Lost in Anonymization — A Data Anonymization Reference Classification Merging Legal and Technical Considerations

Health Policy Portal

Kerstin N. Vokinger, Daniel J. Stekhoven, and Michael Krauthammer

In recent years, advances in technology have enabled research with health data derived from large volumes of electronic health records (EHR) and other health-related data sources to improve innovation and quality in medicine.[1] This has also been accelerated through national and international efforts offering access to repositories containing an increasing amount of clinical knowledge and collaborative platforms harmonizing not only the algorithms used, but also ontologies enabling better interoperability.[2] At the same time there is growing concern that the use of health data for publicly-funded research may lead to exposure of patients' personal information, which potentially increases, among other things, risks for discrimination.[3] Legislators have addressed this issue by implementing regulations to protect patient privacy, often focusing on data anonymization, i.e., the removal or masking of identifiable information.

In this study we analyze, how the regulations in three jurisdictions (United States, European Union, Switzerland) distinguish between different levels of anonymization of health data, and assess whether and how these levels align with technical advancements.

Legal Overview

In the European Union (EU) there is no regulation specifically for health data. A general regulation, i.e., the General Data Protection Regulation (GDPR) regulates and protects the collection, processing, sharing, and storing of any data concerning an *identified or identifiable person*.[4] Also pseudonymized data fall within the scope of the GDPR. Pseudonymized data are data where obvious identifiers have been removed and replaced with a code. Individuals can be re-identified by using a key, therefore, also pseudonymized data are considered as identifiable data. However, the privacy protection regulations of the GDPR do not apply to anonymized or anonymous information, i.e., data where not only the identifiers, but also the key has been removed so that identification of the individual is no longer possible (anonymized data) or information that has been collected in such a manner that the individual is not identifiable (anonymous data). Whether data is considered anonymous or anonymized is tightly linked to the estimated effort needed to re-identify the patient providing the data, including, among other things, the costs, the amount of time required and the available technology.[5] If the effort for re-identification can be considered as "reasonable," the data is qualified as non-anonymized or non-anonymous. Whether the effort for re-identification is "reasonable" must be decided on a case-by-case basis. Since there is a spectrum of interpretation this leads to serious uncertainties in practice.[6]

Also in Switzerland, there is no regulation on federal level that addresses specifically health data. Like in the EU Switzerland has a federal act, the so-called "Federal Act on Data Protection" (FADP) that addresses the regulation and protection of data in general, including health data.[7] Swiss law distinguishes the same "anonymization levels" as the EU: Data concerning an identified or identifiable person fall within the scope of the FADP by contrast to non-identi-

About This Column

Aaron Kesselheim serves as the editor for Health Policy Portal. Dr. Kesselheim is the *JLME* editor-in-chief and director of the Program On Regulation, Therapeutics, And Law at Brigham and Women's Hospital/Harvard Medical School. This column features timely analyses and perspectives on issues at the intersection of medicine, law, and health policy that are directly relevant to patient care. If you would like to submit to this section of *JLME*, please contact Dr. Kesselheim at akesselheim@bwh.harvard.edu.

Kerstin N. Vokinger, M.D., J.D., Ph.D., LL.M., *is an Assistant Professor at the University of Zurich, Switzerland.* **Daniel J. Stekhoven, Ph.D.**, *is at NEXUS, Personalized Health Technologies, Swiss Federal Institute of Technology, (ETH) in Zurich, Switzerland.* **Michael Krauthammer, M.D., Ph.D.**, *is Professor for Medical Informatics at the University of Zurich, Switzerland.*

The Journal of Law, Medicine & Ethics, 48 (2020): 228-231. © 2020 The Author(s)
DOI: 10.1177/1073110520917025

fiable data. Like in the EU, pseudonymized data is considered as identifiable data, whereas anonymized and anonym data are qualified as non-identifiable data. The definitions are like in the EU. Data is considered as anonymized or anonymous if only an unreasonable technical effort can re-identify the data. Also under Swiss law, there is no specific definition of what an unreasonable effort is supposed to mean.[8] There is a scope of interpretation and decisions are made on a case-by-case basis.[9]

By contrast to Europe, the United States (US) has a specific act on federal level that addresses specifically health data, the so-called Health Insurance Portability and Accountability Act (HIPAA). By contrast to the European countries, the US has a different approach to the definition of "identifiable health data". Instead of asking about the effort needed for data reidentification, HIPAA specifies 18 identifiers — e.g., names, email addresses, social security numbers, or medical record numbers — that need to be removed for data to qualify as non-identifiable.[10] This approach leaves no scope of interpretation when deciding whether health data should be qualified as identifiable or not and may lead to less uncertainties in practice. However, studies show that the removal of these identifiers may still enable re-identification of individuals.[11] Alternatively, an expert can review and declare a data set as anonymized. There is no specific professional degree or certification program for designating who is an expert at rendering health information de-identified.[12] Experts may be found in the statistical, computer sciences, or other scientific domains.[13]

In this study we analyze, how the regulations in three jurisdictions (United States, European Union, Switzerland) distinguish between different levels of anonymization of health data, and assess whether and how these levels align with technical advancements.

Technical Analysis

Recent technical advances and the emergence of global efforts towards interoperable data resources result in a situation where data re-identification is increasingly likely, despite best effort to remove identifiable information. The existing data protection laws leave much uncertainty about whether de-identified data sets are within the scope of the laws. To remove such uncertainty, and to enable effective big data research with health information, we propose a move towards a more fine-grained legal definition and classification of the data de-identification steps (Table 1).

Let us assume the following hypothetical data set containing an EHR of a patient, including measurements of heart frequency over time, clinical images, and comprehensively sequenced DNA data. The EHR contains the name of the patient, the address, and other obvious identifiers allowing for direct identification. These obvious identifiers also include names and birthdates printed, for example, on x-ray images. If these obvious identifiers are removed and replaced with a code, the data set would be classified as pseudonymized in the EU and Switzerland, and non-indentifiable in the US. One reason for keeping a code is to be able to contact a patient, who has agreed to be informed about research results having a potential impact on his or her health. This is especially important in the case of incidental findings not directly related to the respective research done.[14] The removal of said code from the data set would — in the traditional perception — render it anonymous in all above described regulations. However, this is only

Table 1

Reference classification for levels of data anonymization.

Classification		Definition
Identifying or identifiable data		Data contains obvious identifiers of individuals.
Reversibly anonymized data	**Pseudonymized data**	Obvious identifiers are removed and replaced with a code; re-identification is possible via a key.
	Pseudo-anonymized data	All directly identifiable information is removed and no key exists to map the records back to the respective individuals. However, linkage to other available data sets enables re-identification.
Irreversibly anonymized data		Re-identification of individuals is impossible.
Anonymous data		Data has been collected on an anonymous basis or has been aggregated and re-identification of individuals is not possible.

The Journal of Law, Medicine & Ethics, 48 (2020): 228-231. © 2020 The Author(s)

https://doi.org/10.1177/1073110520917025 Published online by Cambridge University Press

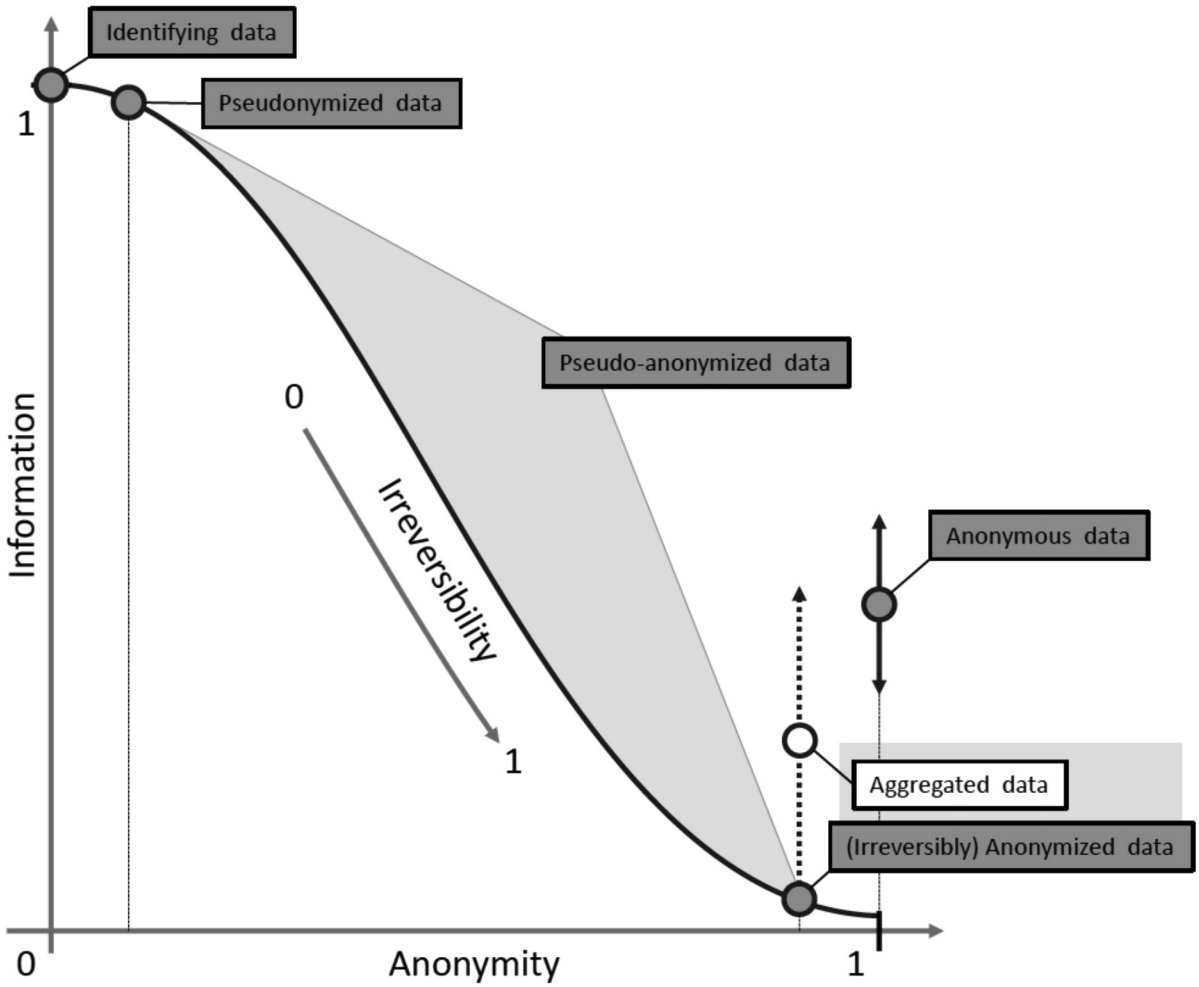
true, if the data set is kept isolated from linking it to other sources of information. This is why we propose the (new) class of *pseudo-anonymized* data. For example, the longitudinal data of heart beats acts as a unique fingerprint to another dataset due to potential linkage. This is possible for most sequentially recorded values of patients.[15] Of course this other data set needs to contain similar heart frequency measurements of the hypothetical patient in our example. The same is true for genetic data, which when sequenced comprehensively enough, will not only allow for linkage to other genomic data repositories, but will also allow to predict traits, such as hair and eye color. Also in this case, linkage would require the existence of said genetics profile to be present in another dataset, so would the personal description. We can reduce, but not eliminate this linkage probability by applying methods, such as data perturbation, which obfuscate the identifying signatures.[16]

If only summary data across patients are released, such as mean glucose levels over time, this can (still) be qualified as an irreversibly anonymized data set. In this case, the information contained in this class can again be substantial, at least on a cohort level.

Generally, it is important to note that there often exists a trade-off between the level of anonymization and capacity to conduct meaningful data analysis that may lead to advancements in medicine. Therefore, a strict application of anonymization may not always be helpful. While identifying data provides the maximum amount of information, but no anonymity, anonymous data provides the maximum degree of anonymity, but the amount of informa-

Figure 1

Schematic overview of classification for levels of data anonymization. While identifying and pseudonymized data contain the maximum degree of information and at the same time are the least anonymous. Pseudo-anonymized data dwell in a gradient dependent on the degree of applied irresversibility. The more anonymity is enforced, the less information is kept. The class of irreversibly anonymized data is reached, when re-identification is no longer possible.



The Journal of Law, Medicine & Ethics, 48 (2020): 228-231. © 2020 The Author(s)

https://doi.org/10.1177/1073110520917025 Published online by Cambridge University Press

tion may be limited. Especially in the domain of medical research, where the ultimate goal is to improve diagnosis, prognosis, and treatment of individual patients, patient-level data is indispensable. A too large degree of perturbation might therefore be unadviced, since it will not only obfuscate the identifying signatures, but also the biological signal under study. This is also true in some of the obvious identifiers. For example, the removal of obvious identifiers, such as ZIP codes, in the generation of reversibly anonymized data precludes research on comparative health issue across geographic regions.

Conclusion

Europe and the US have different approaches for defining "identifiable" or "non-identifiable" health data. However, the legal understanding of "identifiable" and "non-identifiable" health data in all three assessed jurisdictions (US, EU, Switzerland) is not congruent with the technological advancements. Removal of direct identifiers increasingly allow re-identification due to the advances in technology that allow the analyses of large volume data and linkage of different data sets that we refer to as "pseudo-anonymized data."

Ultimately, a legislation that respects technological advances and provides clearer legal certainty will allow a secure environment to drive medical advances while ensuring patient privacy.

Acknowledgements

The authors would like to thank Diana Elena Coman Schid, PhD (ETH Zurich), Franziska Singer, PhD (ETH Zurich), and Nora Toussaint, PhD (ETH Zurich) for their comments on a prior draft.

Note

The authors have not conflicts to declare.

References

1. C. Cassel and A. Bindman, "Risk, Benefit, and Fairness in a Big Data World," *JAMA* 322, no. 2 (2019): 105.
2. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, "Pan-Cancer Analysis of Whole Genomes," *Nature* 578, no. 7793 (2020): 82–93; See also <https://clinicalgenome.org/> (last visited February 27, 2020); The Global Alliance for Genomics and Health, A Federated Ecosystem for Sharing Genomic, Clinical Data," *Science* 352, no. 6291 (2016): 1278–80.
3. See *supra* note 1; W. N. Price and I. G. Cohen, "Privacy in the Age of Medical Big Data," *Nature Medicine* 25, no. 1 (2019): 37–43.
4. See <https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-does-general-data-protection-regulation-gdpr-govern_en> (last visited February 27, 2020); Art. 4 GDPR, *available at* <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1528874672298&uri=CELEX:32016R0679> (last visited February 27, 2020).
5. Global Alliance, *supra* note 2.
6. J.P. Albrecht, "How the GDPR Will Change the World," *European Data Protection Law Review* 2, no. 3 (2016): 287–289.
7. D. Rosenthal and Y. Jöhri, *Handkommentar zum Datenschutzgesetz* (Zurich/Basel/Geneva: Schulthess; 2008).
8. K. Vokinger and U. Muehlematter, "Re-Identifikation von Gerichtsurteilen durch «Linkage» von Daten(Banken). Eine Empirische Analyse anhand von Bundesgerichtsbeschwerden Gegen (Preisfestsetzungs-)Verfügungen von Arzneimitteln," *Jusletter* (2019).
9. *Id.*
10. Price and Cohen, *supra* note 3.
11. L. Sweeney, J. S. Yoo, L. Perovich, K. E. Boronow, P. Brown, and J. G. Brody, "Re-identification Risks in HIPAA Safe Harbor Data: A Study of Data from One Environmental Health Study," *Technology Science* (2017), *available at* <https://techscience.org/a/2017082801> (last visited March 13, 2020); V. Janmey and P. L. Elkin, "Re-Identification Risk in HIPAA De-Identified Datasets: The MVA Attack," *AMIA Annual Symposium Proceedings* (2018): 1329–1337.
12. See <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#determinations> (last visited February 27, 2020).
13. *Id.*
14. On behalf of the ACMG Secondary Findings Maintenance Working Group, S.S. Kalia, K. Adelman et al., "Recommendations for Reporting of Secondary Findings in Clinical Exome and Genome Sequencing: 2016 update (ACMG SF v2.0): A Policy Statement of the American College of Medical Genetics and Genomics," *Genetic Medicine* 19, no. 2 (2017): 249–55.
15. R. V. Atreya, J. C. Smith, A. B. McCoy, B. Malin, and R. A. Miller, "Reducing Patient Re-Identification Risk for Laboratory Results within Research Datasets," *Journal of the American Medical Informatics Association* 20, no. 1 (2013): 95–101.
16. *Id.*

The Journal of Law, Medicine & Ethics, 48 (2020): 228-231. © 2020 The Author(s)

https://doi.org/10.1177/1073110520917025 Published online by Cambridge University Press