J. Mantas et al. (Eds.)

doi:10.3233/SHTI200489

The Mishandling of Anonymity in Terms of Medical Research Within the General Data Protection Regulation

Daniel NASSEH [a,1]
[a] *Comprehensive Cancer Center, Munich, Germany*

Abstract One of the major regulatory factors for health informatics is data privacy protection. In the European Union, a shared set of laws has been implemented – the General Data Protection Regulation. While this set of rules aims at harmonizing the European data privacy protection standards, it fails in properly detailing the handling of anonymized data. This is a problem, as, for example many current research initiatives aim at reusing patient data collected within primary care, but lack a patient consent, hence, might rely on anonymized data as being the only alternative. Within this work, we detail different aspects why the concept of anonymity is wrongly handled within the GDPR and give suggestions how the laws could be adapted.

Keywords. GDPR, anonymization, data privacy protection

1. Introduction

The GDPR [1] is the European Union's approach to harmonize the laws of privacy protection of its participating nations. One of its goals was to create a regulatory toolset [2] against large data gathering cooperations. While it is doubtful that this goal was achieved [3,4], we can see some harmful effects of the GDPR on medical research [5].

This work is about a problem which impacts a subset of projects which want to work with routinely collected patient data, but lack a patient consent. As an example, this can be an issue for large scale approaches like the medical informatics initiative in Germany which would like to reuse most of a hospital's patient data, but due to missing consent cannot use any data retrospectively [6]. Without further ado, the contents within these valuable datasets are lost and have to remain unused. Though, there might be one alternative to this problem, which would be anonymization of the given data [7,8]. Unfortunately, as we will show within the section of methods, the GDPR does not value to detail the handling of anonymization and offers nearly no information concerning this important issue (GDPR: Recital 26). In combination with more severe punishment (GDPR: Article 83) academically this data is in danger to remain irrelevant, hence, as stated before, harming progress in terms of medical research.

[1] Corresponding Author, Daniel Nasseh, CCC-Munich-LMU, Pettenkoferstraße 8a, 80336 Munich, Germany; E-mail: Daniel.Nasseh@med.uni-muenchen.de.

2. Methods

Within the section of methods of this work we tried to identify the most problematic aspects of the GDPR in terms of anonymization. We list these aspects within the section of results and attempt at presenting a solution how the GDPR could be adapted to better support the option of academically working with anonymized patient data.

3. Results

Aspect 1 – Underrepresentation: The GDPR basically ignores the handling of anonymous data. The category has been represented solely within recital 26 which states that anonymous data is not person identifying, hence, does not need to be discussed within the GDPR: "[…] *This Regulation [GDPR] does not therefore concern the processing of such anonymous information* […]."

Aspect 2 – Responsibility: The GDPR's handling of anonymized data would make sense if data could easily be turned into fully anonymous data. That is not the case. With enough background information non-directly identifying data fields (MDAT) could be used to identify a patient whose IDAT (directly identifying data) had been removed (formal anonymization) [7,8]. Hence, patient records have to be altered in a way that they are not easily re-identifiable while conserving as much medical information as possible. Usually the most we can reach is factual anonymity [7] which can be achieved by e.g. altering or generalizing the so called quasi identifiers (typically fields containing dates or locations) [9,10] but with a potential risk of re-identification remaining [11]. Technologies like k-Anonymity aim at grouping individual patient tracks into similar groups of a specific size (k) [8]. Still, there remain ways to also foil this concept as the harsher concepts of l-diversity and t-closeness demonstrate [12,13]. In general, aside from identifying patients based on their individual features, they could potentially be identified just based on their treatment patterns. Working with factual anonymity is theoretically accepted and even supported by the GDPR. The GDPR states: "*To determine whether a natural person is identifiable, 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 (4).*" The problem about this section is, that it is absolutely unclear how to proof that data has been adapted adequately and who bears the responsibility about this decision.

Aspect 3 – Unregulated publication: As has been stated, the GDPR works for full anonymity. An example for that is aggregated data but we usually do not see individual anonymized patient tracks within publications as it is intuitive and has been shown before that even with high efforts they could potentially be re-identified. The question has to be raised: Should there really not be a difference in regulating factual anonymized data and fully anonymized data? At the moment the GDPR allows to treat both factual and fully anonymized data the same way: Hence, factual anonymous data which is, as shown before, potentially re-identifiable, could be under current laws published to the whole world, drastically increasing the possibilities of attack scenarios.

Aspect 4 – Technical challenges: As we have shown truly effective factual anonymity is hard to accomplish. Effective factual anonymity needs well trained informatics specialists. There might be tools like ARX which could help to set up e.g. k-anonymity [14], but as has been shown, depending on the project k-anonymity might not be sufficient. The competence to technically create factual anonymity and the skills to proof that the risk of identifying is diminishingly low is something only distinct expert groups, not available for most academic projects, could supply.

Aspect 5 – Contradiction: The biggest issue about the GDPR might be, that the GDPR contradicts itself. We can deduce that within the GDPR it is accepted that factual anonymity is not full anonymity. While full anonymity cannot lead to re-identification, factual anonymity could potentially be broken (else it would be full anonymity). Now, if data can, even if there is a diminishingly low chance, be re-identified, it is potentially person identifying and should, hence, be part of the GDPR. As recital 26 states, this is not the case, thus, the GDPR contradicts itself.

4. Discussion

The GDPR currently has no sound concept in terms of anonymity. It circumnavigates this shortcoming by claiming: Anonymous data is not person identifying (*see aspect 1*). We showed that this is not a good solution for medical research arising from the 5 presented aspects. With anonymity being the main option to having no patient consent, the resulting effect is often the loss of relevance for retrospective medical datasets without a consent. Applying factual anonymity under current legislation is not a real option as it can potentially, with enough effort and information, be broken (*see aspect 3*). If the threat of re-identifying factual anonymity cannot be excluded, factual anonymity should be seen as at least potentially person identifying (*see aspect 5*).

A solution could be: Within the GDPR fully anonymous data (usually aggregated data) should remain the class of data which cannot be re-identified and which can be used and published without any regulations. Aggregated data though, is not something academic projects usually utilize in the first place. Scientists need to work with microdata (e.g. individual patient tracks), with aggregated data (fully anonymous) being at the end of the chain of work, not at the start. Thus, a new class of data could be introduced: Rule based anonymous data. It would be a sub-class of factual anonymity, but not with the aim of providing maximal security against re-identification, but instead giving a strict ruleset which increases the security level significantly, but is easy to apply and understand (*see aspect 4*) without the unrealistic demand of creating nearly unbreakable factual anonymity if no consent is available. This approach would also supersede thinking about responsibility (*see aspect 2*), because the law would not be under arbitrary discretion but based on strict rules. Since rule-based anonymity would potentially be person identifying, the GDPR would have to be extended by rules for this class. All laws of the GDPR at the same time would apply for this class, hence a free and completely unregulated upload of this kind of data into e.g. the internet would not be an option. The use of rule-based anonymity should then be regulated: Examples could be, that persons working with this data must be specified, the circle of persons working with this data must be kept as small as possible and so forth. Most important regulation could be the purpose which could be interconnected with already existing recitals like considering a legitimate (*GDPR: Recital 47)* or public interest (*GDPR: Recital 54*). More general this could be an option for research projects with a direct benefit for society. We can only echo the conclusion of Mostert et. Al [15] that there should be research exemptions in data protection law.

This work is not about the formulation of such a ruleset – but it could in theory look something like the following: Appliance of a one way encryption on IDAT, identifiers and locations, date fields being trimmed to year, grouping of all data fields containing biometrical data like eye color or size. We have to stress that this is only a naive suggestion for a possible ruleset, but we believe that the concept of rule-based anonymization would be a better alternative than what we currently find within the GDPR. At the moment, the need of a consent without any real options within GDPR appears to be too heavily weighted. Richter et. Al showed within their study, that most patients within their survey would even be willing to abolish a consent completely if their data would be used for beneficial research [16]. In that regard, our call for a new class of anonymity and better handling within the GDPR seems to be modest.

5. Conclusions

As long as the GDPR is not reworked scientist will struggle to use anonymity as a serious alternative to a patient consent. Hence, retrospective data lose their value and their potential to be used as a source for science and research. While this work tries to shed light on the mishandling of anonymity within the GDPR it can only offer possible suggestions how this problem could be solved.

References

[1] General Data Protection Regulation (GDPR), https://gdpr-info.eu (accessed 25.03.2020).
[2] Kimberly A, Houser W, Gregory W, Voss W, Gregory Voss, GDPR: The End of Google and Facebook or a New Paradigm in Data Privacy? SSRN Electronic Journal (2018), 70
[3] Kostiv N, Schechner S, GDPR has been a boon for google and facebook, www.wsj.com/articles/gdpr-has-been-a-boon-for-google-and-facebook-11560789219, (accessed 25.03.2020).
[4] Bershidsky L, Facebook and google aren't hurt by GDPR but smaller firms are, https://www.bloomberg.com/opinion/ articles/2018-11-14/facebook-and-google-aren-t-hurt-by-gdpr-but-smaller-firms-are, (accessed 25.03.2020).
[5] Rumbold JMM, Pierscionek B, The Effect of the General Data Protection Regulation on Medical Research. J Med Internet Res 19(2) (2017), 47.
[6] Gehring S, Eulenfeld R, German Medical Informatics Initiative: Unlocking Data for Research and Health Care, Methods Inf Med 57 (2018), 46-49.
[7] Anonymity. https://www.forschungsdatenzentrum.de/en/anonymity (accessed 25.03.2020)
[8] Sweeney L, k-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, World Scientific 10(5) (2002), 557–570.
[9] Dalenius T, Finding a Needle In a Haystack or Identifying Anonymous Census Records. Journal of Official Statistics 2(3) (1986), 329–336.
[10] Anderson N, Anonymized data really isn't—and here's why not. Ars Technica (2009).
[11] Branson J, Good N, Chen JW, Monge W, Probst C, El Emam K, Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations. Trials 21 (2020).
[12] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M, l-diversity: Privacy beyond k-anonymity, ACM Transactions on Knowledge Discovery from Data 1, ACM, (2007).
[13] Li N, Li T, Venkatasubramanian S, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, ICDE 7 (2007), 106–115.
[14] Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA, ARX--A Comprehensive Tool for Anonymizing Biomedical Data, AMIA Annu Symp Proc (2014), 984-93.
[15] Mostert M, Bredenoord AL, Biesaart MC, Van Delden JJ, Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach, Eur J Hum Genet (2016), 956–960.
[16] Richter G, Borzikowsky C, Lieb W, Patient views on research use of clinical data without consent: Legal, but also acceptable?, Eur J Hum Genet 27 (2019), 841–847.