Journal of Parkinson's Disease xx (2021) x–xx
DOI 10.3233/JPD-202431
IOS Press

Review

Data Protection Using Polymorphic Pseudonymisation in a Large-Scale Parkinson's Disease Study

Bernard E. van Gastel*, Bart Jacobs and Jean Popma
Interdisciplinary Hub for Security, Privacy and Data Governance, Radboud University, Nijmegen, the Netherlands

Accepted 11 May 2021
Pre-press 1 June 2021

Abstract. This paper describes an advanced form of pseudonymisation in a large cohort study on Parkinson's disease, called Personalized Parkinson Project (PPP). The study collects various forms of biomedical data of study participants, including data from wearable devices with multiple sensors. The participants are all from the Netherlands, but the data will be usable by research groups worldwide on the basis of a suitable data use agreement. The data are pseudonymised, as required by Europe's General Data Protection Regulation (GDPR). The form of pseudonymisation that is used in this Parkinson project is based on cryptographic techniques and is 'polymorphic': it gives each participating research group its own 'local' pseudonyms. Still, the system is globally consistent, in the sense that if one research group adds data to PPP under its own local pseudonyms, the data become available for other groups under their pseudonyms. The paper gives an overview how this works, without going into the cryptographic details.

Keywords: Computer security, privacy, big data, data protection

INTRODUCTION

A doctor is not supposed to treat patients as numbers. A medical researcher on the other hand should only see numbers (pseudonyms), not individuals. This is a big difference, which requires that the same person—a doctor also doing research—acts and thinks differently in different roles. It is a legal, data protection requirement to hide the identity of participants in scientific studies. Additionally, people are in general only willing to participate in medical research if their identity remains hidden. Hence, it is in the interest of researchers themselves to thoroughly protect the data and privacy of their study participants, in order to provide sufficient comfort and trust to participate, now and in the future. With the increasing number of large-scale data gathering studies, high quality protections need to be in place, mandated by both legal requirements and ethical considerations.

We study how pseudonymisation can be applied to actual large scale data gathering projects to protect the data of participants. Pseudonymisation is one of many data protection techniques. The aim of pseudonymisation is to decrease the risk of reidentification and to decrease the risk of data being linked to data concerning the same participant without approval. This is difficult, for several reasons.

- Many other sources, outside the research dataset, such as social media, provide publicly available information that facilitates re-identification.

*Correspondence to: Bernard E. van Gastel, Radboud University (iHub), Erasmusplein 1, 6525 HT Nijmegen, the Netherlands. Tel.: +31 24 3652632; E-mail: b.vangastel@cs.ru.nl.

- Medical datasets are often very rich with many identifying characteristics, for instance with DNA or MRI data that match only one individual. Thus, the data themselves form persistent pseudonyms.
- Continuous input from wearable devices provides identifying behavioral data or patterns.

The context of this paper is formed by a large cohort study on Parkinson's disease called Personalized Parkinson's project (PPP, see [2] for an elaborate description of the study). This project aims to overcome limitations of earlier cohort studies by creating a large body of rich and high-quality data enabling detailed phenotyping of patients. The PPP aims to identify biomarkers that can assist in predicting differences in prognosis and treatment response between patients. It is clear that collecting such data is an elaborate and costly undertaking. Maximizing the scientific benefits of these data by sharing them for scientific research worldwide is therefore important, and one of the explicit goals of the PPP project. Sharing sensitive biomedical and behavioural data is a challenge in terms of legal, ethical and research-technical constraints.

To enable responsible ways of data sharing, a novel approach has been designed and implemented in the form of a data repository for managing and sharing of data for the PPP project. This design involves so-called Polymorphic Encryption and Pseudonymisation (PEP, see [8, 9]). The PEP system improves over the current best practices in pseudonimisation techniques as described in [1] by using a stronger form of pseudonimisation based on asymmetric encryption, in such a way that enables sharing of data amongst different researcher groups, while not relying on a third party for pseudonimisation. Sharing of data amongst different research groups requires an easy but secure way to translate pseudonyms as used by one research group to pseudonyms as used by another research group. This implies that pseudonimisation should be reversible, which is not the case for some of the methods described in the aforementioned overview of best practices. Using a third party introduces a large amount of trust into one single party. If that party acts in bad faith, data protections can be circumvented. The infrastructure hosting the PPP data repository is referred to as the PEP-system, managed by the authors. Design and development of the system was done in close cooperation with the PPP team. Of course this approach is not restricted to the PPP study, but this study is the first implementation and thus provides an example for use in future studies, also outside the field of Parkinson's disease.

The working of our PEP system can be laid out after describing the legal requirements to pseudonymisation and constraints for pseudonymisation. We will briefly describe the PEP system, at a functional level, with emphasis on polymorphic pseudonymisation—and not on encryption, even though pseudonymisation and encryption are tightly linked in PEP. Pseudonymisation has become a scientific topic in itself, but this article focuses on the practical usage of the relatively new technique of polymorphic pseudonyms, whereby, roughly, each research group participating in the study gets a separate set of pseudonyms.

PSEUDONYMISATION AND THE GDPR

In health care, it is important to establish the identity of patients with certainty, for various reasons. First, the right patient should get the right diagnosis and treatment, but for instance also the right bill. In addition, the patient's identity is important for privacy/data protection, so that each patient gets online access to only his/her own files and so that care providers discuss personal medical details only with the right individuals. In hospitals patients are frequently asked what their date of birth is, not out of personal interest, but only in order to prevent identity confusion.

In contrast, in medical research identities need not and should not be known, in principle. Certain personal attributes, like gender, age, etc., are useful in certain studies, but the identity itself is not relevant. Indeed, if such attributes are used, they should not be identifying. Treatment and research are two completely different contexts [4], each with their own legal framework and terminology. We shall distinguish 'patients' and 'health care professionals/providers' in care, and 'study participants' and 'researchers' in research. In practice the same person can be active both in health care and in research and switch roles frequently. The awareness of this context difference is therefore an important job requirement.

In general, one can hide an identity via either anonymisation or pseudonymisation. After applying pseudonymisation there is still a way back to the original data set when combined with additional information. With anonymisation this is

impossible. Europe's General Data Protection Regulation (GDPR) uses the following descriptions.

- In Art. 4, pseudonymisation is defined as "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organizational measures to ensure non-attribution to an identified or identifiable individual".
- Anonymisation is described in Recital 26 as "... 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".

The difference is legally highly relevant, since the GDPR does not apply to anonymous data. But it does apply to pseudonymous data.

In larger studies, such as PPP, participants are monitored during a longer period of time and are (re-)invited for multiple interviews and examinations. In such a situation a random number (a 'pseudonym') is assigned to each participant and this number is used in each contact with the participant during data collection. The GDPR thus applies to such studies.

Proper pseudonymisation is a topic in itself (see, e.g., [6]) since re-identification is always a danger [7]. Just gluing together the postal code and date of birth of a participant and using the result as pseudonym is not acceptable: it does not fit the above requirement: "... subject to technical and organizational measures to ensure non-attribution". Pseudonymisation can in principle be done via tables of random numbers, often relying on trusted third parties to perform the translation. However, modern approaches use cryptographic techniques, like encryption and keyed hashing, see e.g., [5], which generate pseudonyms—the entries in such random tables. On the one hand such cryptographic techniques form a challenge, because they require special expertise, but on the other hand, they offer new possibilities, especially in combining pseudonymisation, access control and encryption. Researchers retrieving data from PEP need to authenticate (prove who they are) before they can get encrypted data that they are entitled to, together with matching local pseudonyms and decryption keys. There are commercial pseudonymisation service providers, but outsourcing pseudonymisation introduces additional dependencies and costs and makes such integrated pseudonymisation difficult. This paper describes how so-called polymorphic pseudonymisation protects participants—and indirectly also researchers.

CONSTRAINTS

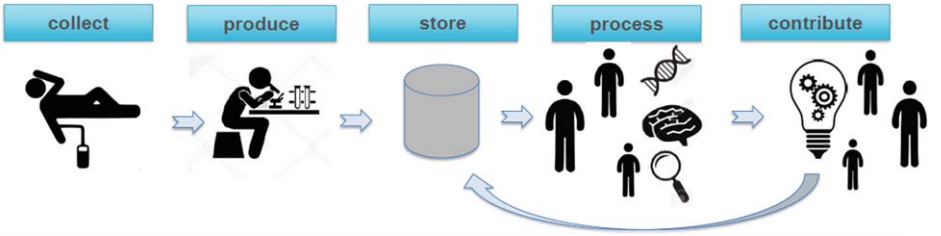
In practice, there are a number of constraints on how pseudonimisation can be applied. There are a number of sources these constraints originate from: biological data properties, from standard practices such as how to handle incidental findings, and how bio samples are handled. These constraints need to be taken into account into a system for data management. We can classify these constraints in the following types: re-identifying due to the nature of the data, re-identifying in combination with additional outside sources, regulation-based constraints, and practical constraints.

Even when pseudonyms are generated and used properly, digital biomedical data itself may lead to re-identification. This may happen in two ways.

1. Such biomedical data often contain patient identifiers. The reason is that devices for generating such biomedical data, like MRI-scanners, are also used for health care, in which it is important to ensure that data are linked to the right patient. Such devices may thus automatically embed patient identifiers in the data.
2. The biomedical data itself may be so rich that it allows for re-identification. This may be the case with MRI-scans of the brain which contain part of the face, or with DNA-data from which identifying characteristics can be deduced, or which can be compared to other DNA-databases

Hence, whatever the (cryptographic) pseudonymisation technique is, technical measures must be in place to remove such identifying characteristics from the data, especially of the first type.

An early experience in the PPP study illustrates the point raised above: through some error, a couple of MRI-scans got detached from their (internal) pseudonyms in the PEP-system. This is a disaster because it means that the scans have become useless. But the MRI-operator was not concerned at all and said: next year, when study participants return for their next visit (and MRI-scan), I can easily reconnect the few lost scans with matching new ones! Such matches do not involve the identities of the study participants but work simply because such an MRI scan is uniquely identifying any participant.



Who	Assessor	Producer	Repository	Researcher
Identity information	Identifiable	Pseudonymised	Encrypted/ Pseudonymised	Pseudonymised
Data	Partly Clear text	Clear Text	Encrypted	Clear text
Linkability	Linkable to person	Unlinkable	Unlinkable	Linkable to pseudonym

Fig. 1. Phases in research data management.

Combining the data with outside sources could also lead to re-identification. Besides the data itself, metadata such as timestamps can be identifying, especially if combined with external additional data sources. Not storing metadata such as time stamps is not always enough: if there is frequent data synchronization, it is easy to determine when certain data was first made available. This is with high probability the day a participant came in for tests. The effect of combining external information sources is not always clear for everybody involved, as can be seen in another experience of our team. Finding study participants is always a challenge, and so the staff of the PPP study suggested to people who had already signed up that they could enthusiastically report their PPP-involvement on social media, in order to further increase participation. We were upset when we heard about this and had to explain how such disclosures on social media undermine our careful pseudonymisation efforts. Data protection, including pseudonymisation, is not solely a technical matter. It can only work with a substantial number of study participants whose participation is not disclosed, at least not in a way that allows linking to actual data in the system, like for instance date and time of visits. The smaller the population of participants, the easier it is to single them out based on just very little information.

There are also practical constraints. The pseudonyms used need to be human-readable and short enough to fit on labels. However, the used polymorphic pseudonyms are based on non-trivial cryptographic properties, namely the malleability of El Gamal encryption. The handling of these pseudonyms within the PEP-system happens automatically and is not a burden for the researchers that interact with the system for storing and retrieving study data. Internally, these pseudonyms are very large numbers (65 characters long, see below). In fact, they are so large that they are not suitable for external usage by humans. For instance, these internal pseudonyms do not fit on regular labels on test tubes. As a result, shorter external representations of these polymorphic pseudonyms are used for input—and also output—of the PEP-system.

There are regulation-based exceptions. In scientific research on medical data, researchers may come across an 'incidental finding' about the health condition of the study participant. Under such special circumstances it may be needed to de-pseudonymise deliberately, and to turn the study participant into a patient. In some studies, like in PPP, separate, exceptional procedures must be in place for such cases.

PEP IN FIVE PHASES

The five phases of research data management via the PEP-system are summarized in Fig. 1, with special emphasis on the identity of the study participant.

These phases will be discussed separately below. They suggest a certain temporal order, but in practice these phases exist in parallel, for instance because study participants are followed during a longer time (two years, in principle) and are monitored (1) during repeated visits at discrete intervals, and (2) also continuously, via a special sensor based wearable device in the form of watches, provided by Verily.

Plasma visit 1	POM1PL0610058
fMRI visit 1	POM1FM2641022
ECG visit 1	POM1EC5844271
PBMC visit 1	POM1PM2962819
DNA visit 1	POM1DN8293883
RNA visit 1	POM1RN6147864

Fig. 2. Examples of short pseudonyms, also printed on a test tube. Such a short pseudonym like POM1PL0610058 consists of five random numbers (here 10058), a prefix (POMPL) identifying the project (POM, Dutch for PPP) and type of data (1PL for plasma taken during the first visit), and a checksum (58) such as used in an International Bank Account Number (IBAN).

The PEP-system makes crucial use of so-called polymorphic pseudonyms for study participants. These are large numbers that typically look as follows—65 characters in hexadecimal form: 0EAD7 CB2D70D85FE1295817FA188D22433C2237D946 964A6B4C063E6274C7D7D.

Such numbers are not meant for human consumption. They have an internal cryptographic structure so that they can be transformed to another such number, called a local pseudonym, to be used by a particular researcher (or group of researchers) of the PEP-system.

This transformation of pseudonyms is done by several, independently managed components of the PEP-system, working together to produce the final pseudonymisation result. This is done "blindly", which means that the components of the system perform their tasks without knowing the identity of the study participant involved[1]: internally the system is able to produce a persistent local pseudonym for each research-group.

In this way each international research group that joins the PPP research effort gets its own local pseudonyms for all the study participants. If for some reason data get compromised or re-identified in such a research group, the effect is local, in principle, and does not affect the whole system. There are further organizational and legal safeguards, implying for instance that participating research groups get access to only the data that is relevant for their research questions, and that the data may only be used for specific and well-defined purposes, but that is a different topic.

We now discuss the five phases in Fig. 1. The collection phase consists of two parts, with repeated site-visits and with continuous monitoring via sensor-based wearable devices. To the participant they have the appearance and function of a watch, hence the term 'study watch' is used in practice. These site-visits typically take a whole day and involve several medical examinations, tests, and interviews. During such a visit, a dedicated assessor accompanies each study participant. During the first visit the study participant receives a study watch, provided by Verily Life Sciences, which is permanently active—except during daily charging and transmission of collected data. The watch contains a serial number, which is linked to a local polymorphic pseudonym associated with Verily. Verily receives the combination <serial-number, local-pseudonym>, but does not learn which study participant gets which watch. Verily collects the data from the watches into its own system, via the serial number, and uploads the sanitized data into the PEP-system via the associated local pseudonym.

For each visit of a study participant the assessor gets an automatically prepared table of short external pseudonyms, connected to the internal pseudonym that is associated with the study participant. At each lab for biospecimens or measurement (say for MRI), the associated short pseudonym is put on the relevant tube or file, see Fig. 2.

The production phase is for cleaning-up and for producing data in such a format that it can be uploaded into the PEP-system. This may involve sanitization like for the study watch data, transforming raw data into data that can be used in further

[1] As an aside: since these components also perform logging, it is cryptographically ensured that nothing can happen in the PEP-system without producing an audit log.

analysis, or performing measurements on biospecimens. The measurement devices in the PPP study are typically also used in a health care setting, for patient diagnosis. This means that the output files often contain identifiers that become persistent if not removed before uploading data to the repository. They have to be removed from the data, before upload to the PEP-system, to prevent unwanted re identification. Similarly, MRI-data must be de-faced, so that the study participants cannot be recognised visually or by face-recognition algorithms. During the storage phase, the sanitised, appropriately formatted data is encrypted, uploaded and stored. This encryption is also done in 'polymorphic' manner but how that works is out of scope here (see [3] for some more information, and [9] for cryptographic details). Actual storage happens in a data centre of Google in Europe. Since all stored data are encrypted and encryption keys are handled internally in the PEP-system, the actual storage provider is not so relevant, since the data is protected from the storage service provider. What matters most is that the encrypted data is available only to legitimate users of the PEP-system, when needed, and is not corrupted. The processing phases exists for research-groups that have been admitted to the PPP project, after submitting and approval of a research plan, and after signing a data use agreement (see [3]). Such a user-group then gets its own local pseudonyms and a decryption key for locally decrypting a download from the PEP-data repository containing the research dataset it is entitled to. Typically, this dataset is necessary and minimised for the approved research plan. Such a research-group thus has access to unencrypted (clear text) medical research data, but only in a locally pseudonymised form. Data may only be processed in a secured processing environment. There is an additional contribution phase in which such a research-group can return certain derived results to the PEP-systems. This group uses its own local pseudonyms for the upload. Once uploaded, these additions become available for all other participating research-groups, under their own local pseudonyms. The PEP-system enures consistence of these pseudonyms, via its blind translation mechanism.

The table in Fig. 1 provides an overview of the identity and encryption status of each of the phases and of the names/roles involved. Only during the collection phase, the identity of the study participant is (necessarily) known, for instance to the assessors, for personal contact and logistic purposes. Personal, identifying data like name and contact information of study participants are stored within the PEP-system, but these data are never handed out, except via a very special procedure, involving a designated supervisor, for emergency reporting of incidental findings.

FINAL REMARKS

We have described the main lines of the PEP-system that is designed and built for privacy-friendly management of research data, focused on medical studies.

The PEP methodology combines advanced encryption with distributed pseudonymisation, and distribution of trusted data with fine-grained access management, allowing access to be restricted to subsets of participants and subsets of different data types. It thus allows cooperation of public and private research organizations on a global level. PEP provides secure access to minimized datasets with local pseudonymisation, in a global setting, including (selected) contributions and research results via those local pseudonyms. Although this article concentrates on pseudonymisation, within the PEP-system pseudonymisation is tightly integrated with access control, audit-logging and encryption of the research data. PEP is currently being used in the large-scale PPP study, which will encompass approximately one terabyte of data per participant, for a total of 650 participants. This final section contains some additional remarks on specific points and comes to a conclusion.

Under the GDPR, every 'data subject' has a right of access, that is, a right to see which data a particular organisation holds (on oneself). This is particularly challenging for a research setting with pseudonymisation. The PPP study supports this right of access by telling study participants that they can come and visit and then get access to their data in the repository, via the special data administrator (supervisor). But the PPP does not give participants online access to their data; that would require dedicated client-side software together with a reliable form of authentication that is coupled to polymorphic pseudonyms as used in the PEP system. Such a coupling is a challenge and does not exist at this stage. Another practical reason is that the PEP-system contains mostly raw biomedical data which can only be interpreted by specialists.

Within the GDPR subjects can always withdraw their consent and ask that data on them is removed. This right clashes with the obligation that scientific research must be reproducible. For reasons of archival and scientific research, this is not possible however.

In the PPP project, a compromise is implemented in the PEP system, that after withdrawing consent the existing data is not removed but is not used any more for new studies. In this way, the data can still be used for reproducing results of already published articles (e.g., in case of doubts or even fraud allegations).

As a final remark, a dedicated software team of 4-5 people has been developing the PEP-system for roughly four years (2017-2020). It is now in stable form and other research teams are starting to use PEP as well. The software has become open source in late 2020 in order to provide maximal transparency about how it works[2]. No special hardware or licences are required to run PEP. However, running PEP does require some guidance, which the PEP-team is planning to provide for the coming years.

CONFLICT OF INTEREST

The authors have no conflict of interest to report.

REFERENCES

[1] Agrafiotis I, Bourka A, Drogkaris P (2019) *Pseudonymisation techniques and best practices*. European Union Agency for Cybersecurity, Greece. Available at: https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices.

[2] Bloem BR, Marks WJ Jr, Silva de Lima AL, Kuijf ML, van Laar T, Jacobs BPF, Verbeek MM, Helmich RC, van de Warrenburg BP, Evers LJW, intHout J, van de Zande T, Snyder TM, Kapur R, Meinders MJ (2019) The personalized Parkinson project: Examining disease progression through broad biomarkers in early Parkinson's disease. *BMC Neurol* **19**, 160.

[3] Jacobs B, Popma J (2019) Medical research, big data and the need for privacy by design. *Big Data Soc* **6**, 1-5.

[4] Nissenbaum H (2009) *Privacy in context. technology, policy, and the integrity of social life*. Stanford University Press, Redwood City.

[5] European Commission: Article 29 Data Protection Working Party (2014) *Opinion 05/2014 on anonymisation techniques*. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[6] Pfitzmann A, Hansen M (2008) Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management — a consolidated proposal for terminology. Available at: https://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.31.pdf. Accessed on May 13, 2021.

[7] Sweeney L (2000) *Simple demographics often identify people uniquely*. Carnegie Mellon University, Pittsburgh. Available at: http://dataprivacylab.org/projects/identifiability/paper1.pdf.

[8] Verheul E (2015) Privacy protection in electronic education based on polymorphic pseudonymization. In *IACR Cryptology ePrint Archive* 2015/1228. Available at: https://eprint.iacr.org/2015/1228.2015.

[9] Verheul E, Jacobs B (2017) Polymorphic encryption and pseudonymisation in identity management and medical research. *Nieuw Archief Wiskunde* **5/18**, 168-172.

[2] For details, see https://pep.cs.ru.nl