Towards a Privacy Compliant Research Interface for Multicenter Medical Data

Arno Appenzeller

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
arno.appenzeller@kit.edu

Technical Report IES-2019-06

Abstract

Big Data analysis gains more and more interest in the processing of e-Health data. The potentially big benefit of those analyses comes with a set of new unknown impacts to an individual's privacy. Still it is important to find a balance between privacy impact and utility of the medical data analysis. To achieve this, this technical report takes a look on different privacy preserving techniques, that could be used for a privacy preserving research interface for medical data. The three techniques Differential privacy, k-Anonymity and Secure multi-party Computation are evaluated on their feasibility for a medical use-case. With those preliminaries some formal definitions are made for a privacy preserving research interface which implements an hybrid approach of the three techniques and a consent based interface.

1 Introduction

The digitization in the health care sector is starting to gain more and more traction. As a consequence of the digitization more e-Health data than ever before is accessible for broad use cases. As the amount of data to a given topic is growing, Big Data research usually start to become interested in those topics. Especially for medical data Big Data promises new therapies and new valuable insights on different diseases [12]. A more or less open question from a technical perspective is data protection regarding medical data. From the law perspective, for example with the European General Data Protection Regulation (GDPR), there is a firm opinion on privacy of medical data. However there are many open question when processing large amount of medical data. In general the GDPR categorizes personal health information as special data. Article 9 Paragraph 1 says: "*Processing of [...] data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited*" [3]. At first this means that the processing of personal health data is not allowed. But Article 9 Paragraph 1 a) to j) has exclusions, which allow the processing of this special category of data. One of these exclusions is, if the affected person consents to the usage of their data. Other reasons that allow the processing, like the processing for public interest, are more ambiguous than the explicit consent. While the GDPR asks for an explicit permission for the use of the data from an affected person, even the processing of a large amount of anonymized data does not guarantee privacy. Furthermore a recent study showed that the combination of 15 different attributes per dataset is enough to identify an exact person in the US [10]. This proofs that even if data is only processed in an anonymized way, additional measures have to be taken if an affected individual does not explicitly consent to a certain risk of de-identification.

Another fact we face when working with medical data is that the data environments are often multi centric. This means that the data of a single patient is split across different clinicians or hospitals. As a consequence data from multiple sites need to be coordinated, which means in most cases that a trusted party is needed as a broker for the data. Furthermore the privacy of the data is an important questions when coming from different sources and the data is potentially used in different sites for different purposes. Besides the challenge of a research interface for multi centric health data, there are other challenges like how to merge the data of a single patient from different sites or how the different data providers can be connected securely. However for this technical report we focus on a potential research interface for multicenter medical data. A main requirement for this is the privacy compliant processing of the personal health information. While maintaining this and providing anonymized/pseudonomynized data when needed, another important thing is to provided a back channel for potential results out of the processed data. Especially if they have important results for an individual.

In this technical report we will have an in-depth look at various techniques to provide privacy on personal data in big datasets while still retaining maximum data precision. Afterwards we will present a concept that combines those mechanism with additional techniques that consider consent to provide a privacy preserving research interface for multicenter medical data. In the end this concept will be concluded and an outlook is provided.

2 Related work

Like mentioned in the introduction a recent study by Rocher et al. showed that 15 different attributes are enough to identify 99.8% of the citizen of Massachusetts [10]. The claim is proven with a statistical model. This applies regardless how incomplete the data is, so anonymization will not provide enough benefit to protect an individuals privacy. So even a training set for a machine learning algorithm can be a privacy risk. Because of this conclusion the authors demand for even higher measures, than for example the GDPR demands, to protect the privacy of individuals.

The project "PAPAYA: A Platform for Privacy Preserving Data Analytics" focuses more on the specific issue of a privacy preserving research interface for medical data [2]. Ciceri et al. introduce a project to create privacy-preserving neural networks. The approach uses a combination of encryption, secure multiparty computation, differential privacy and functional encryption. Different data sources are used to train a neural network. The training data is discarded afterwards. All in all they do not provide an in-depth look of their approach. But they present the idea of using differential privacy for the training data to add noise to the original data.

Another project that provides a research interface for medical data is the MOSAIC project [1]. Bialke et al. describe this in "*MOSAIC - A Modular Approach to*

Data Management in Epidemiological Studies". The authors want to comply with privacy requirements by using study-specific pseudonymisation and giving access for third parties only through a designated interface. An interesting fact about MOSAIC is that it enables a designated back channel for research algorithm. With this the algorithm can give back individual findings that occurred during the processing. Unfortunately the concept is not explained in more detail.

3 Privacy preserving techniques for multicenter environments

The following section presents three techniques that can preserve privacy for large databases. Therefore they can be used for multicenter environments. Finally the three techniques will be evaluated by criteria like accuracy and privacy guarantees.

3.1 Differential privacy

In 2006 Dwork et al. introduced the notion of ε-Differential Privacy [6]. In general Differential Privacy has the goal for a certain data in a statistical database to achieve the same level of privacy as if the data is removed or never was in the database. This means that the data of a single individual needs to be modified so that the individual can not be identified. With this approach privacy can be preserved while still retaining a good utility for the processing of the modified data. The assumption for Differential Privacy is, that the likelihood that there is any disclosure, is a very small number regardless if the data is in the database or not. To be more specific the ϵ in ϵ-Differential Privacy describes the privacy loss when a dataset is released from a database. Therefore a really small ϵ is desired but certainly it remains important to keep the utility of the data. Formally K is a ϵ-Differential Privacy algorithm if the following is valid: All available data are part of the set S. D_1 and D_2 are datasets that have the difference of at most one element.

Definition 3.1.1 (ε-Differential Privacy Algorithm)**.**

$$Pr[K(D_1) \in S] \leq e^{\epsilon} * Pr[K(D_2) \in S] \tag{3.1}$$

The conclusion of this definition is that even if data is removed no output and its consequences in regard of privacy loss becomes significantly less or more likely. Which ultimately means that it does not matter if data is in or not in a database, if K fulfills the requirement of Definition 3.1.1.
With this strong privacy guarantees can be achieved but an important factor is the size of the dataset: The smaller the database the higher the noise added (or the smaller ϵ) has to be to alter/randomise the original data.
Another important question is what a good Differential Privacy algorithm is. This question can not be answered in general because it depends on the use case. If the use case is to process numeric values for statistical operations like sum, median or average a good choice is Laplacian noise. This uses the Laplacian mechanism to add noise to the input data. For this algorithm the ϵ is a measure for the randomization. If $\epsilon = 0$ the privatized data is complete random noise. While in theory this provides obviously the best privacy, the data has no more real utility and leads the Differential Privacy approach ad absurdum.
Differential Privacy can be divided in two different variants. The one is Global Differential Privacy where all original data is stored globally. Only the output of this original data is aggregated to fulfill the requirements of Differential Privacy. For this approach a trusted third party which manages the data is essential. The other variant is Local Differential Privacy. Here every individual or data owner modifies the data before it leaves the origin, so that the original information is nowhere else. For this no trusted third party is needed because the data is already modified when it reaches another party. Besides ϵ-Differential Privacy there also exists (ϵ,δ)-Differential Privacy. This version of Differential Privacy accepts deviations by δ from the original notion like in Definition 3.1.1.
Differential Privacy is a concept that sounds very promising in theory. While there are practical use cases (even Apple [5] and Google [8] are using it in their mobile systems) the real utility depends on the scenario it is used. There is a review paper by Dankar et al. which provides an in-depth look at medical applications but still the conclusion is that besides statistical evaluations it is

very limited [4]. However for a combination of different techniques Differential Privacy is one of the most promising ones.

3.2 k-Anonymity

Another technique to preserve privacy is k-Anonymity. The method was introduced in 2002 by Sweeny et al. [11]. The main principle of k-Anonymity is to alter the existing data of a database, so that they still have utility but it is guaranteed that affected individuals with data in the database can not be reidentified. A collection of datasets can be called k-anonymous if one of the datasets ca not be distinguished from $k - 1$ other datasets.

Example 3.2.1 (4-Anonymity)**.** A $k = 4$ anonymized dataset has at least 4 records for each value combination of certain attributes that k-Anonymity applies to.

There are two methods to achieve k-Anonymity:

- **Suppression**: Parts of the data will be removed, disguised or made indistinguishable (Mapping all data to the same pseudonym e.g.).
- **Generalization**: Modify parts of the data to ranges of values instead of exact values or assign attributes to a more general type.

One issue with k-Anonymity is that there is no general measurement for the privacy guarantee. Furthermore additional domain knowledge is required for suppression or generalization of the data. In some cases there are guidelines that could be used for generalization. For example the Canadian Institute for Health Research published the "*CIHR Best Practices for Protecting Privacy in Health Research*" which helps to generalize medical data.
A medical use case for k-Anonymity is described by El Emam et al. [7]. Here the previous mentioned guidelines from the Canadian Institute for Health Research are used as background knowledge for an algorithm that generalizes medical data. With this the generalization can be performed automatically and it is also possible to measure the information loss compared to the original data. So

the privacy impact on a dataset to which the guidelines apply can be reduced. They also show real world feasibility of the approach by using it to hand over k-anonymous data from pharmacies to commercial data brokers. However the the issue of a universal generalization remains and every use case has to be considered individually.

3.3 Secure multi-party computation

The main principles of Secure multi-party Computation (SCM) were already introduced by Yao in the 1980s [13]. The basic idea of this was to evaluate data from different parties without revealing the data.
According to Lindell and Pinkas there can be two models to achieve this [9]. In one case there is a trusted third party that evaluates the data for the participating parties. The other case has no third party one can trust with its data. In this case a direct communication between the data is needed and it needs to be ensured that the data already leaves the participating parties in a private state.
The typical scenario for SMC is that there are several parties that own private data. All parties want to evaluate their data to a common public result. This can also mean that a third party like a research institute gets this data to do the evaluation. The main issue in this scenario is that there is no trust established between the parties or the parties do not want to reveal their data. A special variation of this scenario exists when there is a third party that does the data processing and returns the value to the parties. However for a medical use case it still remains important that the participating parties do not get the raw data but only the final result.
A concrete example for such a scenario is to calculate the average salary of three parties. When using the secret sharing the typical procedure is that the starting party chooses a secret r. This secret is added to the own salary x and the result will be sent to the second party. The second party adds its salary y and sends it to party number three which follows the same procedure. This can be easily extended to an arbitrary number of parties. Finally after the round trip the first party gets the result back and subtracts r to receive the final value to calculate the average without revealing its salary to the others or gaining knowledge of the others salary.

Another approach to this is using homomorphic encryption. In this case certain mathematical operations can be done with the ciphertext without knowing the secret key or the need to decrypt it. The operations depend on the homomorphic properties of the encryption method. For example an additive homomorphic property would mean that it is possible to calculate $Enc(a) + Enc(b) = Enc(a + b)$. It needs to be considered that for plain encryption those methods would have a lot of weaknesses to adversaries, but the measures are enough to preserve data privacy. A possible scenario for this would be a third party research algorithm that does a cohort analysis for a clinician. For this it needs the data from the clinician and other participants that provide the comparison data to create the cohort. A main requirement is that the third party does not see the plain data. To realize this a key broker is required which gives a common key to all participants. With the resulting chiphertexts the third party algorithm can do its cohort analysis using the homomorphic properties.

An obvious advantage to the previous techniques is correctness of the result which also implies precision. That means while the results achieved with Differential Privacy or k-Anonymity can differ to a certain degree from the real result, SMC always returns the exact result. An issue with SMC is that it has a big overhead in terms of run time. Even simple operations can use a lot of time.

3.4 Evaluation of the techniques

After the introduction of the three different techniques considered in this report, we will do an evaluation of them that considers the strengths and weaknesses of the techniques. Table 3.1 gives an overview of this.

In terms of privacy guarantees both Differential Privacy and k-Anonymity have metrics that make a statement about the degree of privacy. SMC's guarantees are dependent on the encryption mechanism used and can not be generalized. Full accuracy is provided when using SMC while the privacy preserving mechanism does not rely on modification of the data. Differential Privacy's accuracy is affected by the choice of ϵ, where a very large ϵ provides good accuracy but not much privacy. For k-Anonymity no general assumption can be made because the accuracy depends on the generalisation/surpression method. When considering scalable performance Differential Privacy as well as k-Anonymity should provide good results regardless the amount of data while SMC has a lot of overhead because of the encryption mechanism. Lastly it is important if any kind of trusted party is needed to perform the techniques. Differential Privacy and k-Anonymity require a party the manages the data. If considering Differential Privacy it is possible that the local approach is used so the trusted party is only needed for the global approach. Only SMC offers the option to operate completely without a trusted party, if the participants communicate directly with their ciphertext.

Table 3.1: Overview of privacy preserving techniques

		Techniques	
	Differential Privacy	**Secure multi-party Computation**	**k-Anonymity**
Privacy guarantuees	●		●
High Accuracy	○	●	
Scaleable performance	●		●
Trusted Party needed	Partly	No	Yes
Limitations	Choice of ϵ affects properties	Utility and processing time heavily depends on the type of SMC	Requires domain knowledge

4 A privacy compliant research interface

To define a research interface it is important to understand the difference between a non-interactive interface and an interactive one. A non-interactive research interface is one where the data is released once and for all and there is no way to modify the data for a certain request. An interactive research interface can decide the privacy strategy for each query since only the data for the given request is released and the complete data remains hidden through the interface.

We think that for a privacy preserving research interface it is important not to follow an one fits all approach. There are different kind of queries that can require different degrees of accuracy. The main goal should always be: *Preserve as much privacy as possible and lose as less accuracy as possible.* This can only be achieved with a hybrid approach. On the one hand a combination of the previously introduced techniques, that are used for the range of queries where the individual technique, can be used best. On the other hand those techniques all fall in specific use cases and can reach their limit, where no more useful query is possible. Furthermore there can be some requests where both the researcher and the affected person can benefit from data that is not anonymized. You can think of queries that can provide feedback on the individual person. For those queries the person's consent is mandatory.

To include this in the desired fully automated research interface a mechanism is required to map the consent in a digital format. Furthermore this consent should be dynamic so that an affected person can authorize or revoke it at any time. In addition to enable automatic evaluation of this, an enforcement mechanism is needed to evaluate consent for each query. Medical consent in a digital format is a non trivial task with some existing concepts but most of them are far from complete. We will postpone this part which we call *consent based interface* to future work.
We assume that he research interface exposes a set of privacy functions like $\mathbf{P\_SUM}$, $\mathbf{P\_AVERAGE}$, $\mathbf{P\_MEDIAN}$ etc. to do operations on attributes of the data in the database.

Definition 4.0.1 (Privacy preserving functions)**.** A privacy preserving research interface defines a Set $\mathcal{F}$ of privacy preserving function. They all follow the following naming convention $\mathbf{P\_}*$ where $*$ is a mathematical function like $\mathbf{SUM}$ or $\mathbf{COUNT}$.

To perform a query the researcher has to provide additional properties. It needs to be defined if $accuracy$ or $privacy$ to which scale is desired or if an algorithm wishes to provide additional feedback to an individual $feedback$.

Definition 4.0.2 (Privacy preserving configuration)**.** A privacy preserving research interface has a Set $\mathcal{C} = \{accuracy, privacy(x), feedback\}$ which contains the privacy preserving configuration for a request. $privacy(x)$ has $x \in \mathbb{N}^+$ as number to indicate the factor of the privacy impact.

With this a request can be formulated. Such a request uses a query language in an interface specific language where the request attributes from the dataset can be defined. A privacy function out of $\mathcal{F}$ also needs to be used in this query. In addition a configuration needs to be provided to indicate what the requirements for the request are.

Definition 4.0.3 (Privacy preserving request)**.** A request req for a privacy preserving research interface looks like the following: $req = (query, config)$ where $query$ is a query made with a query language $\mathbf{QL}$ that includes $\mathcal{F}$ and $config \in \mathcal{C}$.

With such a request req the interface can now decide depending on $config$ which privacy preserving technique should be used. The following Definition 4.0.4 illustrates this.

Definition 4.0.4 (Evaluation of $config$)**.**

$$config = \begin{cases} \text{if } accuracy \rightarrow \text{use SMC} \\ \text{if } privacy(x) \rightarrow \text{use Differential Privacy} \\ \hookrightarrow \text{or } k\text{-Anonymity depending on } x \\ \text{if } feedback \rightarrow \text{use } \textit{consent based interface} \end{cases}$$

5 Conclusion & outlook

This technical report looks at three different techniques to preserve privacy on an individuals data. All of these three techniques have various advantages and disadvantages. While Differential Privacy and k-Anonymity have good privacy guarantees they can lack accuracy. SMC can provide accuracy on the results but its performance can be a great uncertainty. So there is certainly no one fits all approach. In fact a hybrid approach that combines those three techniques and that chooses the best depending on the requirements for a certain request is proposed. In addition there can be requests where those techniques can not help or do not fit the requirement. Therefore a fallback to the individuals consent is needed. With this definition of a privacy preserving research interface for multicenter medical data the foundation for more in-depth work and experiments with real world e-Health data is made.

While this report provides the fundamentals a real world evaluation needs to be done. It needs be proven that the introduced privacy preserving techniques work good on real medical data. Another issue that remains is a good privacy metric. This is especially required for an informed consent decision of a patient. Considering that the consent based interface needs to be introduced in future work. With this integration a full feature research interface is possible, which remains open for further refinement. Finally this approach should be evaluated against the GDPR. It has to be figured out what is needed to be compliant to it and what an interface should provide to fulfill requirements of the GDPR.

References

[1] M. Bialke et al. “MOSAIC – A Modular Approach to Data Management in Epidemiological Studies”. In: *Methods of Information in Medicine* 54.04 (2015), pp. 364–371.

[2] Eleonora Ciceri. “PAPAYA: A Platform for Privacy Preserving Data Analytics”. In: *ERCIM News* 118 (2019).

[3] European Commission. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. URL: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

[4] Fida K. Dankar and Khaled El Emam. “Practicing Differential Privacy in Health Care: A Review”. In: *Trans. Data Privacy* 6.1 (2013), pp. 35–67.

[5] Apple Differential Privacy Team. “Learning with Privacy at Scale”. In: *Apple Machine Learning Journal* 1.8 (2017).

[6] Cynthia Dwork. “Differential Privacy”. In: (2006), pp. 1–12.

[7] K El Emam et al. “A globally optimal k-anonymity method for the de-identification of health data.” In: *J Am Med Inform Assoc* 16.5 (2009), pp. 670–682.

[8] M. Guevara. *Enabling developers and organizations to use differential privacy*. 2019. URL: `https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html` (visited on 10/25/2019).

[9] Yehuda Lindell and Benny Pinkas. “Privacy Preserving Data Mining”. In: *Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology*. CRYPTO ’00. London, UK, UK: Springer-Verlag, 2000, pp. 36–54. ISBN: 3-540-67907-3. URL: `http://dl.acm.org/citation.cfm?id=646765.704129`.

[10] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. “Estimating the success of re-identifications in incomplete datasets using generative models”. In: *Nature Communications* 10.1 (2019).

[11] Latanya Sweeney. “K-anonymity: A Model for Protecting Privacy”. In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10.5 (Oct. 2002), pp. 557–570. ISSN: 0218-4885. DOI: `10.1142/S0218488502001648`. URL: `http://dx.doi.org/10.1142/S0218488502001648`.

[12] K. Verspoor and F. Martin-Sanchez. “Big Data in Medicine Is Driving Big Changes”. In: *Yearbook of Medical Informatics* 23.01 (2014), pp. 14–20.

[13] A. C. Yao. “Protocols for secure computations”. In: *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science*. IEEE Computer Society, 1982, pp. 160–164.