Commentary

BIG DATA & SOCIETY

Big Data & Society
January–June 2019: 1–5
© The Author(s) 2019
DOI: 10.1177/2053951718824352
journals.sagepub.com/home/bds
SAGE

Medical research, Big Data and the need for privacy by design

Bart Jacobs and Jean Popma

Abstract

Medical research data is sensitive personal data that needs to be protected from unauthorized access and unintentional disclosure. In a research setting, sharing of (big) data within the scientific community is necessary in order to make progress and maximize scientific benefits derived from valuable and costly data. At the same time, convincingly protecting the privacy of people (patients) participating in medical research is a prerequisite for maintaining trust and willingness to share. In this commentary, we will address this issue and the pitfalls involved in the context of the PEP project[1] that provides the infrastructure for the Personalized Parkinson's Project,[2] a large cohort study on Parkinson's disease from Radboud University Medical Center (Radboudumc), in cooperation with Verily life Sciences, an Alphabet subsidiary.

Keywords

Big Data, GDPR compliance, informed consent, medical cohort study, polymorphic encryption, privacy by design

This article is a part of special issue. Below disclaimer should be added after keywords as per style in both PDF and online XML: This article is a part of special theme on Health Data Ecosystem. To see a full list of all articles in this special theme, please click here: https://journals.sagepub.com/page/bds/collections/health_data_ecosystem.

Introduction

In this commentary we will share our experiences in the design process for secure and privacy-friendly data management in a large scale medical cohort study regarding Parkinson's disease. This study is carried out by Radboud University Medical Center, in cooperation with Verily Life Sciences Inc., a subsidiary of Alphabet. The public–private cooperation between an Academic Hospital and a large corporate stakeholder must involve explicit attention to aspects of privacy and data protection, to address concerns raised by participation of a commercial stakeholder with its own private interests.

As researchers involved in technological and legal protection of data and privacy in a broad sense we have become involved in this project. Our direct contribution was the design and implementation of a novel data management infrastructure with a strong emphasis on security and privacy. Participating in this project, together with medical researchers, has enabled us to gain an in-depth view into medical research practices and into the requirements that need to be fulfilled in order to perform such a study in a responsible way, and to contribute to the implementation of these requirements. Sharing lessons learned, especially from a computer science perspective, may provide good practices for future studies of this kind. In this commentary we extrapolate from this Parkinson's study to other medical studies.

Sharing of data within a research context is a prerequisite for gaining new scientific insights. At the same time this raises questions on aspects like data quality (standardization, comparability, methodology), privacy of the subjects participating in a study and rights of these participants regarding their data. Who will have

Institute for Computing and Information Sciences, Radboud University, Nijmegen, the Netherlands

Corresponding author:
Jean Popma, Institute for Computing and Information Sciences, Radboud University, Toernooiveld 212, 6525 EC Nijmegen, the Netherlands.
Email: j.popma@ru.nl

access to this data? Who will have access to study results and benefits? There is always the risk that existing data is used for other purposes than it was originally collected for, like in the Google-NHS case, where health data from hospitals was shared with Google to develop new AI technology (Hodson, 2016). This may lead to unlawful and/or ethically questionable processing of this data. It may be tempting to reuse data collected in one specific context (for instance, academic scientific research) in another, different context (for instance, commercial product development or profiling of individuals) (Nissenbaum, 2011). This will almost certainly result in frictions regarding personal interests of data subjects (security, privacy) on the one hand, versus interests of researchers, corporate actors (intellectual property) and the public interest (societal benefits of scientific progress) on the other.

Most of these frictions can be avoided if the data can be kept in their original context – applying principles of privacy by design (Hoepman, 2014) – in an early stage of defining research plans when all processes needed for adequate protection of personal rights are designed and implemented.

Our primary role in this project is to provide novel (cryptographic) methods for the protection of the confidentiality of data, thereby protecting the privacy of participants in this study. These methods focus on pseudonymisation and encryption, in so-called polymorphic form. The technical details have been described elsewhere and are not of direct concern here (Verheul et al., 2016).

In the first two years in our support-role to the Parkinson's study, it became clear that in order to effectively protect data and privacy of study participants, four basic processes need to be organized and implemented: informed consent, data governance, data use agreements and data security. These four processes were among several areas of interest that had been identified at the start of the project, based on experiences from previous studies. We single out these four processes here since during the project itself it became obvious that there were major pitfalls involved. It is about these four processes that we gained valuable experience that we share in this commentary. In the following, we will discuss these pitfalls more elaborately and we will try to formulate guidelines for responsible processing of shared scientific data.

Distinguishing informed consent from other matters

For the purpose of scientific medical research the only legal basis for processing and sharing of data is informed consent, at least in European countries.[3] A participant in a study must give this consent after he or she is well informed about the consequences (risks) of participating, the purpose of the study, the way data will be used and secured and the way he/she can exercise his/her basic rights to inspection, correction and erasure of data. Consent must be obtained in such a way that subjects can distinguish the consent information from other matters using clear and plain language. People suffering from a specific disease are often highly motivated to participate in a medical research project. Not as much for their own personal benefit but more altruistically, as a personal contribution to the improvement of future treatment (Nobile et al., 2013). In writing consent documentation (including consent forms) for participants there are two issues that cause a potential conflict with the basic principle of informed consent.

The first problem (the ''glossy'' syndrome) is that a researcher recruiting participants for a study usually has an interest to find as many suitable participants as possible. Recruiting a cohort of scientifically relevant size is not trivial. A researcher might be tempted to document the consent information with this in mind, in order to motivate people to participate. Participants might even be stimulated to discuss their participation on social media, in order to promote the project and make recruitment of suitable participants easier. This is not in line with the idea behind informed consent, where a clear and comprehensible description of all the risks for the participant in a study should be documented, clearly distinguishable from other matters and interests. An example of this is the sharing of data with researchers in countries that have a lower level of protection of personal data in their national legislation. This introduces risks of unintended use of the data – for instance for surveillance or profiling purposes. Being clear about all the personal risks involved might discourage potential participants to participate, while not presenting this information would disable a free and informed decision.

The second problem with informed consent (the ''legal'' syndrome) is that even if the information presented to a potential participant is correct and complete, the documentation becomes very comprehensive, in almost legal wordings in order to avoid any future liabilities. Many examples of informed consent documentation presented to potential study participants suffer from either the glossy or the legal syndrome, often from both at the same time.

Within the aforementioned Parkinson's study we have contributed by reviewing the consent documentation shared with potential participants. As a result the glossy and legal components were stripped from the basic consent information, focusing on the participants and their privacy risks. Legal and promotional aspects can and should still be described in other informational

documents, as they can provide relevant information for participants.

Data governance should not end with sharing

Data governance is a process encompassing the organization of tasks, responsibilities and control over data. Here we shall focus on access control for personal (medical) data in a research context using the framework provided by the General Data Protection Regulation (GDPR). Data governance is not a simple task nor is it a one-time effort. The most important responsibility is to keep data within its original context (tightly coupled to the informed consent) at all times, during the entire life span of the data. Through data governance, it must also be ensured that data is handled in a secure way, by providing the means to manage and store this data securely until it is shared. Certainly if data is voluminous this can be an expensive obligation. Often the cost of security and management of data over a prolonged period (it is a good practice to keep data available for up to 25 years after termination of a research project) is underestimated. Potentially, either storage quality and/or data protection are compromised, with all the risks involved. It is clear that data governance is a crucial responsibility, requiring communicational, organizational, legal and technical skills and long-term financial means, which in combination tend to be scarce resources.

Collecting and storing of data is typically followed by sharing the data with researchers. This means that there must be a system in place to ensure that all organizational, legal and technical prerequisites are met. This involves taking informed and transparent decisions on requests for data access by other researchers (Is the intended research and use of the data in line with the consent given?). The GDPR recognizes the roles of controller (the responsible person/organization providing the means and determining the goals of data processing) and data processor (processing the data on behalf of the controller). Their relationship is documented in a processor agreement. This is the general GDPR context. While some use of research data might fit in the controller/processor model used in the GDPR, in most cases sharing of research data involves transfer of controllership over this data. The difficulty is that the consent was given to the original controller, whilst the new controller who uses shared data has no relationship with the data subject (study participant) nor any knowledge of his/her true identity. This means that far-reaching contractual arrangements must be made by the original controller to be able to exercise his/her obligations to the study participant bound by the participant's consent that would hold even when he/she transfers formal controllership over the data. This problem is often overlooked or not addressed properly. In fact an agreement is needed that bears a great resemblance to a processor agreement, but lacks the legal (GDPR-defined) status of such an agreement. Drafting and implementing such an agreement is a tedious task, requiring many resources. Taking care of the data, its use, and communication about the use does not end with the sharing of the data and the transfer of controllership. It is a persisting responsibility for the original controller.

In the case of the Personalized Parkinson's Project, data governance is formally part of the responsibilities of the Principal Investigator at Radboud University Medical Center, enabled by the informed consent obtained from the study participants. The operational aspects are partly exercised through a Research and Data Sharing Review Board, in which all major stakeholders are represented. Any use of the data must be approved by this board. It is a pitfall that data governance receives a great deal of attention when the project is started, but it should remain effective for the entire life cycle of the research data, a period often exceeding the presence of the Principal Investigator him/herself. It must therefore be anchored at the organization level. We observe that few organizations have instruments to accommodate for this.

Data use agreements: The legal basis for data sharing

As stated in the previous paragraph the original controller can only exercise his/her responsibilities, obligations and rights regarding data shared with other researchers through a legally binding data use agreement. Drafting such an agreement is a complicated and expensive legal exercise. First, such a data use agreement must ensure that the use of the data is within the bounds of the original purpose (and the informed consent), and that all obligations of the original controller regarding the data and the data subject can be met after formal controllership has been transferred to the data user. Examples of this include obligations such as reporting of accidental findings that are severe and relevant to the data subject (such as for instance evidence for the presence of carcinoma), a strict policy prohibiting de-pseudonymisation or reporting of data breaches that might lead to exposure of privacy-sensitive information. Second, a data use agreement should also enforce the requirements for data protection, confidentiality and pseudonymisation, as defined in relevant ISO-standards.[4]

The third area that needs to be covered is intellectual property. A minimum requirement in the

aforementioned Parkinson's project is that all derived data (derived from other data or from bio-samples) should be contributed to the study repository and the original controller – through his/her governance responsibilities – must be allowed to share this data with other researchers. Examples of derived or enhanced data are for instance new analyses performed on shared bio-samples, or improved representations of raw fMRI-data. Whether the same holds for intellectual property (IP) and scientific findings is up for debate. In general the data use agreement must be very clear on the rules that apply to IP. In a scientific context researchers will have an interest to publish their results first before sharing them or the data they have used to produce them. In a commercial research context companies may want to protect their findings prior to sharing them, to be able to commercialize the outcomes of their research. It is our opinion that if original data is produced with public funding and voluntary contributions of study participants, the common interest (in a *civic* repertoire, Sharon, 2018) has to be addressed to avoid monopolization of this data.

Last but not least: a data use agreement should always include a right for the original controller to audit the implementation of the obligations specified in the agreement. Non-compliance with these obligations should lead to revocation of the agreement to use the data. The data use agreement is therefore the sole instrument by which data governance can be exercised over the entire data life cycle.

Data security and privacy controls: Scientific research vs. health care

The fourth pillar to build upon is data security and privacy controls. One of the major pitfalls to deal with is that when conducting scientific medical research, one has to be aware that this is a fundamentally different context from the regular health care context that most people involved are familiar with. The same staff involved in regular health care-related patient assessments, lab-analysis or equipment operation are often also involved in the collection and processing of scientific research data. Whereas the regular health care processes require repeated identification of a patient, in order to avoid mix-up of people or treatments, the scientific process requires deep pseudonymisation of all information flows in order to avoid unnecessary identification of participants. This fundamental difference in context – while working with the same staff and analysis processes – is a permanent source of misunderstanding or incorrect handling of data. We have seen that training people to be aware of this difference and acting accordingly requires a lot of effort. A second pitfall is that data needs high levels of protection (information security) and inherent privacy guarantees to prevent leakage or unlawful combination of data. Given the long life span of the scientific data, risks are considerable. There are many known cases of large scale data leakage in recent years (AP, 2017; Joseph, 2017). New technology is required to protect data over their entire lifecycle using principles of security and privacy by design.

In order to facilitate this we – as computer science experts – have worked with the Personalized Parkinson's Project to devise new ways of data creation, storage and sharing. We have developed a technology called Polymorphic Encryption and Pseudonymisation (Verheul et al., 2016). This technology enables strong encryption of all research data in or near to the data source, remaining in an encrypted state during transport and storage of data in the research data repository. No data-management staff, hosting partner or cloud-service provider has the ability to access and decrypt the data. Only legitimate users (approved by the Research and Data Sharing Review Board) receive specific cryptographic keys for those parts of the data they are entitled to. They will be able to download this data in encrypted form from the repository and decrypt them in a secured working environment where data can be analysed. Encryption keys are managed in a distributed way, which means that legitimate users receive parts of their key from different actors, and only in combining the parts a correct key is constructed. Added to that, data is pseudonymised in such a way that every user (actually every project team) receives a unique pseudonym for a study participant. This means that pseudonyms cannot be exchanged between researchers working in different projects. Using this technique data is protected from leakage, hacking or theft while in movement or while stored in the study repository. Pseudonymisation is strong and distributed, making exchange of data and re-identification unlikely.

Far from utopia

All measures and precautions discussed here – technical, organizational and legal – contribute to the protection of the privacy of study participants. None of these measures are fail-safe however. They need to be aligned and reinforce each other. Especially human error or misconduct can lead to serious privacy risks. As an example, a data use agreement may prohibit depseudonymisation of data or linking of data to other sources. However, once a legitimate user has received and decrypted data from the repository, this data can be leaked. And data is often self-identifying, especially if combined with other sources (Elliot et al., 2018). A total genome sequence for instance is unique and can

therefore be used in forensic contexts to identify people with a very high degree of certainty. Also, common daily practice acting as a medical professional is very different from handling data as a researcher, especially when multiple roles are exercised by the same staff. Another example of this is that participants themselves should be asked to restrain themselves in the use of social media concerning their participation in the research project as such exposure undermines any pseudonymisation effort (Elliot et al., 2018). The awareness of such context-specific responsibilities and good practices is of great importance for the long-term protection of the data and the privacy of the study participants.

Conclusion

Responsible management of medical data for research purposes requires a multidisciplinary professional approach. Based on the practical experience from a project like the Personalized Parkinson's Project, good practices and new technological approaches can be combined and field tested. Consent, governance, legal data use agreements and data protection require specialized professionals in areas of communication, medical, legal, organizational and technical fields who should be aware of the complexities of the entire process of creating and sharing scientific data. Only an integrated multidisciplinary approach can offer sufficient guarantees for the protection of personal data and the privacy of study participants, although it is impossible to reduce the risks involved to zero.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/ or publication of this article: This work was funded by the Province of Gelderland grant 2016-006615.

Notes

1. See: https://pep.cs.ru.nl for details.
2. See: http://www.parkinsonopmaat.nl
3. Following the GDPR (http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf)
4. That is, ISO27001 and ISO27002 for information security and ISO25237 for pseudonymisation.

ORCID iD

Jean Popma http://orcid.org/0000-0002-2369-1655

References

AP (2017) Quarterly report of data breaches in the Netherlands, 2nd quarter. Available at: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/2017-10-03_2017_q3_kwartaalrapportage_algemeen.pdf (accessed 31 December 2018).

Elliot M, et al. (2018) Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review* 34: 204–221.

Hodson H (2016) Revealed: Google AI has access to huge haul of NHS patient data. *New Scientist* 29 April, (3072). Available at: https://www.newscientist.com/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/.

Hoepman J-H (2014) Privacy design strategies. In: *IFIP TC11 29th international conference on information security (IFIP SEC 2014)*, 2–4 June 2014. Springer, pp. 446–459.

Joseph R (2017) Data breaches: Public sector perspectives. *IT Professional* 20(4): 57–64.

Nissenbaum H (2011) A contextual approach to privacy online. *Daedalus* 140(4): 32–48.

Nobile H, Vermeulen E, Thys K, et al. (2013) Why do participants enroll in population biobank studies? A systematic literature review. *Expert Review of Molecular Diagnostics* 13(1): 35–47.

Sharon T (2018) When digital health meets digital capitalism, how many common goods are at stake? *Big Data Society* 5(2): 1–12. DOI: 10.1177/2053951718819032.

Verheul E, Jacobs B, Meijer C, et al. (2016) Polymorphic encryption and pseudonymisation for personalised healthcare: A whitepaper. Available at: https://eprint.iacr.org/2016/411.pdf (accessed 31 December 2018).