. 25, 2022) (forthcoming), https://papers.ssrn.com/abstract=4,180,967.

[11] Spiegel, *supra* note 4; Daviet, *supra* note 4 (discussing ways to utilize genetic data in consumer marketing).

[12] Paul Quinn & Liam Quinn, *Big Genetic Data and Its Big Data Protection Challenges*, 34 Comput. L. & Sec. Rev. 1000, 1000–01 (2018). *See also European '1+ Million Genomes' Initiative*, Eur. Comm'n (Mar. 8, 2023), https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes [https://perma.cc/39JT-PAR8] (describing the EU's initiative of building an infrastructure to support genomic research and provide access to big genomic data).

[13] Quinn & Quinn, *supra* note 12, at 1000–01.

[14] *See generally* Zhiyu Wan et al., *Expanding Access to Large-Scale Genomic Data While Promoting Privacy: A Game Theoretic Approach*, 100 Am. J. Hum. Genet. 316 (2017) (discussing this balancing from a game theoretic perspective).

[15] *See generally,* eg Mark A. Rothstein, *Why Treating Genetic Information Separately Is a Bad Idea*, 4 Tex. Rev. L. & Pol. 33, 33–34 (1999) (finding that differential treatment of genetic information from other types of data is impractical and unjustified).

Identifiability and sensitivity are central concepts in how genetic data are currently regulated in both the USA and the EU, as discussed in Part II. They are major themes in both the rationale of the legislation as well as in defining its scope. If either of them is based on false premises, discrepancies between the law and the reality are bound to arise.

Part III disaggregates the concept of identifiability as it applies to genetic data. I argue that genetic data are not always identifying but that this presumption only applies to particular situations or types of genetic data. Furthermore, I argue that to the extent genetic data are identifying, it is not clear that this should entitle it to special protections, when similar prospects of identification are also present with respect to other types of data in our big data age.

Part IV unpacks the theory that genetic data would be uniquely sensitive. On closer look, most genetic data are not sensitive and the rationale of sensitivity only applies to certain kinds of genetic data. It may not even make sense to distinguish genetic data with direct health implications, because similar conclusions can often be made based on other data. It is also unclear to what extent protection of genetic data should or does cover inferences and likelihoods with a genetic component. Overall, these findings point to it being more rational to focus safeguards on types of use instead of the genetic origin or link of data.[16]

Part V brings together the findings regarding identifiability and sensitivity to identify different dimensions of genetic data that should be observed in determining what level of protection should be awarded to what kind of genetic data. These include data amount, uniqueness, informational content, and type of use. Based on these, I build a framework for considering the status of genetic data both in the context of individual use cases as well as for the purposes of improving and interpreting legislation and policy. Going forward, regulation of genetic data should be more nuanced and adaptive to increasing scientific knowledge and novel uses.

II. CURRENT REGULATORY FRAMEWORKS

II.A. Identifiability

Identifiability of genetic data determines the required privacy safeguards under both USA and EU law, but these jurisdictions also present two very different general approaches to privacy regulation.

The USA has adopted a sectoral model for regulating privacy, ie federal regulation is targeted to specified use cases.[17] The most important regulations for genetic data in the USA are the Health Insurance Portability and Accountability Act of 1996 ('HIPAA'[18]) and the Genetic Information Nondiscrimination Act of 2008 ('GINA'[19]). HIPAA provides a national standard for privacy and security of health data in the USA, but its rules only cover health plans, health care clearinghouses, health care providers as well as

[16] This Article can also be framed as promoting contextual privacy assessments in the case of genetic data. *See generally,* eg Helen Nissenbaum, Privacy in Context (2009) (discussing how privacy should be viewed contextually, not as absolute control of private information).

[17] *See,* eg Margaret Foster Riley, *Big Data, HIPAA, and the Common Rule: Time for Big Change?, in* Big Data, Health Law, and Bioethics 251, 260 (I. Glenn Cohen et al. eds., 2018).

[18] 42 U.S.C. § 1320d.

[19] 42 U.S.C. § 2000ff.

their business associates, and these are specifically defined and exclude certain types of entities.[20] GINA prohibits genetic discrimination in health insurance and employment, ie restricts how entities in these fields may use genetic information—the prohibited acts specifically include requesting, requiring, or purchasing genetic information.[21] The Common Rule[22] aims to protect the privacy of subjects of research projects.

HIPAA applies to protected health information (PHI).[23] PHI is subject to restrictions regarding its use and disclosure, but only with respect to the covered entities.[24] More stringent requirements protecting the individual's privacy may be imposed in state laws.[25] PHI includes all 'individually identifiable health information,' which is a subset of health information. It includes any information, including genetic information, that '(i) is created or received by [a covered entity]; and (ii) relates to the past, present, or future physical or mental health or condition of an individual . . . and (a) that identifies the individual; or (b) with respect to which there is a reasonable basis to believe the information can be used to identify the individual.'[26] HIPAA explicitly provides that 'health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.'[27]

Data can escape HIPAA's rules through the process of deidentification. HIPAA singles out 18 factors that can be removed from data to make it deidentified and thus no longer PHI.[28] Alternatively, even if some of these identifiers are retained, an expert may determine that the risk of identification is very small and thus the data are effectively deidentified, but as a starting point it is assumed to be identifiable if any of the identifiers are kept.[29] On the other hand, if all of the specified identifiers are removed, the data are no longer considered identifiable, and it is thus not PHI subject to the HIPAA rules, if 'the covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is

[20] June M. Sullivan & Shannon B. Hartsfield, HIPAA: A Practical Guide to the Privacy and Security of Health Data 10–11, 16–22, 31 (2020).

[21] *See* 42 U.S.C. § 300gg–53. *See also,* eg Louise Slaughter, *Genetic Information Non-Discrimination Act*, 50 Harv. J. on Legis. 41 (2013) (discussing the background, purposes and limitations of GINA).

[22] 45 C.F.R. 46.

[23] *See* 45 C.F.R. §160.103.

[24] 45 C.F.R. §§ 164.502, 164.504.

[25] A great number of states have enacted legislation addressing genetic privacy, either as part of general consumer data protection laws or in more narrow and targeted regulations. Many of these laws address genetic information through regulation of biometric data and define genetic information by reference to analysis of DNA and other nucleic acids. *See Genome Statute and Legislation Database*, NIH Nat'l Hum. Genome Rsch. Inst. (Aug. 3, 2020), https://www.genome.gov/about-genomics/policy-issues/Genome-Statute-Legislation-Database [https://perma.cc/CJR3-KFJ4].

[26] 45 C.F.R. §160.103.

[27] 45 C.F.R. § 164.514.

[28] 45 C.F.R. § 164.514. These are names, geographic subdivisions, dates, phone and fax numbers, email addresses, social security numbers, medical record and health plan beneficiary numbers, account numbers, license numbers, vehicle and device identifiers, URLs, IP addresses, biometric identifiers, facial photographs, and 'other unique identifying number[s], characteristic[s], or code[s]'.

[29] *See,* eg Sullivan & Hartsfield, *supra* note 20, at 23–27; *De-identification of Protected Health Information: How to Anonymize PHI*, HIPAA J. (Jan. 1, 2023), https://www.hipaajournal.com/de-identification-protected-health-information/[https://perma.cc/ZE5D-72AR].

a subject of the information.'[30] Reidentification is not fully prevented under HIPAA, since HIPAA allows the covered entity to code (ie pseudonymize) information so that reidentification is possible.[31] Thus, reidentification remains a risk (or option) under the HIPAA deidentification methods and thus being detached from the HIPAA requirements does not guarantee anonymity—rather, it presents a way to reduce the risks to a level that is considered acceptable.[32]

Overall, identifiability is a key factor in determining the scope of HIPAA protections.[33] However, the formal procedure for deidentification does not necessarily cover the parts of the data that make it identifiable. Thus, there is potential for mismatches regarding what is protected and what is left outside the scope of the regulation. In addition, HIPAA provides the 'actual knowledge' phrase regarding case-specific assessments of identifiability, but it is unclear how frequently and under what kind of decision-making process it is applied. Thus, understanding factual identifiability is material to assessing whether the provided level of protection is adequate particularly in light of the deidentification procedures, data sharing practices, and the overall scope of the legislation.

In the EU, data protection rules are largely harmonized by the General Data Protection Regulation ('GDPR'[34]).[35] The GDPR applies to personal data, which is defined through its capability of identifying an individual.[36] Pursuant to Article 4(1), 'an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier . . . or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.'[37] Genetic data are also a specific subcategory of personal data that relates to a person's genetic characteristics.[38] Article 9(1) of the GDPR prohibits processing of genetic data, whereas Article 9(2) allows it in specific circumstances.

Data are anonymous when they cannot identify an individual, and in that case the GDPR does not apply to the processing of the data. The GDPR only distinguishes between personal and anonymous data, and pseudonymized/deidentified data are considered personal data even though it has been processed to prevent direct identification, ie attribution to an individual. The GDPR does not apply to anonymous data, which does not allow an individual to be identified, and such anonymous data can be used freely. Recital 26 provides that when determining identifiability, 'account should be

[30] 45 C.F.R. § 164.514.

[31] 45 C.F.R. § 164.514.

[32] *De-identification of Protected Health Information*, *supra* note 29.

[33] The Common Rule, too, defines its protections through whether private information is 'identifiable.'

[34] Regulation (EU) 2016/679 of the European Parliament and of the Council of Apr. 27, 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) [hereinafter 'GDPR'].

[35] Member States may add more detailed requirements in certain areas, including the regulation of genetic data. *See* GDPR, art. 9(4). *See generally,* eg Fruzsina Molnár-Gábor et al., *Harmonization after the GDPR? Divergences in the Rules for Genetic and Health Data Sharing in Four Member States and Ways to Overcome Them by EU Measures: Insights from Germany, Greece, Latvia and Sweden*, 84 SEMINARS CANCER BIOL. 271 (2022) (describing the different national rules that affect, among other things, genetic research in the EU).

[36] GDPR, art. 4(1).

[37] *Id.*

[38] *See* GDPR, art. 4(13).

taken of all the means reasonably likely to be used . . . either by the controller or by another person to identify the natural person directly or indirectly.' The standard is thus what is found 'reasonably likely'—such assessment should consider 'all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.'[39] This is often seen not as a genuine limitation, but identifiability is rather assessed from an objective perspective: any data anywhere held by anyone that could result in identification of the data subject makes the data at hand identifying.[40] For the GDPR and GDPR-like frameworks, the open questions regarding identifiability of genetic data are thus, first, what kind of genetic data (if not everything) falls under the terms personal data and genetic data, and, second, when and how these data can be anonymized to the level required by the regulation.

Ultimately, the identifiability of genetic data is crucial under both types of frameworks for whether the regulation applies at all. This is a central question for the practicality of using genetic data. Clinicians, researchers, and companies who wish to use genetic data must assess these questions.[41] Having broad access to genetic data while complying with all applicable legal requirements can be a challenge both in terms of technical feasibility and costs,[42] but restricting potential use to safety zones can impede useful projects.[43] To be justified, such barriers should come with a corresponding benefit to the public or individuals.

II.B. Sensitivity

Sensitivity of genetic data is the central rationale behind current regulations. Many data protection laws provide special protection for 'sensitive data' (called 'special category' data under the GDPR). The concept is based on the recognition of types of data that pose higher risks for individuals in the form of potential for discrimination and other harms, especially from the perspective of vulnerable groups.[44] Under the GDPR, sensitivity is also understood more broadly as anything that would pose risks to the 'fundamental rights and freedoms' of the individual, ie it is not limited to discrimination, although preventing discrimination is a major justification for the limitations

[39] GDPR, rec. 26.

[40] *See* Case C-582/14, Patrik Breyer v. Bundesrepublik Deutschland, ECLI:EU:C:2016:779, ¶¶47–48 (Oct. 19, 2016) (holding that an online media service could 'likely reasonably' obtain the identity of an IP address holder from an internet service provider, because the law allowed a competent authority to obtain that information from an internet service provider in the case of a cyber-attack and potentially share it with an affected online media service). Note that this case was decided under the Data Protection Directive and thus applied the older standard of 'likely reasonably' instead of 'reasonably likely'—however, these have in practice been considered equivalent.

[41] Mahsa Shabani & Luca Marelli, *Re-Identifiability of Genomic Data and the GDPR*, 20 EMBO Reps. e48316, 2 (2019), https://doi.org/10.15252/embr.201948316.

[42] Kieran C. O'Doherty et al., *Toward Better Governance of Human Genomic Data*, 53 Nat. Genet. 2, 2–5 (2021).

[43] Some have viewed that researchers have a duty toward the public and the research subjects to make the best and fullest use possible out of the provided samples and data. *See,* eg Rodriguez, *supra* note 8, at 276. This would imply that either the researchers must have a clear view of what they are allowed to do or that they should constantly seek and test the boundaries.

[44] Quinn & Malgieri, *supra* note 133, at 1585.

regarding special categories of data.[45] Sensitive data are also protected on their own as a fundamental right in addition to the more instrumental discrimination-prevention approach.[46]

Under the GDPR, genetic data are special category data subject to Article 9 safeguards. Article 4(13) defines genetic data as 'personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question.' The Article 4(13) definition is not limited to data derived from biological samples, so genetic data can technically be interpreted expansively to include any information that has genetic implications. Furthermore, the GDPR only contemplate samples 'from the natural person in question,' so genetic data of family members seems to not be encompassed. Overall, the definition excludes genetic data that does not give (i) information unique to that person; or (ii) information about the physiology or health of that person. Thus, the GDPR practically distinguishes between three types of genetic data: genetic data with the special status subject to Article 9, genetic data that are ordinary personal data (identifying but not related to health or physiology), and nonpersonal (anonymous) genetic data.

In the USA, HIPAA applies to health information, to which genetic information is a subcategory. The idea is that health information is sensitive and thus needs protection. Genetic information includes information about the individual's genetic tests, the genetic tests of family members of the individual, or the manifestation of a disease or disorder in family members of such individual.[47] Information regarding age and sex is explicitly excluded from the definition.[48] Furthermore, a genetic test is defined as analysis of human DNA, RNA, chromosomes, proteins, or metabolites, which detects genotypes, mutations, or chromosomal changes. GINA and other nondiscrimination regulations also manifest the sensitivity attributed to genetic information. The definitions of genetic information in the USA regulations are quite comprehensive and scientifically accurate, but on the other hand the application of the rules is limited to certain entities. These limitations, too, along with the formal deidentification procedure under HIPAA, represent judgments regarding the sensitivity of the involved data.

III. DISAGGREGATING THE IDENTIFIABILITY OF GENETIC DATA

III.A. Why Identifiability Matters

Identifiability refers to the question of whether the data can unequivocally be connected to one individual. If data are identifying, the individual who was their source can be identified either directly or indirectly. This is clearly the case when genetic data are included in a patient file with the name of the individual. What remains unresolved is whether and when isolated or deidentified genetic data will remain identifying. When is the link to the individual permanently severed so that the data are anonymous?

[45] *Id.* at 1585–86.

[46] *Id.* at 1586–87.

[47] 45 C.F.R. § 160.103.

[48] There are also some further inclusions and exclusions in the statutes, for example regular diagnostic tests are excluded if they do not specifically detect genotypes. GINA relies on a similar definition. *See* GINA § 201.

The question of identifiability of genetic data is important for two main reasons: (i) it often determines what legislation is applicable (as discussed above), and (ii) it shapes and limits the interests of an individual in the data. This section examines the identifiability of genetic data from the perspective of common assumptions and the technical reality and prospects. I conclude that genetic data—while potentially highly identifying—is not *always* identifying and in any case no more identifying than many other types of data. Ultimately, I call for a more dynamic assessment of genetic identifiability. I begin with a brief discussion of the practical significance of identifiability.

Identifiability is a key concept in determining the benefits of privacy-based use restrictions. In the basic case, the benefit would be the protection of individuals' privacy rights, integrity, and anonymity.[49] A large part of why use of genetic data for various purposes is so controversial is its potential for identifying the individual.[50] Inappropriate uses and leaks of genetic data can produce far-reaching harm for individuals, which is an argument in favor of strong protections and a low threshold for holding the data identifying.[51] Even if genetic discrimination is illegal, the disclosure of genetic details may have adverse personal implications for an individual and their relatives similarly to disclosure of any other information perceived as private or sensitive—and genetic information does not meaningfully change over the course of one's life.[52] These issues tie closely to the questions of sensitivity discussed later, but from a pure identification perspective harms from unwarranted disclosures could include connection of the individual to a location, a crime or investigation, a health condition, an ethnicity, a relative or a research project. Depending on the circumstances, such connections may be something the individual would not want to be made. Data protection laws are designed to provide the individual with the discretion to decide when to allow such connections to be drawn and what those can be used for, subject to certain exceptions.

To the extent the individual is *not* identifiable based on the data (ie the data are anonymous), it may be assumed that the individual has no interest in how their data are

[49] *See,* eg Diana Liebenau, *What Intellectual Property Can Learn from Information Privacy, and Vice Versa*, 30 Harv. J. L. & Tech. 285, 288–90, 297–99, 302–03 (2016) (discussing privacy rights as forms of control, access restrictions, and contextual integrity, and comparing it to the regulation of intellectual property).

[50] Quinn & Quinn, *supra* note 12, at 1016. *See generally also* Benjamin T. Van Meter, *Demanding Trust in the Private Genetic Data Market*, Note, 105 Cornell L. Rev. 1527 (2020) (discussing ways to reduce harms to consumers from use of their 'anonymous' but identifiable genetic data and suggesting a fiduciary duty framework).

[51] *See* Luca Bonomi, Yingxiang Huang & Lucila Ohno-Machado, *Privacy Challenges and Research Opportunities for Genomic Data Sharing*, 52 Nat. Genet. 646, 647–48 (2020) (distinguishing between whether the data is misused to identify an individual in an undesirable context or whether it is used to infer knowledge about an already identified individual, and how these might occur in practice).

[52] Bonnie Berger & Hyunghoon Cho, *Emerging Technologies Towards Enhancing Privacy in Genomic Data Sharing*, 20 Genome Biol. 128, 1 (2019), https://doi.org/10.1186/s13059-019-1741-0; Quinn & Quinn, *supra* note 12, at 1003–04; James Brian Byrd et al., *Responsible, Practical Genomic Data Sharing That Accelerates Research*, 21 Nat. Rev. Genet. 615, 618 (2020). *See generally also* Richard Karlsson Linnér & Philipp D. Koellinger, *Genetic Risk Scores in Life Insurance Underwriting*, 81 J. Health Econ. 102, 556 (2022) (presenting that the prospect of genetic discrimination is not fully eliminated with existing laws).

used as long as it was collected by legitimate means.[53] This is also the starting point of privacy regulations. When the data are considered anonymous, it can legally be shared without regard to data protection laws.[54] The issue arising here is that often, in practice, the possibility of identification is not zero—but when is it low enough that the data should be free of the data protection requirements? Both the EU and USA regimes take some form of a risk-based approach to identification, but arguably the GDPR attempts to extend the identifiability of the data even too far, whereas HIPAA has been criticized for not effectively preventing reidentification, because the standards for deidentification are quite low and can be applied mechanically.[55] Expansive interpretations of the rules make sense from an anti-circumvention and privacy-prioritization perspective, but at the same time there is a danger that non-identifiable data becomes only a theoretical concept in certain contexts, and that valuable uses are prevented with little gains.[56]

These perspectives highlight that understanding identifiability is critical for regulating human genetic data in a sensible manner. It is important that we hold a realistic and accurate understanding of the risks of identification and assess the regulations against that background. To that end, I next discuss how identification may take place in practice.

III.B. How Genetic Data Can Identify Someone

The identification capabilities of genetic data together with modern computational tools came somewhat as a surprise to the scientific community that used to publish genomic datasets openly in the early days.[57] The standard has now shifted to strict data protection requirements, especially under the GDPR, but questions of identifiability remain an ongoing struggle.[58] The question of identification can be approached on three levels: data uniqueness, identification process, and practical identifiability. The first level addresses the required amount or type of genetic data to make the sample unique to an individual. The second is the procedural question of how identification can take place in practice. The third relates to the issue of how small possibilities or cumbersome processes should qualify as a relevant risk of identification.

[53] *Cf.* Jorge L. Contreras, *Genetic Property*, 105 Geo. L. J. 1 (2016) (arguing that property-like rights to genetic data should be replaced by rules regarding liability of researchers). As a borderline example, one could imagine a situation where data is collected with consent for one type of research, anonymized, and then distributed further for a different type of project that may have controversial motivations. One could make a case for how the original research subject has a valid interest in not contributing to or enabling the controversial project even if it is not possible to personally connect them to the data.

[54] Under HIPAA, the lower standards of deidentification are sufficient to break the link between the individual and the data from a legal perspective.

[55] *See,* eg Berger & Cho, *supra* note 52, at 1. *See generally also* Riley, *supra* note 17 (arguing that HIPAA is generally ill suited for the digital age and an environment where data can easily be combined from multiple sources).

[56] *See generally,* eg Contreras, *supra* note 53 (arguing that use of genetic data should be allowed by default and liability for use that is prohibited or goes too far would be sufficient to guard interests of individuals).

[57] Now open publication is only warranted with consent of the data subject (or in some cases, an ethics committee) or if the data is not identifiable, ie not personal data. Lowrance & Collins, *supra* note 1, at 601; Shabani & Marelli, *supra* note 41, at 4.

[58] *See generally* Shabani & Marelli, *supra* note 41 (discussing the identifiability of individuals based on genomic data).

As for data uniqueness, the DNA between two humans is 99.9 per cent identical, so most genetic data cannot be linked to any one person.[59] Yet, because of the large size of the genome and the way the 0.1 per cent of variation is sprinkled around, the consensus today is that any 'meaningful sequences of DNA' are identifiable personal data.[60] The issue is where to draw the line for when the data are no longer unique and identifying. The complete genome is always unique, but individual bases of the DNA are generally not.[61] Studies have shown that fewer than 100 single nucleotide polymorphisms (SNPs) can confirm by statistical means that the data are coming from the same source as an existing sample.[62] It is possible that this number can get even lower under some conditions. Thus, the threshold for uniqueness appears quite low in relation to the amount of variation that exists.[63]

Yet, any individual locus is unlikely to be identifying on its own, even if it relates to a phenotype. Any single genotype or genetic feature is typically not unique to the individual or their family—particularly genetic loci that are well understood and have been studied in larger populations.[64] Even slightly longer stretches of the genome used in simple genetic studies are often not identifying, because they are identical between various people.[65] Genetic data are not identifiable, if they only concern genotypes that are repeating in a population. However, it is usually the inclusion of demographic data that makes datasets valuable for research purposes, so isolating genetic data from other personal data has limited practical significance.[66]

Recognizing the non-identifying nature of limited amounts of DNA does still release from the data protection regulations isolated genetic sequences that do not bear unique identifiers and that are common in a population, even if they originally came from a particular individual and even if they reveal the presence of a disease. This kind of common genotype data that has been deidentified can thus also be considered anonymous. Instead of attempting to view the sequence as personal data, it is more appropriate to assimilate it to diagnostic criteria or general-level demographics: while they may apply to a particular individual and might be their personal data if attached to their file, they do not *point to* that individual on their own—identify them. Thus, looking at data uniqueness shows that not all genetic information is inherently connected to the individual who it came from.

[59] *Genetics vs. Genomics,* Fact Sheet, NIH Nat'l Hum. Genome Rsch. Inst. (Sep. 7, 2018), https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics [https://perma.cc/Z5CN-5FGK].

[60] Quinn & Quinn, *supra* note 12, at 1016.

[61] With the exceptions of identical twins and very rare mutations only present in one individual. *See id.* at 1002.

[62] Zhen Lin, Art B. Owen & Russ B. Altman, *Genomic Research and Human Subject Privacy*, 305 Science 183, 183 (2004) (estimating the threshold to be 30–80 statistically independent SNPs).

[63] Over 1 billion different SNPs have been documented in humans. According to some estimates, a typical human genome would contain about 3–4 million SNPs compared with a reference genome. Most of these have no known phenotypic significance. *See* NCBI dbSNP Build 155, https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi [https://perma.cc/D7VH-68VU]; Tuuli Lappalainen et al., *Genomic Analysis in the Age of Human Genome Sequencing*, 177 Cell 70, 71 (2019).

[64] Dara Hallinan, Michael Friedewald & Paul De Hert, *Genetic Data and the Data Protection Regulation: Anonymity, Multiple Subjects, Sensitivity and a Prohibitionary Logic regarding Genetic Data?*, 29 Comput. L. & Sec. Rev. 317, 319 (2013).

[65] Quinn & Quinn, *supra* note 12, at 1002.

[66] Minh Thu Nguyen et al., *Model Consent Clauses for Rare Disease Research*, 20 BMC Med. Ethics 55, 56 (2019).

On the level of the identification process, there are three main ways identification with genetic data can take place. I call these genomic matching, content matching, and demographic matching.[67] With complex datasets, different combinations of these basic types are possible and likely, as discussed later with respect to practical identifiability.

Genomic matching means that a genetic sequence (or a collection of genotypes) is matched to another genetic sequence (or a collection of genotypes). This is possible if the genetic data are extensive enough to make it unique to an individual.[68] Where the sample is unique, ultimately, a comparable sample could always be obtained from the individual to ascertain whether they are the source of the data.[69] It is debatable whether this possibility falls (or should fall) within the scope of identifiability from a legal perspective.[70]

More commonly, a genetic sample can be identified by matching it against a reference genotype, if such exists in a database or is otherwise accessible. Databases that contain enough information for this are numerous and have been growing fast, for example in criminal and military records and healthcare.[71] If either sample is linked with identifying data, this also identifies the source of the matched data and potentially provides further information about them.[72] Notably, the genomic data need not be complete or connected with an identifier to make it identifiable given the ease of data sharing and access as well as modern computational power and algorithmic tools.[73] For example, forensic databases typically only use 20 short tandem repeat loci for identification.[74]

A factor contributing to the potential for genomic matching is also the legacy data that is around from the time when genomic data were shared openly, without

[67] This kind of grouping has also been presented by Bonomi, *supra* note 51, at 647. Cf. Literature has sometimes also distinguished between identifying factors that are demographic/administrative, descriptive, or indirect. Lowrance & Collins, *supra* note 1, at 601.

[68] In practice, this type of matching has also utilized the correlation of Y chromosome genotypes and surnames, thus making use of familial connections. *See* Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321 (2013); Bonomi, *supra* note 51, at 647.

[69] Quinn & Quinn, *supra* note 12, at 1003–04.

[70] This relates to the question whether it makes sense to hold that genetic data 'exist' as part of the person who they are from or whether they must be extracted and analyzed to be held to exist. Authors who focus on the privacy aspects of biologic/genetic material generally instead of merely considering extracted data might support this proposition. For example, Simana discusses an 'individual genome', which encompasses both biological material and genetic information extracted from an individual. *See generally* Shelly Simana, *Genetic Property Governance*, YALE J. L. & TECH. (forthcoming 2022–2023), https://www.ssrn.com/abstract = 4,193,240.

[71] *See* Lowrance & Collins, *supra* note 1, at 601; *Frequently Asked Questions on CODIS and NDIS*, FBI, https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet [https://perma.cc/ZTM4-RMA9] (for information on forensic databases); Mitja I. Kurki et al., *FinnGen Provides Genetic Insights from a Well-Phenotyped Isolated Population*, 613 NATURE 508 (2023) (for information on large databases combining genetic and health information).

[72] Lowrance & Collins, *supra* note 1, at 601; Quinn & Quinn, *supra* note 12, at 1002–03.

[73] Quinn & Quinn, *supra* note 12, at 1002.

[74] *Frequently Asked Questions on CODIS and NDIS*, *supra* note 71. Short tandem repeats are short genetic sequences whose number (repeat count) varies between different individuals—for example, one person might have 12 and 12 repeats and another 14 and 16 at a specific locus. Such genotypes can be used to distinguish between different individuals, but the repeat count does not have any biomedical implications and thus it cannot be used to infer anything about a person's health, for example. *See generally* Sara H. Katsanis & Jennifer K. Wagner, *Characterization of the Standard and Recommended CODIS Markers*, 58 Suppl. 1 J. FORENSIC SCI. S169 (2013) (describing properties of the genetic markers used for forensic purposes).

realizing their personal nature. Individuals who participated in the early sequencing studies—and their relatives—can potentially be identified through this data.[75] While the number of persons directly affected is relatively small, the familial connections implied can extend these effects to a larger amount of people. Furthermore, this is a practical, cautionary example of how falsely assuming genetic data anonymous today could contribute to making *future* data more identifiable. This has been presented as a reason to approach the question of identifiability cautiously.[76] Overall, genomic matching becomes more likely the more genetic data societies generate and the more accessible the data are.

The second type of identification, content matching, refers to identification resulting from derivation of the individual's attributes *from* the genetic data.[77] Genetic data can be used to infer a person's ethnicity, genetic disorders, and various attributes, such as gender, blood type, hair color and texture, eye color, and certain facial features.[78] If such attributes are sufficiently many and specific, they may single out one person who matches the data.[79] Many of these attributes may be publicly available through various sources (including, eg social media), so once the information has been extracted from the genetic data, powerful algorithms might trace the identity of the genetic sample donor even without a second, matching sample.[80]

The third type, demographic matching, does not relate to the genetic data as such but addresses the fact that genetic data typically do not occur in isolation from other data. Efficient use of genetic data for most research requires other data about the individual.[81] Individual attributes, demographics, and health data are arguably the most important part of genetic datasets, because the genetic sequences alone offer little value without the ability to connect them to any real-world phenomenon.[82] In a typical example, this would be health data about a disease (eg this person has type I diabetes), but it could also be more neutral phenotypic data regarding normal metabolism (either on the molecular or everyday level) or appearance of the individual (eg height, eye color). Medical and research records are virtually always combined with directly identifying personal data, a participant number or other pseudonym, or a set of indirectly identifying data (for example, date and place of birth, a disease status, etc.). While this kind of identification does not follow from the genetic data as such and can also occur without any genetic data being involved (and is thus not the focus of this paper), this is a relevant aspect to recognize for the overall identifiability of genetic data. This is because part of the rationale for protecting genetic data is shielding individuals from being connected to specific genetic content or contexts. Where the

[75] *See* Lowrance & Collins, *supra* note 1, at 601; Shabani & Marelli, *supra* note 41, at 4; Rodriguez, *supra* note 8, at 275; Quinn & Quinn, *supra* note 12, at 1003–04.

[76] Shabani & Marelli, *supra* note 41, at 4.

[77] *See* Lowrance & Collins, *supra* note 1, at 601.

[78] *Id.*

[79] *Id.*

[80] *See generally* Rodriguez, *supra* note 8 (contemplating how genetic identifiability interplays with the public's expectations of privacy and modern computational methods); Gymrek, *supra* note 68 (demonstrating identifiability of deidentified genomic data with publicly accessible sources); Joanne Hinds & Adam N. Joinson, *What Demographic Attributes Do Our Digital Footprint Reveal? A Systematic Review*, PLOS One 1 (2018) (showing that internet use data can be used to reveal a plethora of information about an individual).

[81] Quinn & Quinn, *supra* note 12, at 1001–02.

[82] *Id.* at 1016.

data accompanying the genetic data are sufficient to enable reidentification of the source of the genetic data—as it often is[83]—it may in practice be secondary whether or not the genetic sequence itself points back to an individual.

Identification by demographic matching can occur via linking to nongenetic databases: demographic and health-related data accompanying the genetic data can be compared with these other data sources to find a match there.[84] Statistical methods can be quite powerful in narrowing searches so that an individual can be indirectly identified even if the combination of attributes may appear anonymous.[85] Sometimes only a few demographic factors have been enough for identification based on publicly available information.[86]

This ties closely to the third level of identifiability: the practical identifiability. This means asking how likely it is that an individual *will* be identified rather than focusing on whether it is theoretically possible. Ultimately, identifiability of genetic data is a question of how available and accessible the necessary further data or input must be to make the data identifiable. For example, when data are held inside one organization subject to strict data security measures, it may not be meaningfully accessible to anyone and thus the practical risk of identification is small. However, with larger genetic segments, the practical identifiability may also be high in the contexts where it is often most useful, ie as part of a dataset with comprehensive demographic and other background data, since research interests often support open sharing of (seemingly) anonymous data.[87] Factors affecting the practical identifiability of genetic data thus include (i) the type and amount of data the genetic data are stored with, (ii) ways to access the data (including data security), (iii) the types of personal and anonymous data otherwise accessible to the data controller, as well as (iv) the available computational power and sophistication. Considering that the last two categories evolve quickly and are difficult to assess reliably, there is a danger for error estimates.

This problem has been widely recognized in the research community. Several exercises of data subject identification have been recorded and commented on.[88] Individual contributors to complex DNA mixtures and large-scale research projects have

[83] Nguyen, *supra* note 66, at 56.

[84] Lowrance & Collins, *supra* note 1, at 601. *See also* Hinds & Joinson, *supra* note 80 (discussing how demographic attributes can be derived from seemingly uninformative data like digital footprints).

[85] *See generally* Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, *Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models*, 10 Nat. Commc'ns 3069 (2019) (reporting findings that even heavily incomplete datasets often allow identification of a person).

[86] Harald Schmidt & Shawneequa Callier, *How Anonymous Is 'Anonymous'? Some Suggestions Towards a Coherent Universal Coding System for Genetic Samples*, 38 J. Med. Ethics 304, 306 (2012).

[87] Quinn & Quinn, *supra* note 12, at 1016.

[88] For description and discussion of these projects, see generally, eg Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, 4 PLOS Genet. e1000167 (2008), https://doi.org/10.1371/journal.pgen.1000167 (documenting identifiability through summary statistics of genome-wide data); Yaniv Erlich & Arvind Narayanan, *Routes for Breaching and Protecting Genetic Privacy*, 15 Nat. Rev. Genet. 409 (2014) (presenting several computational techniques for finding out the identity of study subjects of genetic studies); Yaniv Erlich et al., *Identity Inference of Genomic Data Using Long-Range Familial Searches*, 362 Science 690 (2018) (predicting that a relatively small database would suffice to find familial matches for a majority of the population); Arif Harmanci & Mark Gerstein, *Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles Through Genomic Deletions*, 9 Nat. Commuc'ns 2453 (2018), https://doi.org/10.1038/s41467-018-04875-5 (discussing identifiability of aggregated sequencing data); Shabani & Marelli, *supra* note 41, 3 (summarizing several re-identification studies).

sometimes been identified in proof-of-principle projects even based on pooled data, which has forced databases to limit the amount of genetic data they share.[89] These examples highlight the dynamics and uncertainty surrounding practical identifiability. While the potential for identification is considerable, it is not all-encompassing nor a unique feature of genetic data—as noted above, the demographic data plays a significant role in the equation.

Thus, considering all the different levels—data uniqueness, possible identification processes, and practical identifiability—identifiability of genetic data is not straightforward but often requires elaborate analysis of the content of the data as well as accessibility of other data that may be linkable to the genetic data. Not all situations where genetic data appears result in the genetic data coming out as identifiable when these aspects are considered. This fact has been insufficiently recognized in legal debate. Next, I consider the possibilities of decreasing the chances of identification and even anonymizing otherwise identifiable genetic data to supplement this picture.

III.C. Anonymizing Genetic Data

There is a consensus that genomic data—ie more or less the entire genome of an individual—can never be truly anonymized.[90] The potential for anonymization is less settled with respect to incomplete genomes. As discussed above, modern computational methods have enabled identification based on ever smaller amounts of genetic data and accompanying information. Modern computational tools may also enable effective anonymization of genetic data.

In this context, anonymous means that the individual who was the source of the data cannot be identified. There are two basic approaches to anonymization attempts: decreasing *access* to identifiable data and changing the *content* of the original data.

The access-based approach relies on data security measures and pseudonymization techniques to prevent actual identification.[91] Pseudonymization means that any identifying data are key-coded and kept separately from the rest of the dataset, with limited access. It has sometimes been presented as a better option than attempted anonymization due to the uncertainties regarding genomic identification: the genomic data could potentially be matched to another sample at some point and thus remain at risk of identifying an individual. If all identifying information is deleted, the data subject will lose all possibilities to access and control the data. In contrast, if an identifying key is maintained as a link between the data and the source, the individual may still exercise some forms of control.[92] These same issues arise if identifying data are merely deidentified, ie stripped of directly or most likely identifying content without regard to the factual identifiability of the remaining data. There are also other, more elaborate tools to restrict access to certain portions of genetic data to reduce risks of identification.[93]

[89] Schmidt & Callier, *supra* note 86, at 306 ('[I]t is possible to reidentify seemingly anonymous DNA sequences by linking them with other publicly available qualifiers such as gender, age or zip code and then matching the linked DNA with records containing further identifying information, such as census records').

[90] O'Doherty, *supra* note 42, at 6.

[91] *See* Berger & Cho, *supra* note 52, at 2–3; Shabani & Marelli, *supra* note 41, at 3–5; Lowrance & Collins, *supra* note 1, at 601–02; Bonomi, *supra* note 51, at 649–50.

[92] O'Doherty, *supra* note 42, at 6.

[93] *See,* eg Bonomi, *supra* note 51, at 651 (reviewing different technical possibilities and noting that there are big differences in terms of vulnerability and cost between the options).

Some proposals, for example, involve black box type algorithmic tools that would limit the output viewable by a user to variables essential for drawing a conclusion rather than sharing the underlying more informative content and individual-level data.[94]

The content-based approaches are directed to permanently changing the nature of the data to make it non-identifying. These techniques fall into the categories of input-limitation and data degradation.

Input-limitation can be viewed as a more sophisticated version of deidentification and pseudonymization. It means that parts of the data that could make it identifiable are either permanently deleted or not collected to begin with. This could mean limiting the proportion of genomic data used and shared as well as collecting only strictly necessary demographic attributes. This would reduce the degree of data uniqueness, making less likely that the data singles out an individual. There is, however, an administrability issue, for the amount of data required for identification depends on the 'region and extent of genome covered, the density of mapping, the rarity of variants, the degree of linkage disequilibrium, and other factors.'[95] It would thus often be cumbersome to verify sufficiency of the measures, but this problem is not unique to input-limitation but concerns all means of anonymization. These approaches would also be contrary to the genome-wide and big data trends that are based on generating large amounts of data and sorting out what is relevant later. Large genome-wide studies are one of the only ways to accurately study rare conditions, so this type of limitations to research could be a big loss.[96]

Data degradation methods include data aggregation as well as various processing methods that reduce identifiability. As today's high-profile research tends to use genomic data, genome-wide approaches, and big data, there is a limit to how practical any data degradation methods are. The problem regarding value loss is particularly encountered with statistical degradation, ie aggregating the data on a very high level.[97] It is also very difficult to determine what the exact level of aggregation or degradation should be in order to achieve effective anonymization, and uncertainties are likely to remain.[98] Many aggregated data formats have been shown to remain identifiable despite aggregation attempts, partly because a very limited amount of identifiers can lead to a particular individual with modern computational tools.[99]

One processing-based degradation method is to snip any larger genomic regions into smaller pieces so that they cannot be put together. As a downside, this can also be counterproductive for the purposes for which the data may be valuable.[100] There may also remain a risk that the pieces will be recombined due to the statistical correlations between the genotypes at different genomic loci. The same problem arises with techniques that rely on only removing or masking certain variable loci—the masked genotypes may often still be inferred due to the statistical correlations.[101]

[94] *See generally* Charlotte Bonte et al., *Towards Practical Privacy-Preserving Genome-Wide Association Study*, 19 BMC Bioinformat. 537 (2018) (describing a method for restricting research output to significance of a variable).

[95] Lowrance & Collins, *supra* note 1, at 601.

[96] Bonte, *supra* note 94, at 537.

[97] Lowrance & Collins, *supra* note 1, at 601.

[98] Shabani & Marelli, *supra* note 41, at 2.

[99] Harmanci & Gerstein, *supra* note 88, at 7–8. *See also* Rocher, *supra* note 85.

[100] Lowrance & Collins, *supra* note 1, at 601.

[101] These correlating regions are called haplotypes. Bonomi, *supra* note 51, at 650.

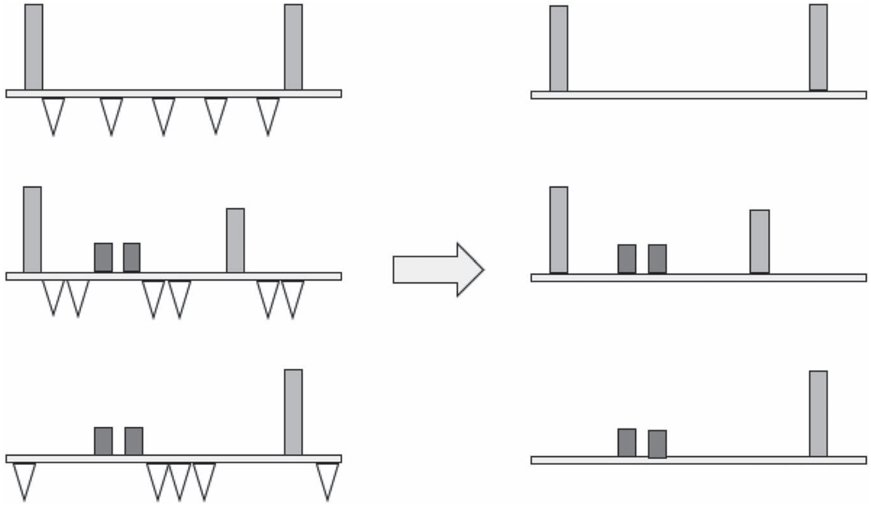

Figure 1. Removal of identifying genotypes. The lines are genetic data from different individuals and the bars rising from the lines represent variation of interest for the purposes of the study. The triangles below the lines represent small deletions or similar genotypes that enable identification of the study subject – and whose removal significantly reduces the potential for identification without affecting the informational content that matters for the study.

There is still hope for anonymization of broader samples. In practice, for example, it has been shown that personal RNA-Seq (gene expression) datasets can be anonymized by the removal of small deletion genotypes, which are largely responsible for making the original data identifiable.[102] Apparently, this can be accomplished without distorting the informational value of the data and as an integral part of normal data processing operations, rather than requiring burdensome additional steps.[103] While not without holes,[104] this is an example of technical means that can potentially be developed to deidentify even large sets of genetic data without making the data obsolete. Yet, it remains unclear whether reidentification could still be possible by contemporary and future machine-learning algorithms, for example.[105] The basic principle of this method is presented schematically in Figure 1.

In addition to the access- and content-based approaches, anonymization can be viewed from a purely practical and empirical perspective. From a practical perspective, the fact is that not anyone can run the analyses required to identify someone based on their genetic data: it requires skills, effort, computational tools and it may not even be legally possible.[106] As noted above, the amounts of data available online are starting to

[102] Harmanci & Gerstein, *supra* note 88, at 7.

[103] *Id.* at 7–8.

[104] Identification based on data obtained by other techniques (eg ChIP-Seq instead of RNA-Seq) is not fully eliminated by this method. *Id.*

[105] *Id.* at 8.

[106] Lowrance & Collings, *supra* note 1, at 602 ('The ease of identifying people from DNA or genomic data, without breaking laws, should not be overstated; it takes competence, perhaps a laboratory equipped for the purpose, computational power, perhaps linking to other data, and determined effort.').

be vast, which can make identification easier, but it is still not something a layperson could do on their laptop.[107] Practical risks mostly relate to data breaches, inadequate data security, and careless disclosures.[108] Data can also be released lawfully under court proceedings, law enforcement searches, and freedom of information laws—but to discourage extensive linkage and creation of identifiable data, those releases should be limited to only the strictly necessary genotype, for example.[109] The practical identifiability of genetic data can also be reduced indirectly by having fewer potentially identifying sequences openly shared—but this may also come at a cost to transparency and research.

From an empirical standpoint, some commentators have suggested looking to real-life datasets to determine whether and when identification is possible to find the boundaries.[110] A lot of the time identification is possible when the data are extensive enough.[111] This observation might support a risk-based approach.[112] Rather than focusing on the technical possibility of identification, the likelihood of someone actually carrying out the process and accessing the information would be factored in. At least sometimes the information required for identification would be difficult to access or analyze, and this might allow a conclusion that the data are anonymous for practical purposes.

One additional point to note here is the significance given to a potentially erroneous identification. What if someone attempts genetic matching, narrows the results down to one person, and *thinks* to have identified the data subject—but got it wrong? One view is that the harms of identification are not dependent on whether or not the data are correct, particularly if the harm is of social nature.[113] This would support deidentification approaches that retain the possibility of reidentification to prevent attribution of the data to an incorrect source.[114] The opposite view would be that as long as plausible uncertainty exists with respect to the correctness of the identification, the data should not be held identifiable.

These perspectives show that there are many ways to reduce the identifiability of genetic data to an arguably tolerable level, although uncertainties remain. It might be possible after all to render a genomic dataset anonymous under these approaches. Still, the cost of doing so would likely require clear benefits at the other end from the more liberal sharing and use of the data. Alternatively, more relaxed data protection requirements resulting from the data being anonymized might create savings that make

[107] *See* Shabani & Marelli, *supra* note 41, at 2 ('Yet, factoring in the time, effort and expertise needed, such attacks may still not be conclusive of the actual likelihood of re-identification'); Schmidt & Callier, *supra* note 86, at 306 (noting that often identification is only possible if the same individual provides another genetic sample that can be matched to an existing one, and these data would normally not be available to a research team); Erlich, *supra* note 88. *But see,* eg Gymrek, *supra* note 68 (using only publicly available data to complete identification).

[108] Lowrance & Collings, *supra* note 1, at 602.

[109] *Id.*

[110] Shabani & Marelli, *supra* note 41, at 2–3.

[111] *Id.* at 3.

[112] *Id.* at 2–3. Also harm-based approaches have been discussed as alternatives.

[113] *See* Karlsson Linnér & Koellinger, *supra* note 52, at 13 (considering a scenario where an insurance policy would be terminated due to 'false belief of low genetic risk based on an inaccurate genetic test').

[114] O'Doherty, *supra* note 42, at 6.

these procedures worthwhile.[115] A dynamic combination of content-based and practical assessments might be the optimal way to balance concern for bad faith identification attempts with practicality of using the data for acceptable purposes.

Not all approaches are compatible with current laws and legal interpretations. HIPAA provides this kind of opportunity for making an individualized risk-based determination that the risk of identification is low, but under the GDPR, access-based approaches do not qualify as anonymization, and the GDPR does not normally take into account the actual probability of identification.[116] Limitations to the scope of genetic data as personal data could still be achieved through strictly empirical perception of likelihoods and practicality. Such assessments have been explored by Wan et al. through game theory: one finding was that the potential pay-off from (and thus the incentive to engage in) reidentification can be affected by several policies and technical means.[117] This highlights the need to assess identifiability not only theoretically but also from the perspective of technical feasibility and practical data security. Such approach significantly diminishes the potential to consider all genetic data identifiable within the meaning of the GDPR and similar regulations.

The Court of Justice of the European Union ('CJEU') has also made some interpretations of the scope of personal data that become interesting when applied to genetic data. In *Nowak*, the CJEU stated that data are identifying 'where the information, by reason of its content, purpose or effect, is linked to a particular person.'[118] This seems to mean that there is no link to a *particular* person to the extent the data could have come from a number of different people—which is curious in the case of genetic data that could also identify a family or a twin without singling out just one person.[119] Furthermore, making a judgment under this standard requires a case-specific determination. This may be more costly and cumbersome for data controllers and processors than following a clearcut rule but offering this possibility may lead to a more efficient outcome than determining that anonymization is not possible at all.

Overall, identifiability of genetic data is not merely a technical question but intertwined with the data protection regulations and their definitions and standards. The scientific view of whether and on what conditions identification is possible is essential in making legal interpretations and neither of these should exist in isolation. The legal requirements should be responsive to developments in scientific understanding and technical possibilities. The next section further draws these different strings together and discusses how identifiability could be viewed more dynamically.

[115] Additional value from anonymization is created by the ideal of personal data minimization and respect for privacy and autonomy, to the extent such are deemed to be implicated under human rights frameworks—if the data are anonymized, they do not concern a person and thus the rights and freedoms of that person are not at stake.

[116] *See* Case C-582/14, Patrik Breyer v. Bundesrepublik Deutschland, ECLI:EU:C:2016:779, ¶¶47–48 (Oct. 19, 2016); GDPR, rec. 26.

[117] *See generally* Wan, *supra* note 14.

[118] Case C-434/16, Peter Nowak v. Data Prot. Comm'r, ECLI:EU:C:2017:994, ¶35 (Dec. 20, 2017). The case was about a person wanting access to written answers they had given in an exam. It is quite obvious the CJEU did not think of genetic data when writing the opinion.

[119] I do not go into the full details of the discussion regarding the identification of relatives and groups based on genetic data. In general, it has been called out as a severe weakness that European data protection laws do not protect genetic data of groups, only individuals. *See,* eg Hallinan, *supra* note 64, at 322–23.

III.D. Dynamic Approach to Genetic Identifiability

Genetic data are not always identifiable and, even when they are, they are not necessarily more so than other types of data. Because of this, regulation of genetic data should recognize more nuances in how tightly the genetic data are linked to the person from whom it originates. While a level of caution may be desirable in announcing any genomic data anonymous, it also seems that effective anonymization may be possible. Thus, there is limited justification for singling out genetic data as uniquely identifying. Instead, there are alternative ways to approach the identifiability of genetic data. These include the practical, context-dependent, risk-based, and dynamic perspectives, which are separate but overlapping ways to assess the personal nature of genetic data.

The *practical* perspectives have been described above with respect to both identifiability and anonymization. They focus on the practical feasibility and likelihood of accessing a particular dataset and drawing inferences from it. For example, if the data required for identification are only derivable by a highly sophisticated user and access to them requires illegal data breaches, the data might as well be deemed not to exist from an identification perspective.[120]

Context-dependent means that the identifiability of genetic data is not necessarily assessed against the whole universe of existing (and future) information. Rather, it is considered in terms of how accessible any identification-enabling data realistically is to a typical or a particular user and how likely it is that an identification attempt might be made. This can also be called a *risk-based* or *controller-subjective* approach, which takes into account the actual risks of identification as well as considers this risk from the perspective of the entity in possession of the genetic data. Under this approach, for example, rigorous data security measures could be deemed to diminish the identifiability of the data.[121]

Dynamic means that the data are not permanently labeled as either identifiable or anonymous. Instead, this is assessed continuously in view of the uncertainties involved. It would be possible to change the status of the data and the respective practices if concerns arise or new information is obtained. This would also involve a practical component: high-risk data might be assessed more actively, whereas older archived information might be sealed behind appropriate data security measures without similar continuous oversight.

Considering the reliance on technical procedures and assessments in determining the identifiability of genetic data, it would make sense to adopt a more dynamic and open-minded approach to how we treat genetic data. Sequencing technologies and analysis tools evolve constantly, so the legal approach should also remain dynamic. This means that identifiability should also be assessed observing current and foreseeable technologies and the types of data factually available rather than categorically for all

[120] As a weakness of this argument, it has been seen in practice that data breaches and leaks are not that rare, but patient information frequently becomes subject to unauthorized disclosures—and the remedies tend to be unsatisfactory. *See* Ifeoma Ajunwa, *Genetic Testing Meets Big Data: Torts and Contract Law Issues*, 75 Ohio St. L. J. 1225, 1225–27, 1252 (2014).

[121] Under the GDPR, such conclusion is currently not allowed since identifiability is an inherent characteristic of personal data, whereas data security just relates to how it is protected in practice.

genetic or genomic data.[122] A dynamic approach would also be preferable, because the identifiability of genetic data depends on multiple factors, including 'specific characteristics of datasets, the context in which the processing occurs, the technologies, expertise and incentives available and the mitigation strategies adopted.'[123] To ensure this approach is possible, regulators and commentators should recognize existence of different types of data and the risks and potentials of each. If it is technically possible to largely mitigate the risk of anyone identifying a person based on genetic data, the critical level of the risk of identification becomes crucial to understand and set at a tolerable level. The acceptable risk level should be reflected on the legal standards of deidentification.

Even with the many identification possibilities based on genome-wide data, most often identifiability would be more about probabilities than clear and certain identification.[124] One thing to ask, then, is how certain the identification must be,[125] and how much risk we are willing to tolerate in order to enable uses that are perceived as socially valuable. This problem with the required level of certainty highlights why genetic identifiability should not necessarily be treated any differently from other types of identifiability—at least not without appropriate justification.

Sometimes, the view is that the protection of an individual's genetic data should be almost absolute,[126] but such views hardly survive closer scrutiny and comparisons to how we treat other types of information: first, strictly individualistic conceptions quickly run into problems when the inferential and familial content of genetic data is considered.[127] Second, genetic data are not entirely unique in terms of its potential for inferences and predictions—instead many of the concerns apply to big data more generally.[128] Applying a principle of absolute genetic privacy threatens to expand the concepts of health data and genetic data in a nonsensical way, since a number of other factors also allow similar identification and inference of health information or likelihoods of health-related and genetic facts. For example, a facial photo of a person allows anyone who sees it to conclude something about that person's

[122] *See* Byrd, *supra* note 52, at 619 (noting that genomic identifiability is not static but depends on available resources).

[123] Shabani & Marelli, *supra* note 41, at 4.

[124] George Church et al., *Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection*, 5 PLOS GENET. e1000665, 3–4 (2009), https://doi.org/10.1371/journal.pgen.1000665.

[125] *See generally* Rosemary Braun et al., *Needles in the Haystack: Identifying Individuals Present in Pooled Genomic Data*, 5 PLOS GENET. e1000668 (2009), https://doi.org/10.1371/journal.pgen.1000668 (noting that, at the time, false positive rates of genetic identification were high due to the assumptions used in the process). This relates to the problem that negative identification (ie showing that the identified person was *not* the source of the data) might sometimes be impossible, especially where there is no pseudonym or similar means to identify the actual source.

[126] *See generally,* eg Róisín Á. Costello, *Genetic Data and the Right to Privacy: Towards a Relational Theory of Privacy?*, 22 HUM. RTS. L. REV. 1 (2022) (discussing privacy jurisprudence and case law that is based on an individualistic theory of self-control).

[127] *See generally,* eg Trevor Woodage, *Relative Futility: Limits to Genetic Privacy Protection Because of the Inability to Prevent Disclosure of Genetic Information by Relatives*, 95 MINN. L. REV. 682 (2010) (noting several problems with a highly protective privacy approach and calling for regulation that enables beneficial uses and prevents misuse).

[128] *See generally* I. Glenn Cohen & Harry Graver, *A Doctor's Touch: What Big Data in Health Care Can Teach Us About Predictive Policing* (Harv. Pub. L. Working Paper No. 19–41, 2019), https://ssrn.com/abstract = 3,432,095 (outlining ways in which big data is used to make predictions in nongenetic contexts).

health—including genetics—and potentially also other types of sensitive personal data.[129] However, as long as no one explicitly documents these inferences, that information is usually not held to exist and be available from a legal perspective and instead it is treated merely as a photo.[130]

Would it be possible to consider genetic data in the same way? Just because it is theoretically possible to infer or derive some information does not mean that information is in the possession of the data controller. This argument favors regulating actual use and processing instead of providing strict rules for a particularly defined subgroup of data.[131] Interpretation of identifiability of genetic data as expansively as some of the current law and scholarship suggests is inconsistent with how we treat other types of data and calls into question many underlying assumptions regarding what data can be deemed anonymous—leading to numerous practical problems.[132] If the underlying genetic exceptionalism is dropped, it makes more sense to focus on what kind of data *uses* are allowed instead of focusing on the status of genetic data as necessarily identifiable and highly sensitive.[133]

In summary, genetic data have high capability to identify an individual, but the practical risk of identification is very context-dependent and often quite low. Thus, it would make sense to aim for a dynamic understanding of genetic privacy instead of adopting rigid rules—either absolute privacy or formal deidentification steps. The resulting questions of appropriate risk levels and allowed uses tie closely with the discussion of how sensitive genetic data should be deemed. The next section unpacks the concept of sensitivity and its role in these issues.

IV. DISAGGREGATING THE SENSITIVITY OF GENETIC DATA

IV.A Rationale of Sensitivity

Genetic data are not always sensitive and, even when they are, they are not inherently more sensitive than many other types of data. To substantiate this argument, I first overview how sensitivity manifests in current regulation of genetic data and why sensitive data are protected. I then unpack the informational content of genetic data, noting that not all of the data are health related and showing that 'health-related data'

[129] For example, a facial photo might allow the conclusion that the person has or does not have Down's syndrome, which is a genetic condition—ie the photo would reveal some level of genetic information.

[130] For example, photos taken at events or public places are not usually considered to be anybody's health data despite the potential to infer health details from them.

[131] *See generally* Daniel J. Solove, Data Is What Data Does: Regulating Use, Harm, and Risk Instead of Sensitive Data 12 (Jan. 11, 2023) (draft), https://papers.ssrn.com/abstract=4,322,198 (arguing for regulating use of data instead of categorizing it as either sensitive or nonsensitive).

[132] One example of a similar issue is location data. Location data are abundant and everywhere despite their potential to reveal sensitive details about the lives of individuals. They are also very hard to effectively anonymize if identifiability is interpreted expansively to include also theoretical connections that might be drawn between different datasets. Yet, the GDPR does not usually treat location information as sensitive data—whereas some US laws do. *Id.* at 29. Similar comparisons can be made, eg to photos and internet use data.

[133] *See generally id.* (arguing for abolishing the category of sensitive data); Paul Quinn & Gianclaudio Malgieri, *The Difficulty of Defining Sensitive Data—The Concept of Sensitive Data in the EU Data Protection Framework*, 22 Ger. L. J. 1583 (2021) (promoting data protection rules that take into account the purpose for which sensitive data is to be used).

generally is a very ambiguous concept. Ultimately, a look at big data and inferences highlights that genetic data are not as unique as often presented.

What does it mean in practice that a piece of information is sensitive? Sensitivity can be viewed to include the aspects of personal autonomy, protection against use, protection of content, and protection of a right to stay anonymous/not be identified.[134]

From the perspective of personal autonomy, the idea is that sensitive data are so intimate and deeply connected to the person that it must be specifically safeguarded so that the person to whom it relates retains agency.[135] On this note, it has been suggested that genetic data should be protected more than any other category of sensitive data and that the legal protections provided would be inadequate because they allow processing of genetic data under the same conditions as any other health data.[136] This stance has been also put forward by the European Court of Human Rights in *S. and Marper v. The United Kingdom*, where it noted that genetic data—as well as cellular samples where it could be extracted—were of highly personal nature and contained much sensitive information and unique personal data.[137] The unambiguous grouping of genetic data as sensitive under the GDPR has been praised as an improvement to previous legislation, but at the same time many still view it as inadequate protection from a human rights perspective since it does not provide absolute control to those who the data concerns while being a significant invasion of their privacy.[138] Thus, here the focus is on the view that genetic data are uniquely personal and, in some way, capture the essence of a person. One criticism of this framework is that it lacks effective case-specific assessments of sensitivity of the data.[139] This has of course been purposeful: also the CJEU has noted that the EU legislature meant 'to assign a wide scope to [personal data], which is not restricted to information that is sensitive or private, but potentially encompasses all kinds of information, not only objective but also subjective . . . provided that it "relates" to the data subject.'[140]

Views focusing on protection against use encompass the fears against unauthorized and harmful uses of genetic data, particularly against discriminatory practices.[141] Many of these fears have been based on misconceptions regarding genetic determinism,[142] but many have also been explicitly validated and protected against by subsequent data protection laws. Fears of genetic discrimination relate particularly to stigmatizing genetic information, such as information on being a Huntington disease carrier, and the

[134] *See generally* Ignacio Cofone, *Nothing to Hide, but Something to Lose*, 70 U. Toronto L. J. 64 (2020) (discussing the rationale for privacy protections through debunking of the 'nothing to hide' argument).

[135] *See generally* Wendy Bonython & Bruce Baer Arnold, *Privacy, Personhood, and Property in the Age of Genomics*, 4 Laws 377 (2015) (discussing genomic information from the perspective of personhood and dignity, emphasizing human rights and fairness).

[136] Hallinan, *supra* note 64, at 319.

[137] S. and Marper v. The United Kingdom, App. Nos. 30,562/04 and 30,566/04, ¶¶72, 75 (Dec. 4, 2008), https://hudoc.echr.coe.int/fre?i = 001–90,051 [https://perma.cc/C3TE-F9X5].

[138] Hallinan, *supra* note 64, at 326. It should also be noted that not *all* genetic data are covered by the GDPR and its rules on genetic data, as discussed above.

[139] *See,* eg Solove, *supra* note 131, at 32.

[140] Case C-434/16, Peter Nowak v. Data Prot. Comm'r, ECLI:EU:C:2017:994, ¶34 (Dec. 20, 2017).

[141] *See generally* Rivka Jungreis, *Fearing Fear Itself: The Proposed Genetic Information Nondiscrimination Act of 2005 and Public Fears about Genetic Information*, 15 J.L. & Pol'y 211 (2007) (discussing the rationales of protecting against genetic discrimination).

[142] *See id.* at 230–31.

legal safeguards against discrimination do not fully alleviate these fears.[143] However, protection of genetic information as a separate category may not be the most sensible approach because of the way it is intertwined with manifestation of symptoms and other health information.[144]

One aspect of sensitivity is also what people *feel* is sensitive. Data show that consumers do not have a uniform perception of genetic privacy, and a number of consumers would be ready to treat genetic data similarly to other types of personal data as long as it is not used to their detriment.[145] It has also been reported that many people would be willing to access and use their genetic information to understand and optimize their health, but at the same time they are concerned about their privacy and potential misuse of the data.[146] According to some views, the procedural restraints on the use of genetic data are sufficient safeguards against infringement of privacy rights.[147] Overall, consumer perceptions toward use of genetic data appear to not always match with policies and legislation, with the ideas that data subjects have regarding how their data are factually used or with the factual risks for reidentification and misuse of genetic data.[148] While public perceptions of sensitivity and acceptable use are generally a cornerstone of privacy laws, they can also be assessed with some criticism in the context of such scientific information as genetic data.[149] These examples show that a more expert-driven determination of the sensitivity might be fruitful for genetic data, since the public does not uniformly understand genetic data to be particularly sensitive but it also does not necessarily base its sentiments on logic.

As for protecting the content of sensitive data, the content itself might be something the data subject would not want to be known by anyone, even if it is not used in any way. It is essential to have a sense of what kind of content genetic data entails to understand the scope of this interest. The question of what information can be extracted from the data is also relevant for smaller amounts of data that may not in themselves be identifiable, but which may result in a loss of privacy if disclosed.[150]

[143] *See generally* Annet Wauters & Ine Van Hoyweghen, *Global Trends on Fears and Concerns of Genetic Discrimination: A Systematic Literature Review*, 61 J. Hum. Genet. 275 (2016) (discussing the causes and content of fears regarding genetic discrimination).

[144] *Id.* at 281. This also relates to the issue that sequence data is not the only source of genetic information. *See supra* notes 129–130 and accompanying text.

[145] *See generally* Ellen W. Clayton et al., *A Systematic Literature Review of Individuals' Perspectives on Privacy and Genetic Information in the United States*, 13 PLOS One e0204417 (2018), https://doi.org/10.1371/journal.pone.0204417 (describing perceptions of US consumers regarding the use of their genetic information). However, these perceptions are distributed unevenly among different population groups and, eg racial minorities tend to have more critical views. *Id.* at 12. Conceptions of privacy and the protected values also differ between countries and cultures. *See generally* Lee A. Bygrave, *Privacy and Data Protection in an International Perspective*, 56 Scandinavian Stud. L. 165 (2010).

[146] Slaughter, *supra* note 21, at 62–63.

[147] Hallinan, *supra* note 64, at 326.

[148] Clayton, *supra* note 145, at 16–17.

[149] *See,* eg Daviet, *supra* note 4, at 20 (asking whether the public really does not care whether their genetic data is used in, eg marketing, or whether they are too ignorant to understand what is happening and thus need protection).

[150] Shabani & Marelli, *supra* note 41, at 2.

One question is whether genetic data should be differentiated from other types of data at all. This debate is sometimes called genetic exceptionalism.[151] Some, especially older, literature tends to present a mystified view of DNA and implicitly views it as part of the individual's essence, which often leads down a path where the individual is assumed to have a very strong interest in all uses of their genome.[152] Similar views may also arise in modern debates: in the era of big data, it is difficult to make the argument that a genomic sample would not be sensitive even if it is not *used* for the analysis of disease information, because such information is still relatively easily derivable from it.[153] Scholars have noted that 'data used in genomic research are by necessity personal and sensitive, as samples can unambiguously be traced back to an individual.'[154]

The extent to which these statements apply to *all* genetic data has been underexplored. A closer look at the content of genetic data calls into question the special status awarded to genetic data—either in and of itself or as a subcategory of health data. Assessments of sensitivity largely depend on the actual or assumed content of genetic data. Thus, the content is a central topic to understand in evaluating the best means to protect genetic privacy while allowing reasonable uses.

IV.B. Informational Content of Genetic Data

The informational content of genetic data is variable and potentially vast. Genetic tests can provide information regarding existing or developing conditions (eg Huntington's disease, PKU), risks of developing a condition (eg BRCA genes and breast cancer), risks of genetic conditions in future offspring (eg cystic fibrosis, sickle cell anemia), presence of genetic or genomic abnormalities (eg deletions), family relationships (eg paternity tests), and ancestral makeup (eg geographic origin and ethnicity of ancestors).[155] However, genetic data also contains information on common traits not considered particularly sensitive, like eye color, ear shape, types of cellular enzymes, and typical gene expression patterns. In addition, not all genetic data even has any meaningful content apart from the genetic sequence—at least not for present audiences.

Understanding this pool of potential content is crucial for commenting on whether all genetic data should be treated equally or whether different types could be treated differently. One obvious distinction between types of genetic data is whether the data are health related or not—as noted above, most legislation specifically targets

[151] *See generally* Sonia M. Suter, *The Allure and Peril of Genetic Exceptionalism: Do We Need Special Genetics Legislation?*, 79 Wash. U. L. Q. 669 (2001), (arguing that genetic exceptionalism is unnecessary and instead genetic information should be protected like any other medical information). *See also* Clayton, *supra* note 6, at 8 (noting that most commentators have been critical of genetic exceptionalism but legislators still keep enacting laws that distinguish genetic information from other data). *Cf. generally also* Nicolas P. Terry, *Big Data Proxies and Health Privacy Exceptionalism*, 24 Health Matrix 64 (2014) (questioning exceptionalism related to health data more generally, because with big data it is increasingly difficult to distinguish health information from other types of information).

[152] *See,* eg Catherine M. Valerio Barrad, *Genetic Information and Property Theory*, Comment, 87 Nw. U. L. Rev. 1037, 1085 (1992) (supporting property rights to genetic information due to its 'central importance of individual autonomy and self-determination').

[153] Quinn & Quinn, *supra* note 12, at 1005.

[154] Fruzsina Molnár-Gábor & Jan O. Korbel, *Genomic Data Sharing in Europe is Stumbling—Could a Code of Conduct Prevent Its Fall?*, 12 EMBO Mol. Med. e11421, 3 (2020), https://doi.org/10.15252/emmm.201911421.

[155] Regulations Under the Genetic Information Nondiscrimination Act of 2008, 75 Fed. Reg. 68,911, 68,916 (Nov. 9, 2010) (codified at 29 C.F.R. pt. 1635).

health-related genetic data and this is what most people think of when they discuss genetic data.[156] Also other than health-related genetic data can be considered sensitive (eg ethnicity data) and such may fall under specifically protected classes under applicable legislation. The arguments presented here regarding health data largely apply to those types of data as well, but for illustrative purposes the focus below will be on health. Health also warrants some special consideration because information on certain health risks is often most accurate and accessible by genetic test, the fact that a health-related trait is genetic has special implications, the information has potential to affect the data subject's life in several ways and health-related data generally are so abundant in contemporary society.[157] This being said, it is not at all clear what qualifies—or should qualify—as health information in the context of genetics.

Genetic tests testing for disease genotypes of a particular patient clearly produce health data. However, in research, the link between a genotype and a disease is often not yet established. At the level of data collection, the genetic data might not yet be personal or health related as such. But it is possible to consider that the data are turned into health data if, after the analyses, a genotype is associated with a health-related phenotype (ie trait). Establishing such links is indeed the goal of many genetic studies.[158] But not all studies come out with a finding of a plausible link, or the associated trait might not directly relate to health—or it might only become health related in subsequent studies or when combined with other information. To understand this better, let us consider a concrete example.

A classical genetic study could be set out by sequencing a specific locus in the genome (eg a part of a gene) and collecting the blood level of a specific metabolite from the participants. Several considerations affect the sensitivity issues involved. The first question is whether or not the genetic data collected—ie information on the sequence or structure of the specific locus—has any known function. Such function might have been identified in previous studies. For example, the gene can be associated with numerous functions—which might thus potentially be relevant also for the studied locus. There can also exist other studies regarding the locus in question, linking it to outcomes with respect to other conditions than the one studied here—thus, genotyping individuals with respect to this locus potentially reveals something about their susceptibility for another condition. In case previous studies exist, it might justifiably be asked how strong or direct any connections should be to make our data health data. Is a simple statistical association found in one study enough? Or should there be a broader scientific consensus and an empirically proven mechanism of effect?[159] In case no such

[156] *See,* eg Yael Bregman-Eschett, *Genetic Databases and Biobanks: Who Controls Our Genetic Privacy?*, 23 Santa Clara Comput. & High-Tech. L. J. 1, 7 (2006) (declaring that genetic data is a subclass of medical information, but at the same time going on to discuss types of genetic data that are not medical by nature).

[157] Some of these points can be countered by arguments presented later regarding why health-related genetic data may not be so special—consider the example from above regarding how genetic information could be inferred from a photograph. *See supra* note 129.

[158] Quinn & Quinn, *supra* note 12, at 1002.

[159] This relates to the significance of the correctness of the data: some forms of privacy harm may occur even if—or especially if—the private information turns out to be incorrect. This could be an argument in favor of protecting personal data more broadly without necessarily looking into its content or correctness, because discrimination and other forms of harm can also be based on erroneous facts or conclusions. *See* Danielle Keats Citron & Daniel J. Solove, *Privacy Harms,* 102 Boston U. L. Rev. 793, 839 (2021) (discussing the harms that may result from dissemination of incorrect information).

prior study regarding the polymorphism exists, genetic data regarding this locus would not be health data on their face value—at least not yet.

What is known about the genotyped locus is only a first aspect of information to be considered. Where the study also collects information about the metabolite in the blood, it could be asked what is known of that metabolite. Perhaps the level or type of the metabolite itself reveals further information about the genome of the study subject. For example, certain properties of the metabolite could be linked to another genotype, which may have health implications—or the metabolite itself could have known health implications, which can now also be attributed to a genotype. There may also be differences regarding what is known regarding the phenotype resulting from the metabolite: how the phenotype or the metabolite manifests in the body or life of a person may either be significant or barely understood, if any effects even exist.

Ultimately, after the study has been completed, if the trait/locus is simple enough or well understood, the genotype can possibly be inferred from the phenotype and vice versa, and the individual can be associated with certain likelihoods of having further genotypes and phenotypes.[160] For genome-wide studies, the basic considerations are the same, but the amount of loci is immensely larger and most of them will turn out not to have any connection to the measured phenotype. Despite that, genotypes of those loci will be part of the data and will at least theoretically allow inference of traits or likelihoods associated with these loci either at present or in the future.

This thought experiment highlights the immense number of paths one could take in a genetic study to find health-related information relating to genetic data. Despite this information often being embedded in the genetic data, it can also exist separately and independently from the genetic data, in which case it would typically be considered 'merely' health data or even regular personal data, not genetic data. Similar logic applies to other types of genetic data: it is difficult to decide where genetic data ends and other types of data begin, and when each should be awarded the status of 'sensitive data' or even genetic data, where the definition of genetic data is not limited to nucleic acid sequences. The next section addresses these issues.

IV.C. Scope of Health-Related Data

In view of these various levels of genetic information, two aspects have to be noted: not all genetic data has such informational value that it makes sense to consider them sensitive—even if the data have some connection to health. Furthermore, as the regulation of genetic data is tied to the provision of health-related information in many contexts, it is important to discuss what types of information are and should be considered health-related genetic data—and where the link to health is so vague that it should be out of the scope of this data group.

First, not all genetic data are that sensitive. Based on the above discussion, much of the information produced by genetic analyses relates to nonsensitive everyday traits (such as eye color or ear shape that can also be detected by naked eye), normal cellular functions that do not tell anything of consequence about the individual or markers that only provide information in view of statistical correlations. For example, Hallinan et al. have argued for the position it does not make sense to categorically raise genetic

[160] Hallinan, *supra* note 64, at 319.

data to a special status, especially not higher than other sensitive information.[161] The informational content of genetic data can be vast, but many of the things contained in the data are not particularly meaningful in our society. Furthermore, it would be practically impossible to prohibit or even significantly restrict 'processing of all data with a genetic element,' since this would extend to a vast amount of data that are constantly used in modern society, including information about a person's sex, looks, or origin.[162] Regulation efforts based on the notion of absolute protection of genetic information are in danger of only imposing pretextual control due to this discrepancy with scientific reality.[163]

To the extent some genetic data can unambiguously be considered health data, not all health data are equally sensitive.[164] Generally, sensitivity correlates with the health implications and social stigma (including potential for discrimination) associated with the condition.[165] It is difficult to draw exact lines to which genetic conditions should be regarded as serious or not,[166] but some genetic conditions have fundamental effects on one's life and health (eg chromosomal abnormalities), whereas others are practically meaningless (eg one genetic marker with a low association to a risk of a manageable disease). Yet, as a flipside, the more prominent and serious a genetic condition is, the less likely the carrier is to be able to keep information about it private even if they wished to.[167] Even if such information is considered sensitive health information, the privacy protection should not be contemplated in isolation of the social reality.

Drawing a line to the scope of health data also creates difficult dilemmas. The main issue is whether the link of the information to a health status is real or too distant to count.[168] The EU's Article 29 Working Party (WP29) held a very broad understanding

[161] *Id.* at 326 ('Depending on context, many other forms of data can be far more privacy sensitive than many forms of genetic data. Knowledge of an individual's HIV status, for example, is far more privacy sensitive than, for example, knowing an individual's eye colour.').

[162] *Id.* at 327.

[163] This is a common line of criticism regarding the GDPR, see, eg Solove, *supra* note 131, at 41 (criticizing the GDPR for singling out arbitrary categories of 'sensitive data' while not accounting for other types of data that can be used as proxies for the same purposes); I. van Ooijen & Helena U. Vrabec, *Does the GDPR Enhance Consumers' Control over Personal Data? An Analysis from a Behavioural Perspective*, 42 J. CONSUMER POL'Y 91 (2019) (noting some shortcomings with individual control and the GDPR).

[164] *See* Solove, *supra* note 131, at 32–33.

[165] *Id.*

[166] *See generally*, eg Felicity K. Boardman & Corinna C. Clark, *What is a 'Serious' Genetic Condition? The Perceptions of People Living with Genetic Conditions*, 30 EUR. J. HUM. GENET. 160 (2022) (examining which genetic conditions qualify as serious for the purposes of prenatal diagnosis).

[167] For example, the person may require disability-related accommodations in their everyday life, or the symptoms or treatment of the condition may be visible on their body in a way that does not allow the person to exercise autonomy over who has access to this health/genetic information. *See generally* Jackie Leach Scully, *Disability and Genetics in the Era of Genomic Medicine*, 9 NAT. REV. GENET. 797 (2008) (providing a nuanced overview of the interplay between genetics and disabilities); Aisling De Paor & Peter Blanck, *Precision Medicine and Advancing Genetic Technologies—Disability and Human Rights Perspectives*, 5 LAWS 36, 11 (2016), https://doi.org/10.3390/laws5030036 (highlighting the need to account for the social dimensions of disability, not just the genetics).

[168] *See* Quinn & Malgieri, *supra* note 133, at 1598 (differentiating between health-related data that are 'intrinsically' sensitive and data that have a 'computational distance' to sensitive information).

of health data in its 2015 health data guidance.[169] With regard to likelihoods, the WP29 stated that disease risks are health data even if they are inferred from data that itself is not health related.[170] This would mean that health data are created whenever a controller *uses* personal data to identify a health-related risk—or, possibly, a lack of such risk, since also confirmation that someone is 'healthy' qualifies as health data.[171] The guidance also notes that general data from which inferences could potentially be drawn 'do not have to be treated as health data . . . [but] the systematic analysis of such . . . for the purpose of diagnosis/health risk prevention or medical research certainly qualifies as the processing of health data.'[172] At the same time, the guidance holds that 'data about the purchase of medical products, devices and services' is health data when a health status *can be* inferred from the data.[173] Thus, the WP29 fails to clarify the difference between whether health inferences only *can* be drawn or whether they in fact *are* drawn in a specific context. Later in the same document, the WP29 lists both raw data based on which health data can be derived as well as any conclusions drawn about a person's health as 'health data.'[174]

In its profiling guidelines, the WP29 noted that sensitive health data may be created by inference based on a person's food shopping—but at the same time it implied that such health data are only held to exist (and be subject to respective data protection rules) if such data in fact 'are inferred.'[175] This guidance would thus imply that merely the potential for combining pieces of information to deduce sensitive data does not make the initial data sensitive. The CJEU's decision in *Nowak*[176] assigned a wide scope to personal data and potentially increased the scope of health-related data as well, but there are conflicting views regarding whether it really increased protection of inferred data.[177] Thus, the exact scope of health data in the EU has been murky for some time, but judicial interpretations tend to be expansive. Currently, the grounds for holding

[169] *Health Data in Apps and Devices*, Annex to Letter from the ART 29 WP to the European Commission, DG CONNECT on mHealth, Art. 29 Data Prot. Working Party 2–3 (Feb. 5, 2015) https://ec.europa.eu/justice/article-29/documentation/other-document/files/2015/20150205_letter_art29wp_ec_health_data_after_plenary_annex_en.pdf [https://perma.cc/KJX9-L3SG].

[170] *Id.*

[171] *Id.* The treatment of information regarding the *absence* of a disease as health information is very problematic if potential health-related inferences are included within the scope of health data. This type of information is virtually impossible to avoid in various contexts. For example, simply knowing that a person eg has a specific job quickly rules out several health conditions that they might have. Thus, this interpretation is unworkable when combined with the treatment of inferences and potential inferences as health data.

[172] *Id.* at 3.

[173] *Id.* at 2.

[174] *Id.* at 5.

[175] *Guidelines on Automated Individual Decision Making and Profiling for the Purposes of Regulation 2016/679*, Art. 29 Data Prot. Working Party 15 (Feb. 6, 2018), https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id = 612,053 [https://perma.cc/5PPY-CDUZ].

[176] Case C-434/16, Peter Nowak v. Data Prot. Comm'r, ECLI:EU:C:2017:994, (Dec. 20, 2017). Earlier case law of the CJEU could be interpreted to reject the idea that inferences would be covered and only support the personal data status of the underlying input data. *See* Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI,* 2019 Colum. Bus. L. Rev. 494, 522–27 (2019).

[177] Potentially, inferences receive limited practical protection because data subject rights only apply to them depending on the context. Wachter & Mittelstadt, *supra* note 176, at 541–42.

any information unequivocally nonsensitive or not health related under the GDPR are shaky. Similar line-drawing issues also arise in other jurisdictions.[178]

Thus, not all genetic data are health-related or otherwise sensitive. Furthermore, it is unclear what exactly should qualify as health related or sensitive. The categorical singling out of genetic data on this basis thus appears artificial. There are still certain aspects of genetic data that deserve more detailed discussion—these relate to the interconnections of genetic data with different likelihoods and inferences, which are discussed next.

IV.D. Likelihoods, Inferences, and Sensitivity

Genetic data can be used to make predictions about the future of the person from whom the data was extracted. Most genetic information is not deterministic and only allows conclusion of likelihoods.[179] These likelihoods operate on three different levels. There is the temporal aspect of the data possibly *becoming* sensitive in the future, the present sensitivity of a likelihood of *having* or *developing* a certain trait, and the potential for *inferring* something that qualifies as sensitive either from or relating to a person's genetic data. Examination of these aspects reveals that even where genetic data are unambiguously health related or otherwise within the definition of 'sensitive,' their content is often no more sensitive than other data.

Starting with the temporal issue, it remains unresolved whether the mere possibility of the genetic data having informational value at some point in the future is enough to render the data sensitive at present. We cannot completely rule out that parts of the genome currently considered 'junk'[180] could in the future be determined to have a function. If there is genetic variation, there could also be phenotypic consequences. Thus, the genetic data might become 'personal' upon the discovery of such consequences in the sense that, if the identity of the source is available, the genetic data would provide information on an identifiable natural person. Scholars have long warned that genetic data that are seemingly harmless in one context may become highly sensitive as our understanding and analysis methods develop and it becomes easier to extract meaningful information from small segments of the genome.[181]

Thus, is the potential to *become* health data enough to require treatment of the data as such even in the absence of information that would at present link the genetic data to a health status? One might be inclined to say yes just to be safe—after all, there is the risk that if insufficient protection is awarded now, it cannot be reinstated when it could be relevant. For example, the European Court of Human Rights has considered that

[178] The HIPAA defines PHI quite broadly as information, including demographic data, relating to a person's 'physical or mental health or condition.' 45 C.F.R. § 160.103. However, some commentators have noted that this would practically only encompass medical diagnoses, not everything health related more broadly. Some state laws are broader. Solove, *supra* note 131, at 30.

[179] *See generally* Pamela Sankar, *Genetic Privacy*, 54 Annu. Rev. Med. 393 (2003) (arguing that genetic exceptionalism has largely been based on a false belief in genetic determinism).

[180] Noncoding portions of the genome (that are not translated into proteins) are sometimes termed 'junk DNA', but these regions are also constantly attributed various functional roles. There probably are parts of the genome that genuinely 'do nothing,' but it is difficult in practice to ensure whether a sequence is truly obsolete or if we only have not discovered its role yet. *See generally,* eg Alexander F. Palazzo & T. Ryan Gregory, *The Case for Junk DNA*, 10 PLOS Genet. e1004351 (2014), https://doi.org/10.1371/journal.pgen.1004351.

[181] Hallinan, *supra* note 64, at 319.

merely the possibility for future use is enough to make storing (presently meaningless) data a privacy violation.[182] The GDPR, too, is not limited to data presently considered sensitive—ultimately all that matters is identifiability and whether the data are linkable to one of the special categories. For HIPAA, similar logic can be applied although its deidentification framework encourages more restrictive views regarding future implications of the data. In addition, HIPAA only applies to health information to begin with, and that comes with assumed sensitivity as the rationale. Holding the possible future sensitivity of data as equal to present sensitivity has also evoked criticism because it does not account for the probability of such sensitivity practically arising.[183]

The second aspect of sensitivity to be noted here is the informational content of likelihoods. Consider that a genetic test reveals that Mary has a 20 per cent likelihood of developing colon cancer. Is the genetic data behind this prediction more than, less than, or equally sensitive as knowing for certain that Mary has colon cancer? What if the likelihood resulting from the genetic data is 50 per cent? Or 80 per cent?[184] One might say that likelihoods are not as relevant as knowing a certain diagnosis, because they do not reveal any factual information about the data subject—they are just statistical predictions based on some genetic loci.[185] Still, intuitively, a higher probability is more sensitive than a lower one, since it is more likely to be true and, perhaps more importantly, more likely to be acted upon—which can result in harm for the individual even if the risk turns out not to materialize.[186] On the other hand, a low probability of one fact is a high probability of the opposite: simplistically, a 20 per cent probability of developing colon cancer means that the person does *not* develop it with an 80 per cent likelihood. This information may be equally useful—or harmful.[187] One framing of this issue is to consider it in terms of whether knowing a genetic risk or inferring likelihoods from genetic data involves a 'loss of privacy' for the data subject, how extensive that loss is, and whether all situations of privacy loss are—or should be—captured by data protection laws.[188] Thus, while the size of the genetic risk is not

[182] S. and Marper v. The United Kingdom, App. Nos. 30,562/04 and 30,566/04, ¶¶71–73 (Dec. 4, 2008), https://hudoc.echr.coe.int/fre?i=001-90,051 [https://perma.cc/C3TE-F9X5].

[183] *See* Quinn & Malgieri, *supra* note 133, at 1612.

[184] To be noted, many genetic tests do not provide a clear numerical value but instead only characterize a risk as being either reduced, average, or elevated. Typically having a risk-increasing genotype is talked of as having a genetic predisposition or susceptibility for a condition. The predictive value of different genetic loci varies, and some associations are more straightforward than others, so it is appropriate to consider how different levels of probabilities should be treated from a privacy perspective. See, eg Sakari Jukarainen et al., *Genetic Risk Factors Have a Substantial Impact on Healthy Life Years*, 28 Nat. Med. 1893 (2022) for an example of how various genetic risk factors are discussed and computed together.

[185] *See* Douglas H. Ginsburg, *Genetics and Privacy*, 4 Tex. Rev. L. & Pol. 17, 23 (1999) (noting that mere probabilities are not as sensitive as certain information, and genetic data usually only provides probabilities).

[186] As noted above, discrimination and stigma can also result from erroneous assumptions—the making of which does not necessarily require a privacy violation. *See* Citron & Solove, *supra* note 159, at 839; Karlsson Linnér & Koellinger, *supra* note 52, at 13.

[187] This relates to the WP29's notion that also knowing someone is 'healthy' is sensitive health data. *Health Data in Apps and Devices*, *supra* note 169, at 2–3.

[188] See generally Jeffrey M. Skopek, *Untangling Privacy: Losses Versus Violations*, 105 Iowa L. Rev. 2169 (2020), who differentiates between privacy losses—where the outcome is that some private information is accessed or disseminated, regardless of how that takes place—and privacy violations, where a privacy right is breached due to the way private information is accessed—independent of what kind of privacy loss occurs, if any. Talking in these terms, interpretations of the GDPR are increasingly seeking to cover all situations of privacy loss, also by inferences that do not involve privacy violations. However, in the era of big data such level of protection is difficult to accomplish by focusing on data types.

decisive for how sensitive the data are deemed, it is also not entirely meaningless and could be observed in some situations.

The sensitivity of the underlying genetic data can also be questioned by noting that one might make similar predictions based on demographic factors that are not sensitive under any classification. At the same time, it must be recognized that genetic information does include an additional level of intimacy compared with demographic data, since it implies some form of causation. For example, demographic data might reveal that males of a certain age in a certain area are generally susceptible to type II diabetes, but this type of statistics are generally thought of as mere correlations that do not carry private information about an individual's risk. However, genetic data implies a personal characteristic that provides a causal link to the disease, even if the link is weak or subject to other contributing factors. Moreover, this link can potentially expand to family members, which is sometimes emphasized as the factor that makes genetic data unique.[189]

As a further complexity, genetic tests are increasingly relying on polygenic risk assessment rather than single variants.[190] This means that the risk for developing a disease is based on, say, 3 or 20 different genomic loci that are computed to give an overall prediction of risk.[191] This can be viewed from two perspectives: on one hand, predictions based on multiple variants are going to be more accurate and thus their informational value for the individual will be more relevant and potentially more sensitive. On the other hand, this highlights the fact that the underlying genetic variants do not provide meaningful (or sensitive) information on their own. Instead, it is the predictions and likelihoods regarding health outcomes that have the potential to be sensitive. Thus, sequence and genotype information need not necessarily be protected in their own right but rather protection of the personal information derived from it would be enough.

Ultimately, these issues boil down to the role assigned to inferences in relation to health information and genetic data. The question of inferences has received much attention in the EU lately and these developments provide representative examples. The USA has traditionally been more focused on identified persons and their personal information than the potential for identification or inferences, although this is slowly changing with state-level privacy bills.[192] The California Consumer Privacy Act (CCPA), for example, protects inferences drawn from data as personal information.[193] The same is true for the Colorado Privacy Act, which also explicitly regulates sensitive data inferences.[194]

[189] *See,* eg Marisa A. Leib-Neri & Anya E. R. Prince, *Privacy and the Genetic Community*, 22 Am. J. Bioethics 70 (2022) (highlighting the familial and communal aspects of genetic data).

[190] *The Future of Genetic Testing: How a Love of DNA Led to More Comprehensive Tests*, Myriad Genet. (Apr. 23, 2022), https://myriad.com/myriad-genetics-blog/future-of-genetic-testing/ [https://perma.cc/4BQ9-E89R].

[191] *Id.*

[192] Solove, *supra* note 131, at 7.

[193] *See* Jordan M. Blanke, *The CCPA, 'Inferences Drawn,' and Federal Preemption*, 29 Rich. J. L. & Tech. 53, 73–77 (2022) (discussing the California Consumer Privacy Act's protection of inferences and how the respective provision is interpreted expansively, potentially even more so than the GDPR).

[194] *See* Colo. Priv. Act, 4 C.C.R. §904–3 (2023), https://coag.gov/app/uploads/2023/03/FINAL-CLEAN-2023.03.15-Official-CPA-Rules.pdf [https://perma.cc/8PGW-XTXA].

I noted above that the EU's interpretations have not been entirely consistent in how the scope of health-related information and inferences are treated. In light of the CJEU's recent decision in *OT v. Vyriausioji tarnybinės jaikos komisija*, it seems that the mere possibility of health-related inferences makes the underlying data health data.[195] The CJEU was asked to determine whether data capable of revealing a detail that falls within the sensitive special data categories 'by means of an intellectual operation involving comparison or deduction' should also be considered sensitive data.[196] The CJEU held that, pursuant to the language and purposes of the GDPR, any data 'liable indirectly to reveal sensitive information concerning a natural person' was to be treated as sensitive (special category) data.[197]

Within the GDPR framework, this may seem like the only possible interpretation in view of the text and existing interpretations. Protecting inferences has been deemed necessary for the coherence of data protection laws, because otherwise they would be easy to circumvent.[198] Yet, the line of deduction has to stop somewhere. This limitation should be explicitly recognized, which the CJEU has not done. The authoritative interpretations of the GDPR currently hold that sensitivity should be determined objectively, based on the technical possibilities of combining certain data to arrive at a sensitive piece of information.[199] The practical problem is that technology allows inference of sensitive data from almost anything.[200] Quinn and Malgieri have pointed out that the GDPR mainly approaches sensitive data from a 'contextual' or 'objective' perspective, not giving room for a purpose-dependent interpretation.[201] A purpose-dependent interpretation would give more leeway to controllers, but it would also avoid at least some of the problems with the infinite reach of the sensitive label under the inference-focused interpretation.[202]

To take the argument even further, Solove has argued that a separate group of sensitive data does not make sense at all, because so much sensitive data are derivable from nonsensitive data by algorithms and big data tools—and consequently the category of sensitive data swallows almost everything.[203] Privacy laws that award special protections based on inferences that can be made from the data do not work in 'our age of inference,' where algorithms can easily extract 'sensitive' data from harmless

[195] Case C-184/20, OT v. Vyriausioji tarnybinės jaikos komisija, ECLI:EU:C:2022:601 (Aug. 1, 2022).

[196] *Id.* at ¶120.

[197] *Id.* at ¶127.

[198] Solove, *supra* note 131, at 21.

[199] *See* Quinn & Malgieri, *supra* note 133, at 1594 (noting that in the GDPR only biometric data is defined through the purpose to which it is collected for).

[200] *See generally* Joanne Hinds & Adam N. Joinson, *What Demographic Attributes Do Our Digital Footprint Reveal? A Systematic Review*, 13 PLOS ONE e0207112 (2018), https://doi.org/10.1371/journal.pone.0207112 (discussing inferences that can be made from digital footprints); Tal Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 SETON HALL L. REV. 995, 1013 (2017) (describing how big data tends to make special data categories absorb more and more things); Solove, *supra* note 131, at 22 (giving numerous examples of 'harmless' data that can be used to infer specific sensitive data).

[201] *See,* eg Quinn & Malgieri, *supra* note 133, at 1590–94. The authors slightly misleadingly use the term context-based as an opposite to purpose-based, but the basic distinction is whether the focus is on the type or origin of the data itself or on the use or context where it appears.

[202] *See id.* at 1594–95 (discussing the difficulties of proving a data controller's intent and the potentially changing purposes of processing); Solove, *supra* note 131, at 32–33, 35.

[203] Solove, *supra* note 131, at 18–19.

everyday data.[204] A solution would be the refocusing of privacy laws on the 'use of personal data and [making them] proportionate to the harm and risk involved with those uses.'[205] This view is based on the notion that 'sensitive data is not more harmful than non-sensitive data,' considering that nonsensitive data can largely be used for the same purposes as sensitive data.[206]

Health data, particularly, can be inferred from almost anything, because almost all of the everyday things people 'do, buy and eat' reflect information about their health status and affect their health.[207] Thus, arguably, the GDPR would require almost *all* data to be treated as sensitive personal data—but at the same time, this has evidently not been the purpose of the regulation.[208] Rather, it clearly assumes that the identification of sensitive data and its differentiation from everything else would be easy.[209] In view of the above-described realities of genetic data and big data more generally, it would be more fruitful to regulate genetic data in a context-dependent manner, focusing on the way genetic data are used or reasonably foreseeably intended to be used.[210] Since it is possible to infer all kinds of information from genetic data and genetic data from all kinds of information—especially if likelihoods are included—looking at the data category alone simply does not make sense. The content and the use of the data should be accounted for in what is deemed allowed.

While genetic data are rarely more sensitive than other types of data, it is important that genetic data are not used to harm anybody.[211] Where genetic data are considered sensitive, the sensitivity appears to be tied to certain ways of using, sharing, and making inferences from the data. Thus, from a sensitivity perspective, there is no special need to protect against the collection and storage of genetic data as such. Rather, efforts should be focused on making sure that genetic data are not used in ways that are inconsistent with the individual's interests—same as with any other personal information.[212] Regulation of use based on the risks and harms sounds more complicated that declaring a whole category of information 'sensitive,' but the sensitivity tag comes with its own line-drawing issues that are not insignificant, as has been discussed above.[213]

In conclusion, a nuanced view of the content of genetic data shows that not all genetic data need be treated the same. The next section further develops the

[204] *Id.* at 19, 35–41 (discussing how, eg metadata, addresses, personality, and photos can be sensitive due to the ways they can be used).

[205] *Id.* at 5.

[206] *Id.* at 5, 29–30.

[207] *Id.* at 24.

[208] *Id.* at 27–28.

[209] *Id.*; Quinn & Malgieri, *supra* note 133, at 1611.

[210] *See generally* Quinn & Malgieri, *supra* note 133 (arguing that the concept of sensitivity loses its meaning in a big data environment and it would be better to award special protection for certain purposes rather than data types).

[211] This is protected against under GINA and similar nondiscrimination laws, albeit imperfectly. *See generally* Mark A. Rothstein, *Time to End the Use of Genetic Test Results in Life Insurance Underwriting*, 46 J. L. Med. & Ethics 794 (2018) (describing how genetic information can be used when granting life insurances without much interference from privacy laws).

[212] *See* Clayton, *supra* note 6, at 36 (calling for regulation of genetic information to be refocused from access questions to questions of acceptable use). *See also* Quinn & Malgieri, supra note 133, at 1609 (noting that since the lawful processing of sensitive data under the GDPR is already largely based on the type and purpose of use, it would be relatively simple to adopt an entirely use-based approach).

[213] Solove, *supra* note 131, at 5.

differentiation between types of genetic data and builds a comprehensive view of the aspects that should be considered.

V. A COMPREHENSIVE VIEW OF GENETIC DATA

V.A. The Various Dimensions of Genetic Data

The above analysis has shown that both identifiability and sensitivity of genetic data are complex issues. Types of genetic data differ with respect to how identifiable and sensitive they are. I suggest that genetic data should be regarded on multiple different but interconnected continuums. These include the amount of data, the identifiability or uniqueness of the data, the informational content of the data, and the type of use the data are put to. Genetic data that differs in terms of these aspects could be treated differently.

The first aspect is what *amount* of genetic data is at hand. It could be an individual locus or a short stretch of DNA. Typically, these would be sequenced with the purpose of diagnosing or predicting a specific disease, or in research to verify a causal link between a locus and a trait. The genetic data themselves would usually not be identifiable, but the content can be sensitive for the individual in the case of a pathologic sequence.[214] On the other extreme, the genetic data could also encompass the whole genome of a person. Now the data would be highly identifying (unless specifically processed to reduce that), and the individual would usually have a high interest in knowing what use the data are put to. However, as 99.9 per cent of the data would be identical to any other human,[215] not all of the data would be sensitive—even the 0.1 per cent of variable regions would include sections that carry either neutral or meaningless information. Yet, it must be assumed that the data would also encompass content deemed sensitive. The data could also be anything between a single locus and the whole genome—for example, a portion of a chromosome or a set of SNPs from different parts of the genome. In addition, it may in some cases be relevant to know whether the data in question is from one person or whether the dataset consists of multiple parallel sets of genetic data from different individuals. This will at least be significant for the use the data are meant for.

Looking at the amount of genetic data highlights that identifiability and sensitivity of the data are separate questions and it is not always straightforward to determine when genetic data are either of those. A single locus may be sensitive but not identifiable (eg a genotype causing a disease), whereas a larger piece may be identifying but not sensitive (eg DNA fingerprint consisting of meaningless noncoding variation). Thus, neither identifiability nor sensitivity is an inherent property of *all* genetic data and the interest of an individual to different types of genetic data may vary. While the amount of genetic data is rarely decisive for how it should be treated from a data protection perspective, it is a useful preliminary question for recognizing what kind of interests could even potentially be involved.

The *uniqueness* dimension of genetic data was extensively discussed in the context of identifiability: most of the bases of a genome are not unique to one person, but

[214] As an example, knowing that Mary has a BRCA1 mutation that makes her susceptible to breast cancer is normally considered sensitive. This is the kind of genetic data primarily targeted by regulations.

[215] *Genetics vs. Genomics*, *supra* note 59.

the more genetic data are collected, the more likely it is that the data will contain a unique combination of genotypes. Furthermore, this dimension encompasses the data that is provided or collected together with the genetic data. For example, a commonly occurring but variable genetic sequence might be unique in a dataset that combines it with a set of identifiers, even if those identifiers do not yet make the data identifying. Thus, this identifiability factor directs attention to the nature of the data rather than its origin—and looking at the uniqueness of a line of data can often provide a useful starting point for whether its identifiability should even be examined in more detail.

The *informational content* of genetic data is the dimension that arises from the above discussion of sensitivity. On the information continuum, on one end we have genotypes with clear causal links to specific phenotypes, for example a disease or a physical characteristic. On the other end, we have sequences that are nonvariable or do not perform any known functions and that might theoretically be removed from the genome or replaced with any other sequence. In the middle, there are different likelihoods and susceptibilities. Thus, a genotype could have an unambiguous effect on the individual, have some predictive value, or have no relevance at all. The informational content is also subject to change as science and our understanding progress.

Data that has some informational value may or may not be either identifying or sensitive—it depends on the context. On the purely informational level, 'mutations in gene X cause disease Y' is not personal data, because there is no link to an identifiable natural person. The same applies to knowing that 'someone' has a disease genotype. Normally, it is only the knowledge of a particular person that makes the data potentially sensitive.[216] The sensitivity of this kind of personal data depends on the informational content: knowledge of a disease or other health status would usually be considered sensitive also without the genetic component, so it is logical that the genetic data carrying the same information is considered sensitive. However, as discussed above, not all genetic data are health-related. Knowing a person has blue eyes is not sensitive, so why should it be sensitive that they have a corresponding genotype? To be noted, the GDPR and HIPAA can be interpreted to not cover this type of genetic data, as it is not related to health or physiology—however, it might become covered as regular personal data under the GDPR or under state laws.

Above, I have noted several questions that arise regarding how likelihoods should be treated. One framing of the informational content of likelihoods is to differentiate between input and output data.[217] Input data would be the original raw data obtained from the data subject, ie any genetic sequence (usually extracted from a biologic sample) and accompanying demographic and phenotypic information. Output data relates to the inferences made from the input data, including (but not necessarily limited to) the results of a research project or the finding of a risk factor based on published research. If the research establishes a link between a genotype and a phenotype (a health

[216] This issue gets slightly more complicated when we recognize that genetic information can also be inferred from nongenetic sources, for example from facial characteristics, like in the example of Down's syndrome. In that case, mere knowledge of a correlation of a readily observable trait with (sensitive) genetic information could enable identification of persons to whom that sensitive data relates. Restricting dissemination of information regarding this kind of correlations would be problematic in terms of freedom of speech and freedom of science, but inappropriate uses of these correlations (eg for discrimination) should still be prohibited.

[217] Wachter & Mittelstadt, *supra* note 176, at 571–72.

condition, for example), it is questionable whether this research result is information about any individual sample although it is sometimes framed as such. As a starting point such results only establish correlations and likelihoods on the population level, not diagnoses or predictions regarding any individual.[218] This notion links back to the question of how sensitive we should deem this kind of correlations to be, if their predictive value in an individual's life is very low despite a scientifically interesting and statistically significant link.

Finally, genetic data should be considered on the level of *use*. Traditionally, it is assumed that genetic data will be used either to identify an individual or an individual's relatives (eg in forensics) or to obtain information about an individual's health (eg genetic tests in healthcare). These are not the only uses of genetic data.[219] A basic distinction should be made regarding individual and statistical uses of data.[220]

By individual use, I refer to uses that aim to either identify a particular individual or to obtain specific information about an individual. The former includes uses in forensics, paternity tests, and other fingerprinting-type applications that ascertain the identity of an individual or their relation to another individual. The latter consists of use in individual healthcare, such as diagnostic tests to determine presence or absence of a disease or disease susceptibility variant and 'for fun' genetic tests that do the same, often also with respect to non-disease variants, like appearance and origin/ethnicity markers. Individual use could also include situations where genetic data are used to make decisions about an individual. Statistical use, in turn, refers to data uses that are not about the individual but merely use data from the individual to generate population level information and scientific principles or estimates. Normally this use would take place in the context of scientific research, but it could potentially also happen for the purposes of product development, for example.

To regulate different uses, the protected interests and risks in each use must be examined.[221] If identification is the privacy-invasive act that needs protection against, then it might make sense to have very rigorous rules regarding when it is permissible to identify a person based on their genetic (or other) data. In this scenario, however, the kind of statistical use described above would not be a problem despite it involving processing of the individual's potentially identifying genetic information.[222] Furthermore, if the concern is the dissemination of undesirable knowledge to the individual, there are other means to prevent and control that.[223] The individual could also be concerned

[218] *See* Quinn & Quinn, *supra* note 12, at 1010 (noting that 'data generated within research projects may be aggregate in nature and thus may not relate to a specific individual as such. Conclusions are more likely to be in the form of (potentially very) low level correlations between various DNA sequence variations that may have limited relevance to specific individuals and may not even be considered as personal data.').

[219] *See* Clayton, *supra* note 6, at 20–26 (describing several additional ways genetic data is or could be used).

[220] This distinction is also made by, eg Bonomi, *supra* note 51, at 647.

[221] *Cf.* Quinn and Malgieri frame this issue as sensitivity of particular uses through the involvement of sensitive data in that use and propose that it be assessed by asking both (i) whether the controller intends to use/infer sensitive data, and (ii) whether it is objectively plausible that the use will involve/infer sensitive data. Quinn & Malgieri, *supra* note 133, at 1609–10.

[222] The HIPAA, for example, allows relatively broad sharing of data for research.

[223] *See generally,* eg Laura Flatau et al., *Genomic information and a Person's Right Not to Know: A Closer Look at Variations in Hypothetical Informational Preferences in a German Sample*, 13 PLOS ONE e0198249 (2018), https://doi.org/10.1371/journal.pone.0198249 (showing that individuals have varying concerns and preferences regarding genetic knowledge).

about specific people or institutions (eg employers, insurers) knowing their genetic information, and this risk could be (and at least in part also has been) tackled with clear restrictions and consent requirements regarding the sharing of genetic data and its use if shared. Thus, the risks and sensitivities often viewed as inherent to genetic data can be addressed by considering and restricting the uses that are allowed.

In conclusion, there are several layers and dimensions in what science and legislation call genetic data or genetic information. We have seen that identifiability and sensitivity can vary depending on the amount and type of genetic data involved, and the privacy interests associated with each create different conflicts depending on what kind of use the data are intended for. Noting all these various dimensions, it seems inappropriate to group all genetic data and its uses together and apply only one set of principles. The next section contemplates in more detail how these dimensions should be taken into account.

V.B. A Framework for Evaluating Genetic Data

This final section pulls together the strings of the above perspectives to further an accurate and purposeful understanding of genetic data in legal contexts. I first develop the dimensions stemming from the analysis of identifiability and sensitivity to build a framework under which content and uses of genetic data could be assessed. I then reflect on the implications of these findings for current regulations and debates concerning genetic data.

First, privacy considerations are not necessarily triggered on all levels of processing genetic data. It is largely the sensitive information embedded in genetic data that clashes with different use cases. There are also uses that implicate privacy interests for nonsensitive data that has any informational value as well as all identifying data. The data or content that only operates on the levels of individual genome (as in, the biological material in one's cells) or the extracted genetic data that is non-identifying generally do not involve privacy-intense considerations on their own.[224]

At this point, I make brief note specifically about the status of biological samples in this assessment. Biological samples typically include the entire genome of an individual and consequently the potential for all the inferences that could be made based on the sequenced genome. Sometimes, this fact is presented as a basis for raising biological materials to an 'extra-sensitive' category, but the reality of how abundant biological material is points to the opposite: every time people touch or eat anything or visit somewhere, they leave behind biological material. Generally, this material is not considered a biological sample unless it is specifically collected and used for something. This kind of use, however, tends to be heavily regulated with respect to procedural requirements for collecting the sample (eg consent for participation in research) or using it for an acceptable purpose (eg analysis of crime scene samples for law enforcement purposes). Thus, a comparison to biological samples supports focusing on types of use rather than mere existence of data.

Combining the above-outlined dimensions of genetic data with the basic types of uses (individual and statistical), genetic data can be grouped into different types in

[224] The risks and debates relevant for these levels are more about informed consent, research ethics and ownership and other rights to data—not so much privacy.

Table 1. Potential privacy interests for different types of genetic data.

	Identifying data		Non-identifying data
	Informational data	Quasi—/non-informational data	
Individual use	Strict safeguards necessary	Strict to moderate safeguards necessary	Not possible
Statistical use	Strict to moderate safeguards necessary	Moderate to low privacy interest	No privacy interest

terms of the strength of the privacy interest involved. Such directional classification is shown in Table 1. It first differentiates between whether the data are identifying in the GDPR sense, ie considered together with all the data that accompanies it or that is reasonably connectable to it, or not. If the data are not identifying, then there is no privacy interest and individual use is not even possible (right column).

With respect to data capable of identifying the individual, a distinction can be made between data with (high) informational value and data that carries little to no tangible information. Under this model, non-informational data could be processed for statistical (research) purposes with few problems. Such use might include processing of collections of genetic markers with no real relevance for the individual. Some level of privacy interest should still be recognized, mostly in the form of consent, autonomy, and data security requirements. These lighter procedural safeguards would arguably be enough for the lower end of the privacy interest spectrum.

Despite the low informational value, this kind of data requires appropriate safeguards when used to make predictions or decisions about an individual. In this context, safeguards refer to mandatory data security measures and data subject rights.[225] An example of the 'quasi-informational' data also falling into this box could be a low predictive value genetic test that relates to a cellular characteristic. Individuals presumably wish to have some control over this type of information, even if the content is such that it bears no direct relevance for their life. For this kind of data, moderate safeguards such as a requirement to balance the benefits of the use with individual preferences[226] might suffice, since the data or the use are not directly implicating.

This category would also encompass identification-based applications of genetic data, where high protections against misuse are generally required and which are generally perceived as invasive even if the content of the genetic data involved is

[225] Different jurisdictions have different approaches to the exact rights of data subjects and the obligations and liabilities of the data controller and processor. Some of them are in line with the classifications suggested here while others are not. Safeguards against inappropriate processing have been discussed, eg by Quinn & Malgieri, *supra* note 133, at 1600–04 (discussing some of GDPR's processing safeguards for sensitive data).

[226] This mechanism might look similar to the legitimate interest basis under GDPR art. 6. Currently, said basis for processing is not as such available for genetic data as defined in the GDPR.

nonexistent.[227] In that case, again, it is not so much the inherent nature of genetic data that makes the use invasive but rather the nature of the use. Strict safeguards might include regulatory purpose limitations, procedural safeguards, confidentiality, strong data subject rights, and a threat of punishments. Overall, the choice of the exact safeguards required for individual use of noninformational data would depend on the specifics of the contemplated use and data. Thus, there is room for further dividing this category into more specific situations. In this context, the main takeaway from the existence of this category is that individuals' interests in their genetic data can sometimes be sufficiently accounted for by procedural and purpose-related safeguards without a need for necessarily involving the individual as an active decision-maker.

I would reserve the highest levels of privacy protection for data with high informational value (left column). This would include data with clear relevance for a person's health, for example. I agree that it is generally appropriate to have strict safeguards in place for processing of such data—the difficult line-drawing relates to the question of how certain and direct the genetic information should be to qualify as health data of an individual, as discussed above. The data might also be other than health related, and in that case the type of use would govern. For use in the care or examination of or decision-making regarding an individual, strict safeguards are appropriate based on the implications for a person's life from the use. These safeguards should include, for example, a statutory basis for the use as well as involvement of the individual as a decision-maker. Statistical use, in turn, might occasionally be relieved from some of the requirements because the use does not similarly conflate with the private life of the person.[228] Yet, because of the data being identifying and conveying information about the person, at least a moderate level of privacy risk management should be required. This should involve, at least, confidentiality, data security, and access requirements.

Overall, under this model—as also discussed throughout this paper—it would be for the data controller to determine the identifiability and sensitivity of the genetic data in view of the contemplated use of the data and the risks related to that use. Optimally, there would be rules in place to enforce an appropriate level of safeguards based on the scientific and risk-based assessment. In a number of contexts, it may be very burdensome to conduct a detailed assessment of the properties of the data being collected or used, especially if there is a lot of them. In that case, it would be up to the data controller to balance the losses from choosing to apply the strictest data protection standards against the costs of applying a more nuanced framework. In practice, these matters regarding costs and amount of work may be significant, but this should not prevent the possibility of making individualized assessments when desired. After all, some use cases may be practically prohibited by adoption of the strictest standards—and compliance with them is not free of costs either. Where certain types of uses

[227] *See generally,* eg Erin Murphy, *Relative Doubt: Familial Searches of DNA Databases,* 109 Mich. L. Rev. 291 (2010) (arguing against familial matching in forensic DNA databases on various grounds, including privacy).

[228] This suggestion may be controversial, partly depending on the geographical location, and the respective sensitivities of the public. For example, some authors hold the view that this type of statistical use would be *more* invasive than individual use and raise more troubling bioethical questions. *See,* eg Bonython & Arnold, *supra* note 135, at 394. Unpacking these differences requires falling back to the question of what exactly is sought to be protected in each scenario.

are recognized as typical, authorities or industry organizations could issue rules or guidance regarding how these various aspects should be interpreted in those cases to lighten the burden of individual controllers.

Ultimately, it would be up to the courts to decide whether the assessments were appropriate, but this seems like an area where expert-driven guidance and interpretations could provide useful starting points for both concrete actions as well as legal standpoints. The assumption is that such interpretations would be more contextual, consequentialist, and risk-based than existing policies. The problems with relying on rights-based approaches and deontological ethics include, firstly, the increased potential for inferences discussed above. Where different data categories cannot functionally be distinguished from one another, strong individual rights will either become pretextual or infect everything beyond what anyone finds necessary.

Secondly, risk-based approaches are better suited to observing collective interests such as familial connections that are often mentioned as the unique feature of genetic data. An individual's rights and interests in community/family level genetic data are ambiguous, difficult to define, and most likely would require genetics-specific rules to make workable. A risk-oriented view, on the other hand, allows a data controller or the public to observe and manage collective risks and interests even when it would not make sense to assign this authority to any individual. Referring to the above framework, communal or familial genetic data might occur in the context of both individual and statistical use, but it makes sense to limit the individual's decision-making to use regarding them personally, while public policy should govern the collection and acceptable uses of statistical data concerning groups. Where such uses are deemed desirable by the public, it seems illogical to assign the veto to individual data subjects, if the data originating from them is not directly identifying or sensitive.

To link these views back to EU and US law, the issues with GDPR are that, on one hand, its definition of genetic data is not entirely coherent with the science. There is also some discrepancy in how the language of the definition is interpreted in practice. On the other hand, the GDPR in general is increasingly interpreted to cover all indirect inferences with limited possibilities for contextual, risk-based approaches. Referring to the framework proposed here would refocus the data protection of genetic data to questions of individual's interests and permissible uses rather than definitions.

The views presented here also have implications for the assessment of anonymization, lawful bases of use, and permissible storage of genetic data. The requirements of the GDPR can be interpreted differently depending on how the identifiability, sensitivity, and other dimensions of the genetic data are framed.[229] This is relevant because value losses for research and its applications may be significant if the regulations are interpreted too restrictively[230]—which is why overprotection of genetic data is undesirable even while concerns regarding sufficient privacy safeguards are valid and recognized. Thus, whether specific genetic data are inside or outside the scope of the GDPR is a crucial matter, and the dimensions presented here are tools for the process of determining that. While changes in legal interpretations would be helpful, most likely

[229] *See generally,* eg Quinn & Quinn, *supra* note 12 (explaining how the GDPR affects genetic research, for example through the principles of purpose limitation, data minimization, and storage limitation).

[230] *See id.* at 1007 (explaining the significance of being able to link and use genetic research data in novel and originally uncontemplated ways and the connections to good research practices).

some legislative reforms would also be required to build a fully coherent regulatory scheme.

With respect to the USA, the main issue with the current sectoral framework is that it lacks a comprehensive approach to the types and uses of genetic data, leaving loopholes that become increasingly relevant as genetic data further proliferates and potentially becomes used on many levels of society. In that scenario, a patchwork of different principles on the state levels also seems undesirable. Moreover, even regulatory schemes like HIPAA and GINA face the same issues of drawing the line to which data or information counts as genetic data or health-related data in view of the inferences available to skilled processors, and it is not clear that the balance struck by the regulations with respect to their scope and privacy protections is optimal.[231] The multidimensional view of genetic data presented here can thus be used to draft better informed legislation and guidance also in the USA, in addition to interpreting existing regulations in a new light. These discussions should involve the questions of what the law currently targets and what the law should target, how the law is interpreted in practice, what is happening in practice, and ultimately whether we are happy with that. To answer these questions in a sensible manner, an accurate understanding of the nature of genetic data is necessary.

In addition to privacy regulations, this nuanced view of genetic data is also relevant for other debates. A persistent line of discussion is the nature of an individual's right to their own genetic data as well as the rights of those who create, possess, and use such data. For example, the argument that an individual should have property rights to their own genetic data becomes complicated when one recognizes all the dimensions of genetic data—it would often be very unclear what the exact information owned by a person would be.[232] Similarly, the perspectives provided here can be used to assess whether, when, and to what extent a particular use of genetic data must observe the rights and interests of individuals whose data are used.[233] Such assessments can be made in specific use cases as well as on the theoretical and policy levels. Table 1 presents a tool to approach these questions. Here, it must be observed that the term 'genetic data' is detached from the definitions of specific regulations to cover all genetic sequences and genotype information, not merely identifying health-related data.

Furthermore, these tools help with the question of whether genetic data should be treated as exceptional or whether it might be acceptable to treat it as more everyday data.[234] It appears that definite exceptionality can only be defended for a subgroup

[231] *See generally,* eg Mark A. Rothstein, *The End of the HIPAA Privacy Rule Currents in Contemporary Bioethics*, 44 J. L. Med. & Ethics 352 (2016) (arguing that HIPAA provides very limited privacy protection); Mabel Crescioni & Tara Sklar, *The Research Exemption Carve Out: Understanding Research Participants Rights under GDPR and U.S. Data Privacy Laws*, 60 Jurimetr. 125 (2019) (discussing problems in reconciling research and data subject rights in the EU and USA regulatory frameworks); Ashley Huddleston & Ronald Hedges, *Liability for Health Care Providers under HIPAA and State Privacy Laws*, 51 Seton Hall L. Rev. 1585 (2020) (discussing problems with accountability and liability for unauthorized disclosures under HIPAA and state laws).

[232] *See,* eg Mary J. Hildebrand, Jacqueline Klosek & Walter Krzastek, *Toward a Unified Approach to Protection of Genetic Information*, 22 Biotech. L. Rep. 602, 604–06 (2003) (discussing the pros and cons of awarding property rights to genetic information while evading the question of what part of the information would actually be protected by such rights).

[233] The genomic research community has been very active in trying to find responsible ways of sharing data that also benefit and accelerate research. *See generally* Byrd, *supra* note 52.

[234] *See,* eg Bregman-Eschett, *supra* note 156, at 8 (summarizing some of the main points in the exceptionality debate).

of genetic data, albeit a significant one: whole genomes. Other forms of genetic data are increasingly difficult to distinguish from other types of data available and used in contemporary society and the inferences that can be drawn from them. For example, location data can reveal probabilities regarding health information (eg visits to a cancer clinic or a gym), information regarding racial or ethnic origin (eg area of living and movement), or criminal behavior (eg speeding or trespassing).[235] Similarly, photographs can be used to infer much of the same information and more.[236] Thus, the question is whether these potential inferences should be treated as more severe than otherwise simply because they are based on genetic data. I do not think that is justified.

The above framework presents a process for assessing privacy interests encompassed in genetic data. The framework has been built as specific to genetic data, because traditionally genetic data has been viewed to involve special considerations and this Article was partly built around debunking those common assumptions, while also recognizing the importance of some of the issues. This Article also argues that genetic data are not as special as some instances seem to think, which can be exemplified by applying the above framework. This demystified status should also manifest in the outcome of the assessments made using the framework, where it might otherwise not be intuitive. Furthermore, even if genetic data are not viewed as inherently special, practical considerations dictate that it should still sometimes be discussed separately to account for its properties and the ways it can be used—this kind of specialness also applies to other complex data types. Thus, there is no conflict in recognizing both that genetic data are not inherently exceptional *and* accounting for its specific properties in individual use cases. In case another data type suffers from similar misconceptions regarding what kind of privacy interests are involved, the presented framework could potentially also be applied to that data.

While the analysis and the framework provided here point to a need to rethink the way we regulate genetic data, it is not possible to build a fully cohesive privacy rule within the scope of this Article. However, the tools and views presented here should be used to elaborate on what such rules could be in subsequent scholarship, without taking for granted the exceptional status of genetic data in terms of its identifiability and sensitivity.

VI. CONCLUSION

This paper has illuminated the multidimensional nature of genetic data that tends to be overlooked in regulation and legal literature. Genetic data are a complex topic to regulate and having a better sense of what is being discussed and targeted in practice should result in increased efficiency both in terms of privacy protection and allowing productive use. Specific uses (and misuses) of genetic data remain subject to their own

[235] This is recognized in the European Data Protection Board's (EDPB) Guidelines but the extent of the inferences that could be made from location and similar data seems to be inadequately addressed, since the focus is on practical data security, access, and purpose of processing questions. In this sense, the EDPB seems to emphasize the practical perspective of allowing reasonable uses rather than a strictly protective view that is more commonly seen with respect to health-related data. See *Guidelines 01/2020 on Processing Personal Data in the Context of Connected Vehicles and Mobility Related Applications*, at ¶¶56, 67–68, European Data Protection Board (Mar. 9, 2021), https://edpb.europa.eu/system/files/2021&#x2013;03/edpb_guidelines_202001_connected_vehicles_v2.0_adopted_en.pdf [https://perma.cc/D6KY-5YLS].

[236] *See* Solove, *supra* note 131, at 39–41.

debates,[237] but this paper increases the nuance in those conversations by pointing out aspects of genetic data that the debaters should be aware and concerned about.

The analysis above has shown, firstly, that identifiability needs to be viewed from a more disaggregated and dynamic perspective. Not all genetic data are identifiable: many parts of genetic sequences are universal, common in specific populations, or otherwise incapable of being connected to an individual on their own. Even genomic datasets could potentially be processed to reduce identifiability to a legally negligible level. Furthermore, to the extent genetic data are identifiable, they are often not more so than other types of data commonly used and available in our big data society. Thus, it would make sense to treat genetic data with the same level of protection as that awarded to other types of data with similar content or use rather than applying rules specific for genetic data.

Secondly, genetic data can be evaluated from the perspective of sensitivity. Not all genetic data are sensitive: much of them concern common everyday traits, molecular markers with little personal significance and no health implications, features with unknown functions, and some genetic data do not even have any meaningful informational content. There is no compelling need to treat this type of data with as much oversight as truly sensitive data. Furthermore, to the extent genetic data are sensitive—either at present, potentially in the future or through inferences—they are often not more sensitive than other types of data. Big data and modern algorithms allow extraction and inference of data with similar levels of sensitivity from all types of mundane activities, including photographs, location data, and internet use data. It is inconsistent to protect one with extreme caution but allow another to be used freely or with little restrictions—even if we accept that genetic data still detain some level of exceptionality in how intimately and permanently they are tied to a person.

As a solution, I propose a more nuanced understanding of the types of genetic data and detaching from the notions of extreme identifiability and sensitivity. Different dimensions can be found in terms of the amount, uniqueness, and informational content of genetic data. Different types of genetic data might be appropriate to use for different purposes and in different ways while maintaining an appropriate level of privacy and individual autonomy.

Holding these nuances in mind, I propose the following actions and policy changes in how we approach genetic data. First, the definitions of genetic data and health data

[237] *See generally,* eg Murphy, *supra* note 227 (discussing use of forensic DNA databases for familial matching); Jacqueline Moran, *Privacy Perspectives on Direct-to-Consumer Genetic Testing in the Era of Big Data: Role of Blockchain Technology in Genomics,* 22 TUL. J. TECH. & INTELL. PROP. 185 (2020) (discussing the privacy problems with direct-to-consumer genetic testing); Ron J. Whitener, *Research in Native American Communities in the Genetics Age: Can the Federal Data Sharing Statute of General Applicability and Tribal Control of Research Be Reconciled,* 15 J. TECH. L. & POL'Y 217 (2010) (discussing the access and rights of indigenous populations to genetic research and its results); Nancy J. King, Sukanya Pillay & Gail A. Lasprogata, *Workplace Privacy and Discrimination Issues Related to Genetic Data: A Comparative Law Study of the European Union and the United States,* 43 AM. BUS. L. J. 79 (2006) (discussing the prevention of genetic discrimination and the underlying rationales and principles); Bonython & Arnold, *supra* note 135, at 381–83 (discussing ownership and patenting of genetic information and inventions); Colleen Conboy, *Consent and Privacy in the Era of Precision Medicine and Biobanking Genomic Data,* 46 AM. J. LAW. MED. 167 (2020) (discussing privacy issues with biobanking); Cindy Cornelis, *Medical Confidentiality and Disclosing Genetic Information to Family Members,* 39 MED. & L. 419 (2020) (discussing rights of relatives to access genetic information).

in privacy laws should be reassessed and reflected against science and the potential for inferences. Information with a genetic link or origin should be treated consistently with how the same information is treated when it arises in nongenetic contexts.

Second, it should be more explicitly recognized that not all genetic data are personal, and data controllers should be encouraged to make individual assessments of the personal nature and the related risks of the genetic data they process. For example, where limited amounts of genetic data have been collected for research, the data could be repurposed and shared between various projects with fewer administrative hurdles if it can be deemed non-identifying—either inherently or by technical processing. Furthermore, even larger amounts of genetic data could be used for statistical purposes without data subject rights being implicated, if the informational content of the data is not meaningful for the individual.

Third, I propose promoting improvements in data security and informing data subjects instead of broadening legal definitions and interpretations. The concept of genetic data (and health data) becomes obsolete if there are no real limitations to its reach. Emphasis on practical measures would be more effective in preventing harm and ensuring data subject autonomy and sense of security.

Finally, policy discussions should focus on the use of genetic data instead of merely its nature and inherent properties. Policy makers should aim to determine what uses of what types of genetic data are allowed instead of trying to maximize protection of all data. As novel applications of genetic data are developed, their acceptability could be determined based on the risks arising from identifiability and sensitivity of the data rather than focusing on the fact that genetic data are used. A one-size-fits-all approach in regulation of genetic data—even health-related genetic data—appears outdated in today's environment.

FUNDING

This research has been supported by the Fulbright Finland Foundation (ASLA-Fulbright Graduate Grant 2022-2023).