Science & Society

EMBO reports

Re-identifiability of genomic data and the GDPR

Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation

Mahsa Shabani[1,†] & Luca Marelli[2,3,†]

Human genomic data have become an important and rich resource for biomedical and clinical research. At the same time, concerns about the identifiability of genomic data have been central to discussions regarding adequate protection of personal data and privacy. Addressing such concerns is paramount for research and clinical data repositories, as well as for ensuring interoperability of standards across jurisdictions. However, in spite of increased scholarly and policy scrutiny during the past decade, questions remain about when and if genomic data can be truly irreversibly de-identified.

"... the new law in the EU mandates that data that has been merely pseudonymized is regarded as personal data that falls under its scope, while anonymous data would not be subject to the regulation."

These discussions have acquired renewed salience in Europe after the EU Regulation 2016/679, also known as the General Data Protection Regulation or GDPR (https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN), came into effect. At its core, the GDPR mandates a decentralized, context-specific and risk-based approach to data protection with emphasis on the accountability of data controllers (Arts. 5(2) and 24) [1]. Additionally, under a so-called "research exemption", the GDPR allows for some flexibility for the processing of personal data for scientific research (Art. 9(2)(j)), and it relaxes the stringent requirements for specific consent (Recital 33) and data storage (Art. 5 (1)(e)). Moreover, it allows EU Member States to introduce further provisions for the processing of genetic, biometric, and health-related data (Art. 9(4)).

Processing of genetic data under the GDPR

The GDPR lists genetic data as "special categories of personal data" or sensitive data (Art. 9), which makes their processing for research purposes (Art. 9(2)(j)) subject to the adoption of adequate organizational and technical safeguards, such as pseudonymization (Art. 89(1)) [1,2]. Pseudonymization is defined in Art. 4(5) as the process through which data "can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person." The GDPR explicitly defines data that have undergone pseudonymization as personal data, thus falling within the scope of the regulation.

Regarding anonymous data, the regulation states that the principles of data protection "should not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable" (Recital 26). In determining what should be considered as non-identifiable data, the GDPR mandates that "account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly." The regulation further states that "to ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of 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."

"The crucial question therefore is whether genetic data can always be considered as identifying, or which conditions are required to achieve irreversible de-identification..."

In summary, the new law in the EU mandates that data that have been merely pseudonymized are regarded as personal data that fall under its scope, while anonymous data would not be subject to the regulation. It further stipulates that, in order to be processed anonymously, data must be stripped of any information that could

1 Department of Public Health and Primary Care, Center for Biomedical Ethics and Law, University of Leuven, Leuven, Belgium. E-mail: mahsa.shabani@kuleuven.be
2 Life Sciences & Society Lab, Centre for Sociological Research, University of Leuven, Leuven, Belgium. E-mail: luca.marelli@kuleuven.be
3 Visiting Scientist, Department of Experimental Oncology, IEO, European Institute of Oncology IRCCS, Milan, Italy
†These authors contributed equally to this work
DOI 10.15252/embr.201948316 | EMBO Reports (2019) 20: e48316 | Published online 24 May 2019

lead to identification, either directly or indirectly—for instance, by combining data with other available information; that this process of de-identification must be irreversible; and that de-identification assessment should focus on outcomes, rather than means or procedures. Thus, determining the status of genomic data, namely whether and at which conditions it should be considered as identifiable or not, has significant implications for clinicians and researchers who use and share such data. In particular, processing identifiable data requires both a suitable legal basis, such as consent, and adequate organizational and technical safeguards. On the contrary, the processing of irreversibly de-identified data is not subject to the GDPR, although other provisions may still apply [3].

> *"Earlier studies demonstrated that fewer than 100 single nucleotide polymorphisms (SNP) are sufficient to distinguish an individual's DNA record"*

The crucial question therefore is whether genetic data can always be considered as identifying, or which conditions are required to achieve irreversible de-identification of genetic data? And can we ever confidently consider the latter as non-personal data for the purpose of GDPR? Here, we address these questions by providing an overview of the key factors that impinge on the identifiability of genomic data, namely precedents of re-identification, the context of the processing, and data governance models. Additionally, we discuss the implications of this overview in light of the regulatory framework established by the GDPR.

Identifiability: concept and scope

In order to determine whether genetic data are identifiable or not, it is crucial to define identifiability and, accordingly, re-identification. (Re-)identification is often understood as either identity disclosure or attribute disclosure, that is, as either revealing the identity of a person, thus breaching his or her anonymity, or disclosing personal information, such as susceptibility to a disease, which is a breach of privacy [4–7]. This difference can be highly relevant: even if patients or participants consent to share their identifiable genomes, they may not consent to the detailed (and unanticipated) characterization that may occur through disclosure of their genomic data [7].

The concept of identifiability can also be defined in relation to individuation and distinguishability. While the first refers to the connection between a record and an individual, such as a unique code assigned to a tissue sample, the latter is related to the ability to distinguish records from one another [4]. In the GDPR, distinguishability is referred to as "singling out" (Recital 26). However, the latter is defined in the regulation as a method for re-identification, arguably implying that singling out an individual does not equate, *per se*, to her or his re-identification [8].

Earlier studies demonstrated that fewer than 100 single nucleotide polymorphisms (SNP) are sufficient to distinguish an individual's DNA record. This has significant implications for open-access platforms that allow public queries, such as the Beacon Network (Table 1). In this regard, it is important to determine the minimum amount of data that can be processed and shared, without posing a risk of re-identification. Moreover, genetic and genomic data convey information not solely on the individual, but also on their relatives and their ethnic heritage. In turn, as a growing number of studies have shown, this raises major issues with regard to the disclosure of ethnicity, and the identity of siblings or other relatives in the context of forensic investigations. As with many data protection legislations worldwide, though, the GDPR maintains a narrow focus on the individual, thus shunning such issues and concerns.

Precedents of re-identification attacks: evidence of risk or harm?

One approach to probe the extent to which data can be considered (re-)identifiable is to study previous attempts to demonstrate that individual genomic data, even if included in datasets of aggregated information, cannot be irreversibly de-identified (Table 1). Such studies provide an evidence-based approach to what should be considered identifiable data based on materialized occurrences—rather than theoretical or merely perceived risks—of re-identification. Yet, factoring in the time, effort, and expertise needed, such attacks may still not be conclusive of the actual likelihood of re-identification. Conversely, however, the rapid progress of technologies and expertise, as well as incentives and motives, such as cybercrime, could expand the scope of re-identification from proof-of-principle studies to more widespread attempts that would threaten the privacy of concerned individuals.

> *"... factoring in the time, effort and expertise needed, such attacks may still not be conclusive of the actual likelihood of re-identification."*

These re-identification studies led to different interpretations of the likelihood of the risk and the need for revisiting current data protection and privacy policies. For example, after the re-identification demonstration by Homer and colleagues in 2008, the National Institute of Health (NIH) and the Wellcome Trust moved individual genotype data and aggregate genotype frequency data from open-access to controlled-access databases. In response, scholars engaged in discussions on the consequences of this restriction on research and the importance of achieving the right balance between data access and data protection [9].

Finally, some commentators advocate a harm-based approach, which focuses on evidences of actual harms resulting from data misuse or security breaches after re-identification—for instance, individual distress or financial damage. A 2014 "Review of Evidence Relating to Harm Resulting from Uses of Health and Biomedical Data" by the Nuffield Council on Bioethics and the Wellcome Trust provided an extensive overview of potential harms resulting from re-identification attacks. The question here is whether it would be sufficient to limit the concept of harm to "misuse of data" by third parties, or to any potential breach of privacy irrespective of its consequences.

Identifiability: context matters

The peculiar characteristics of specific genetic datasets, such as the type of data—for example, germline versus somatic tumor variants—sample size, or rareness of the genetic variant considered, represent a key

Table 1. Examples of studies showing re-identifiability and distinguishability of genomic data.

Lin Z, Owen AB, Altman RB (2004) Genomic research and human subject privacy: *Science* 305: 183	Suggested that human beings can be uniquely identified from just 30 to 80 statistically independent SNPs
Homer N, Szelinger S, Redman M, *et al* (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 4: e1000167	Demonstrated that specific individuals could be distinguished in genome-wide association study (GWAS) data through summary statistics (allele frequencies)
Cassa CA, Schmidt B, Kohane IS, *et al* (2008) My sister's keeper?: genomic research and the identifiability of siblings. *BMC Med Genomics* 1: 32	Demonstrated risk of revealing one's siblings' identity through one's SNPs
Schadt EE (2012) The changing privacy landscape in the era of big data. *Mol Syst Biol* 8: 612	Demonstrated that it is possible to derive genotypic information and identify an individual in large-scale collections of genomic profiles from publicly available RNA data
Hae KI, Gamazon ER, Nicolae DL, *et al* (2012) On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am J Hum Genet* 90: 591–598	Demonstrated that quantitative trait GWAS results can be linked directly to human research participants if a matched sample is available
Gymrek M, McGuire AL, Golan D, *et al* (2013) Identifying personal genomes by surname inference. *Science* 339: 321–324	Demonstrated that participants could be re-identified by linking STRs on the Y chromosome with data found in publicly available datasets
Schloissnig S, Arumugam M, Sunagawa S, *et al* (2013) Genomic variation landscape of the human gut microbiome. *Nature* 493: 45	Indicated that individuals might have a unique metagenomic genotype
Shringarpure SS, Bustamante CD (2015) Privacy risks from genomic data-sharing beacons. *Am J Hum Genet* 97: 631–646	The study shows that in a beacon with 1,000 individuals re-identification is possible with just 5,000 queries
Lippert C, Sabatini R, Maher MC, *et al* (2017) Identification of individuals by trait prediction using whole-genome sequencing data. *Proc Natl Acad Sci USA* 114: 10166–10171	Developed model to predict phenotypic traits (e.g., facial structure, voice, eye and skin color, height, weight, and BMI) from common genetic variation in WGS data
Erlich Y, Shor T, Pe'er I, *et al* (2018) Identity inference of genomic data using long-range familial searches. *Science* 362: 690–694	Predicted that with a database size of ~3 million US individuals of European descent (2% of the adults of this population), over 99% of the people of this ethnicity would have at least a single 3rd cousin match

factor for assessing the likelihood and severity of the consequences of re-identification.

At the same time, the context in which genetic data are being processed is an additional crucial element to consider for assessing (re-)identifiability. Whereas genetic and especially genomic data can be considered as highly identifying, no single piece of data taken in isolation represents an inherent or perfect identifier. As the report "Identification and genomic data" from the PHG Foundation notes, "what identifies someone is a combination of the uniqueness of the data and the nature of the connection between different data. Identifiability is the outcome of a network of associations: datasets that allow more connections to be drawn are more easily likely to result in identification." Accordingly, key for determining the likelihood of identification are contextual factors such as the institutional setting in which processing occurs, the stage of the processing, the availability of cross-referenceable datasets, and incentives and availability of resources for re-identification.

A first set of these contextual factors relates to the spatial and temporal configuration of the processing, namely the institutional setting and the stage at which processing occurs. The institutional setting—such as data processing in health care, research, consumer genomics, or forensics—is a key element impinging on re-identification, owing to distinct policy, ethical and regulatory frameworks that may prescribe different duties for data controllers, and different safeguards for data subjects. For instance, the processing of genetic data in biomedical research settings, which mandate ethics review of research projects and follow well-established ethics standards, is subject to additional layers of oversight that offer further guarantees against attempts at, and the risk of, re-identification. Yet, ethics review bodies may differ in their approaches to handle risks and protect research participants, thus potentially creating diverging standards. In the same vein, standards may diverge where different entities are conjointly involved in data processing activities, or in case of changes in the institutional configuration and governance arrangements, following for instance a database's change of ownership [1].

The discussion of privacy risks in different processing stages such as data collection, sequencing, storage, computation, interpretation, and analysis is another relevant point to assess consequences of re-identification. A breach of privacy in any of these stages would reveal various types of information that could differently affect research participants and related individuals. Notably, once genomic and related phenotypic data are stored in large data collections, they may pose heighten risks for the confidentiality of data subjects and relatives, owing for instance to unwanted disclosure, unauthorized access, or third-party attacks.

In addition, the risk of re-identifiability of datasets can increase significantly through cross-reference with publicly available datasets. While research databases alone may not pose privacy risks to research participants, linking them to other public databases, such as hospital records and ancestry DNA databases, greatly increases the likelihood of re-identification of both participants and relatives. Finally, underlying incentives and availability of resources have been also shown to represent important factors for the risk of re-identifiability.

Identifiability in the view of access management

Aside from approaches such as the Personal Genome Project (https://www.personalgenomes.org) or openSNP (https://opensnp.org) that seek to promote data sharing by making genetic data available in the public

domain, responses to the risk of re-identification have mostly revolved around two sets of strategies, namely the adoption of technical safeguards and the implementation of adequate governance frameworks. Hence, when discussing re-identifiability in relation to genetic databases, attention should be also paid to the associated governance models for managing data access [10].

> *"While research databases alone may not pose privacy risks to research participants, linking them to other public databases [. . .] greatly increases the likelihood of re-identification . . ."*

Access models could be considered as "organizational measures" that, along with technical measures, are mandated by the GDPR for safeguarding data as well as discharging data controllers' accountability obligations (art. 89) [2]. More to the point, institutional rules establish terms and conditions for accessing datasets, including the obligation of not attempting to re-identify data subjects. The development of the most suitable model—open access, registered access, or controlled access—with related terms and conditions could thus provide an added layer of protection and create disincentives against re-identification attempts. The enforceability and effectiveness of legally binding documents such as data access agreements is, however, a matter of ongoing discussion.

Discussion

How, against this backdrop, should we assess the (re-)identifiability of genomic data in light of the GDPR? Generally, the data protection framework established by the regulation seems to give adequate weight to the concerns raised in the scientific literature, while distinctively recognizing the importance of contextual factors for re-identification.

First, identifiability of genomic data has been shown to depend on multiple factors, such as the specific characteristics of datasets, the context in which the processing occurs, the technologies, expertise and incentives available, and the mitigation strategies adopted. In light of this, de-identification of genetic data should be regarded as a dynamic exercise: for any given dataset, it cannot be achieved once and for all; rather, attending risks should be periodically re-assessed by data controllers at every stage of the processing [3]. This aligns with the risk-based approach to data protection adopted by the GDPR through "accountability" (Art. 5(2), art. 24) and "data protection by design and by default" (Art. 25) principles, which entrust data controllers to ensure that appropriate protection measures are designed and implemented throughout all data processing activities [1].

Second, the GDPR attributes high importance to contextual factors that, as traced above, represent crucial elements for assessing the (re-)identifiability of genetic data. In particular, it does not provide procedural guidance to prescribe how the de-identification process should or could be performed. Rather, in line with its overall decentralized approach, the GDPR adopts an outcome- and context-based criterion to determine whether personal data should be considered as irreversibly de-identified, which requires to factor in "all," "likely," and "reasonable" means that could be deployed to reverse de-identification (Recital 26).

Here, the GDPR differs from other data protection legislations, most notably the US Health Insurance Portability and Accountability Act (HIPAA). Within its Privacy Rule, the Safe Harbor standard for achieving the de-identification of personal data singles out 18 distinct identifiers—such as names, medical record numbers, and biometric identifiers—the removal of which would make the resulting information "not individually identifiable" (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html). This approach, which does not specifically include genetic data in the list of identifiers, has been criticized for, among other things, still allowing de-identified data to become re-identifiable through triangulation from other datasets. However, contrary to HIPAA, the burden of proof in case unwanted re-identification of anonymized genetic data occurs is, under the GDPR, always with data controllers, for the latter are responsible and accountable to ensure that de-identified genomic data remain as such throughout the full spectrum and context of any processing activity.

In addition, the GDPR's outcome- and context-based criterion of anonymization generates further difficulties impinging on data controller's accountability obligations, which could potentially lead to a "double bind" or "catch 22." The latter arises inasmuch as the processing of anonymous data—which is not subject to the technical and organizational safeguards (e.g., controlled-access models) required by the GDPR—inherently incentivizes data-sharing practices under open-access models. At the same time, however, the absence of safeguards and the increased circulation of data through open-access databases are factors that, in and of themselves, may increase the likelihood of re-identification of the individual, and thus de-anonymization of the dataset.

> *". . . when discussing re-identifiability in relation to genetic databases, attention should be also paid to the associated governance models for managing data access."*

A potential consequence of this could be that data controllers take a conservative approach and consider genomic data as, in principle, always identifiable. This might not necessarily favor scientific research or clinical use, as it could lead to restrictions in sharing and accessing data. To mitigate such risk, data controllers can refer to sectoral codes of conduct or international guidelines by professional societies (Art. 40–41), which provide specific guidance for the concrete implementation of the GDPR, for instance with regard to the processing of genomic data in scientific research. In addition, adherence to such guidance can represent one of the means for fulfilling controllers' accountability obligations. Examples of such guidance in the field of health research are as follows: the BBMRI-ERIC GDPR Code of Conduct for health research (http://code-of-conduct-for-health-research.eu), guidelines being developed by Alliance Against Cancer (https://www.alleanzacontroilcancro.it/en/commissione-acc-gdpr/) for personal data processing in the context of medical and research activities of Italian research hospitals (IRCCS), and guidelines for epigenetic data in the case of

Further readings

Identifiability: Concept and scope

Aamot H, Kohl CD, Richter D, *et al* (2013) Pseudonymization of patient identifiers for translational research. *BMC Med Inform Decis Mak* 13: 75

Bolognini L, Bistolfi C (2017) Pseudonymization and impacts of Big (personal/anonymous) Data processing in the transition from the Directive 95/46/EC to the new EU General Data Protection Regulation. *Comput Law Secur Rev* 33: 171–181

De Cristofaro E (2014) Genomic privacy and the rise of a new research community. *IEEE Secur Priv* 12: 80–83

Evans BJ, Jarvik GP (2017) Impact of HIPAA's minimum necessary standard on genomic data sharing. *Genet Med* 20: 531–535

Greenbaum D, Sboner A, Mu XJ, *et al* (2011) Genomics and privacy: implications of the new reality of closed data for the field. *PLoS Comput Biol* 7: e1002278

Malin B, Loukides G, Benitez K, *et al* (2011) Identifiability in biobanks: models, measures, and mitigation strategies. *Hum Genet* 130: 383

Malin BA (2005) An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *J Am Med Inform Assoc* 12: 28–34

Price WN, Cohen IG (2019) Privacy in the age of medical big data. *Nat Med* 25: 37

Quinn P, Quinn L (2018) Big genetic data and its big data protection challenges. Comput Law Secur Rev 34: 1000–1018

Ram N, Guerrini CJ, McGuire AL (2018) Genealogy databases and the future of criminal investigation. *Science* 360: 1078–1079

Rodriguez LL, Brooks LD, Greenberg JH, *et al* (2013) The complexities of genomic identifiability. *Science* 339: 275–276

Schmidt H, Callier S (2012) How anonymous is 'anonymous'? Some suggestions towards a coherent universal coding system for genetic samples. *J Med Ethics* medethics-2011-100181

Weil CJ, Mechanic LE, Green T, *et al* (2013) NCI think tank concerning the identifiability of biospecimens and "omic" data. *Genet Med* 15: 997–1003

Identifiability and access management

Milius D, Dove ES, Chalmers D, *et al* (2014) The International Cancer Genome Consortium's evolving data-protection policies. *Nat Biotechnol* 32: 519

Muddyman D, Smee C, Griffin H, *et al* (2013) Implementing a successful data-management framework: the UK10K managed access model. *Genome Med* 5: 1

Phillips M, Dove ES, Knoppers BM (2017) Criminal prohibition of wrongful re-identification: legal solution or minefield for big data? *J Bioeth Inq* 14: 527–539

Relevant guidelines, policy recommendations, and reports

Dyke SO, Cheung WA, Joly Y, *et al* (2015) Epigenome data release: a participant-centered approach to privacy protection. *Genome Biol* 16: 142

Ogbogu U, Burningham S, Ollenberger A, *et al* (2014) Policy recommendations for addressing privacy challenges associated with cell-based research and interventions. *BMC Med Ethics* 15: 7

PHG Foundation (2017) Identification and Genomic Data

Povey S, Al Aqeel AI, Cambon-Thomsen A, *et al* (2010) Practical guidelines addressing ethical issues pertaining to the curation of human locus-specific variation databases (LSDBs). *Hum Mutat* 31: 1179–1184

The Nuffield Council on Bioethics (2014) Collection, linking and use of data in biomedical research and health care: ethical issues

The Nuffield Council on Bioethics and the Wellcome Trust (2014) A Review of Evidence Relating to Harm Resulting from Uses of Health and Biomedical Data

U.S. Department of Health & Human Services (2012) Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule

Yim S-H, Chung Y-J (2013) Introduction to international ethical standards related to genetics and genomics. *Genomics Inform* 11: 218

rare diseases by an international working group addressing ethical issues relating to human Locus Specific Variation Database (LSDB).

While these soft laws could provide valuable tools for facilitating data sharing while addressing the risk of de-identification of genomic data, data controllers should still consider that de-identified data can always present residual risks, owing to advancements in technology and expertise, which should be dealt with appropriately and proactively in the concrete context of their processing. Adopting adequate safeguards is highly important considering the detrimental consequences that breach of privacy may have in undermining the reputation of data controllers as well as the trust by research participants.

Acknowledgements

M.S. is supported by a post-doctoral fellowship from Research Funders-Flanders (FWO). L.M. has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 753531.

Conflict Of Interest

The authors declare that they have no conflict of interest.

References

1. Marelli L, Testa G (2018) Scrutinizing the EU General Data Protection Regulation. *Science* 360: 496 – 498
2. Shabani M, Borry P (2017) Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. *Eur J Hum Genet* 26: 149 – 156
3. Article 29 Data Protection Working Party (2014) Opinion 05/2014 on Anonymisation Techniques.
4. Skopek JM (2015) Reasonable expectations of anonymity. *Va L Rev* 101: 691 – 762
5. Erlich Y, Narayanan A (2014) Routes for breaching and protecting genetic privacy. *Nat Rev Genet* 15: 409 – 421
6. El Emam K (2011) Methods for the de-identification of electronic health records for genomic research. *Genome Med* 3: 25
7. Greenbaum D, Sboner A, Mu XJ, Gerstein M (2011) Genomics and privacy: implications of the new reality of closed data for the field. *PLoS Comput Biol* 7: e1002278
8. Mourby M, Mackey E, Elliot M, Gowans H, Wallace SE, Bell J, Smith H, Aidinlis S, Kaye J (2018) Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. *Comput Law Secur Rev* 34: 222 – 233
9. P3G Consortium, Church G, Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J, Bobrow M, Weir B (2009) Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet* 5: e1000665
10. Shabani M, Knoppers BM, Borry P (2015) From the principles of genomic data sharing to the practices of data access committees. *EMBO Mol Med* 7: 507 – 509