A Scalable and Pragmatic Method for the Safe Sharing of High-Quality Health Data

Fabian Prasser, Florian Kohlmayer, Helmut Spengler, and Klaus A. Kuhn

***Abstract*—The sharing of sensitive personal health data is an important aspect of biomedical research. Methods of data de-identification are often used in this process to trade the granularity of data off against privacy risks. However, traditional approaches, such as HIPAA safe harbor or k-anonymization, often fail to provide data with sufficient quality. Alternatively, data can be de-identified only to a degree which still allows us to use it as required, e.g., to carry out specific analyses. Controlled environments, which restrict the ways recipients can interact with the data, can then be used to cope with residual risks. The contributions of this article are twofold. First, we present a method for implementing controlled data sharing environments and analyze its privacy properties. Second, we present a de-identification method which is specifically suited for sanitizing health data which is to be shared in such environments. Traditional de-identification methods control the uniqueness of records in a dataset. The basic idea of our approach is to reduce the probability that a record in a dataset has characteristics which are unique within the underlying population. As the characteristics of the population are typically not known, we have implemented a pragmatic solution in which properties of the population are modeled with statistical methods. We have further developed an accompanying process for evaluating and validating the degree of protection provided. The results of an extensive experimental evaluation show that our approach enables the safe sharing of high-quality data and that it is highly scalable.**

I. Introduction

RECENT developments in biomedical research, e.g., the trend towards precision medicine, require the sharing of sensitive personal data of high quality [1], [2]. Moreover, the sharing of research data is increasingly required by public sponsors [3]. However, the number of health data breaches is growing [4] and there is significant public pressure to ensure that the privacy of patients and probands is protected [5].

Biomedical data is typically protected using methods of *data de-identification*. The basic idea is to transform datasets in such a way that malicious parties are not able to determine the identity of data subjects based on the information contained. This threat to data privacy is called *re-identification* [6]. A typical method for re-identifying records from an insufficiently de-identified dataset are *linkage attacks*. In this process, the adversary correlates the dataset with an additional dataset containing identifying information, e.g., a voter registration list [6].

The broad spectrum of available technologies for countering such attacks and the differences between legal frameworks lead to trade-offs. Two factors have to be considered: reduction of *1)* privacy risks, and of *2)* data quality [7]. The challenge of achieving a reasonable balance between these conflicting objectives has been studied in different scientific fields and a wide variety of solutions have been proposed.

II. Related Work

The *Safe Harbor* [8] method of the U.S. Health Insurance Portability and Accountability Act (HIPAA) [9] specifies 18 rules which describe the removal or alteration of attribute values that are associated with a high risk of re-identification (e.g., names and dates). HIPAA distinguishes between *de-identified datasets* and *limited datasets*. The former type of data can be created by applying all sanitization rules and the result is not considered personally-identifiable under the law anymore. The latter type of data can be created by using only a subset of the rules, resulting in data of higher quality. However, the data is still considered identifiable and it must only be shared when additional safeguards have been implemented.

In other countries, e.g., from the European Union where data privacy is regulated by the European Directive on Data Protection [10], no explicit rules for de-identifying health data have been defined. Here, computational methods to data de-identification, such as *k-anonymity*, are more important. The method requires that data has been transformed in such a way that each record is part of a group of at least k records that are indistinguishable regarding potentially identifying attributes, so called *quasi-identifiers* [11].

Xia *et al.* have performed an experimental comparison of datasets which have been de-identified with HIPAA Safe Harbor and datasets which have been de-identified with computational methods. Their results show that computational methods, including k-anonymity, can produce de-identified data that – at the same time – provide a higher degree of quality and a lower risk of re-identification [7]. Various algorithms which implement this model for biomedical data have been proposed [12], [13]. However, it has also been shown that methods like k-anonymity do not reduce privacy risks to completely zero and that data quality may still be too low for common usage scenarios [14].

Manuscript received August 18, 2016; revised December 14, 2016 and January 30, 2017; accepted February 23, 2017. Date of publication March 23, 2017; date of current version March 5, 2018. *(Corresponding author: Fabian Prasser.)*

The authors are with the Department of Medicine, Technical University of Munich, Munich 81675, Germany (e-mail: fabian.prasser@tum.de; florian.kohlmayer@tum.de; helmut.spengler@tum.de; klaus.kuhn@tum.de).

Digital Object Identifier 10.1109/JBHI.2017.2676880

Data quality is severely impacted by the fact that strict syntactic privacy models, including k-anonymity, do not tolerate unique records in a dataset, which imposes significant structural limitations [7]. More elaborate approaches, see e.g., [15], prevent records from a dataset from having characteristics which are unique within the underlying population, instead of restricting the uniqueness of records within the dataset itself. However, these methods require that the dataset is explicitly represented as a subset of a population table which contains records describing the population. As such tables are typically not available, this is not a realistic assumption. Methods which estimate the characteristics of the underlying population with statistical methods have been proposed for protecting datasets from so-called marketer attacks in which an adversary aims to re-identify a large number of records [16], [17]. However, in contrast to strict privacy models, these methods cannot be used to protect data from targeted attacks by adversaries with detailed background knowledge.

Alternatively, data can be altered in such a way that privacy risks are reduced as far as possible while it is guaranteed that data quality is sufficient for the intended use case [18]. Principles-based rules-of-thumb which have been formulated by experts rather than strict methods of data sanitization are typically used for this purpose [19]. In the domain of official statistics this is called *a posteriori* disclosure risk control. Controlled environments, which restrict the ways recipients can interact with the data, can then be used as a secondary measure to cope with residual risks. This approach combines two methods for protecting privacy: *restricted data* and *restricted access* (also: *safe data* and *safe setting*) [20].

Cryptographic solutions for privacy-preserving data analysis are on one end of the spectrum of restricted access settings. They aim to provide complete confidentiality of input data during analyses. For example, homomorphic encryption schemes have been developed for predictive analyses on encrypted health data [21] and for genomic studies [22]. A significant drawback of such solutions is that they are inflexible, as they need to be designed specifically for each and every application scenario. Moreover, cryptographic solutions are often computationally complex. To also provide output privacy, such methods can be combined with differential privacy, which defines data protection as a property of the data processing method and not as a property of the output dataset [23]. In the biomedical context the model has, e.g., been applied in genetic research for computing significant single-nucleotide polymorphism (SNPs) in genome-wide association studies [24]. However, the approach must also be tailored towards specific processing methods and usage scenarios. Differential privacy is typically achieved via noise addition and it has been argued that the method is not well suited for protecting the type of structured health data considered in this article due to its pertubative nature [25].

Virtual data access environments are on the other end of the spectrum of restricted access settings. The basic idea is that analysts access data processing facilities (e.g., statistics software) via secure remote desktop connections. As no data is actually shared in such environments, it becomes very difficult to extract sensitive information. At the same time, the approach is flexible and it allows data recipients to perform explorative data analysis and hypothesis generation for a wide variety of scientific questions. Virtual access environments are often used to share sensitive data in official statistics. The approach has been recommended by the U.S. National Research Council [26] and it is implemented by the Virtual Microdata Laboratory of the U.K. Office for National Statistics [27].

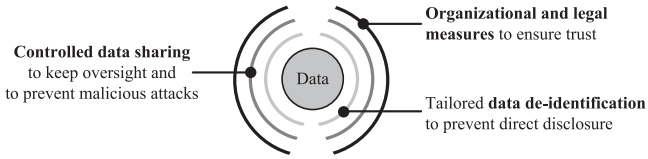

Fig. 1. Combining organizational, legal and technical measures to achieve secure data sharing following the "onion-skin principle".

III. Objectives and Outline

Due to the highly sensitive nature of health data, a combination of organizational, legal and technical measures must be employed to achieve an adequate level of protection [28] and to maintain public acceptance [29]. As is sketched in Fig. 1, we propose a method which follows this approach.

Based on sound organizational processes which ensure the trustworthiness of recipients, our method combines controlled data sharing with appropriate data de-identification to protect datasets from privacy breaches while providing a high degree of data quality and flexibility. We had to overcome multiple challenges in the process of developing this solution.

Firstly, detailed requirements and potential design options for controlled data sharing environments in biomedical research have not been described, yet. We have therefore designed such an environment and analyzed potential attack vectors of the resulting architecture using a well-defined methodology. From the results of this analysis, we have derived important countermeasures which need to be implemented to mitigate the identified threats.

Secondly, we needed a de-identification method which integrates well with controlled data sharing by countering the residual risks of restricted access environments while providing high quality data and scalability. We have developed a method which fulfils these requirements. Our approach focuses on restricting the uniqueness of data characteristics within the underlying population. It does not require the provision of a population table, but it uses statistical methods to estimate properties of the population. We show that this pragmatic solution provides high-quality output data with interpretable privacy guarantees, which offer a sufficient level of protection in our setup.

The remainder of this paper is structured as follows: We briefly describe the trust model underlying our approach and important organizational safeguards in Section IV. Then, in Section V, we present the design of a controlled data sharing environment, perform a structured threat analysis and propose an implementation involving appropriate mitigation strategies. We present a tailor-made de-identification method and

describe how we have developed a scalable implementation in Section VI. Next, in Section VII, we present the results of an experimental evaluation and in Section VIII we describe how residual risks can be analyzed prior to data sharing. Finally, we discuss our method in Section IX, cover future work in Section X and conclude this paper in Section XI.

IV. Trust Model

Our approach assumes that the data custodian has implemented adequate processes and organizational structures to ensure that the recipients of data are trustworthy and that they aim at answering reasonable research questions which require data sharing. A typical approach is to install *data access committees* and to make the recipients sign *data use agreements*, which are contracts that hold them responsible to comply with relevant terms, conditions and regulations [30]. Specifically, recipients will have to take adequate measures to protect the shared data and to make sure that they are fully considering the original informed consent, ethics committee approvals and data access committee decisions. In our scenario it is also important to employ sound and reliable processes for authenticating and authorizing data recipients.

Although ensuring the trustworthiness of recipients is an important building block of privacy protection, it is not sufficient for complying with laws and regulations in many jurisdictions. For example, the European Directive on Data Protection [10] and the European General Data Protection Regulation (GDPR) [31] both require that *"personal data shall be [...] kept in a form which permits identification of data subjects for no longer than is necessary"* [10]. The approach described in this paper relies on further technical measures to achieve this objective.

V. Controlled Data Sharing

A. Basic Design and Threat Analysis

Controlled data sharing environments enable the data custodian to stay in control of a dataset. As the data receiver is only able to access the data but does not receive the dataset itself, it becomes possible to revoke access and to protect data from potential future attacks. This is also required by the European GDPR [31]. Moreover, the data custodian maintains oversight of the sharing process and it becomes less critical to ensure that the recipient handles the data with adequate care.

Without loss of generality, we assume that the researchers are provided with access to a modern web-based data analytics platform. For biomedical data systems such as i2b2 [32] or tranSMART [33] are typical examples. The controlled data access environment is created by installing a proxy between the client and the server hosting the analytics software. We will describe this proxy and analyze potential vectors for re-identification attacks using a *Data-Flow Diagram* (DFD) and a *Threat Tree* (TT). Both concepts are central components of the *LINDDUN* methodology, which is a framework for analyzing and countering privacy threats [34].

Fig. 2 shows a data-flow diagram for the basic design of the proxy. The user accesses the proxy, which is a *Process* in

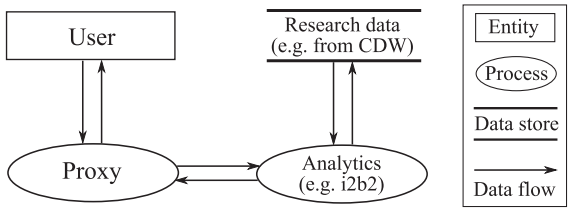

Fig. 2. Data-flow diagram of a controlled data sharing environment.

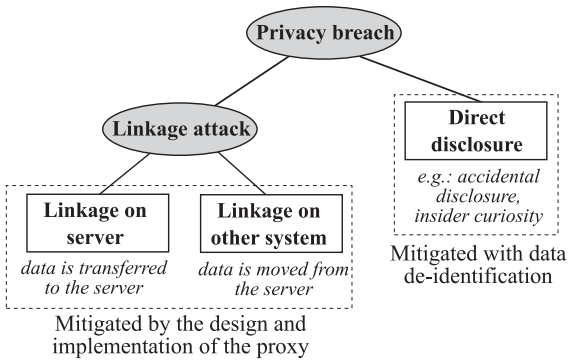

Fig. 3. Threat tree for restricted data access.

LINDDUN's terminology. The proxy redirects requests to the interface of the analytics solution, which runs on top of a data store containing the sensitive information. For the sake of brevity, we will focus on privacy and assume that relevant measures of information security have been put in place. For example, we assume that all systems holding sensitive data have been hardened properly, that communication between services is always encrypted and that every user is properly authenticated and authorized (see above). The remaining objective is to prevent the user from breaching privacy.

Fig. 3 presents a threat tree describing the three attack vectors which we have identified. Techniques for re-identifying data require that the adversary correlates the records of a dataset with additional information. Moreover, even the most robust non-targeted attacks, see e.g., [35], require the adversary to scan a dataset for vulnerable records. This can either be performed *remotely*, by uploading the dataset containing identifying information to the data analytics platform, or *locally*, by downloading the sensitive research data to a system which is under the adversary's control. *Direct disclosure* happens when the user simply recognizes an individual from a given record. Typical examples of this threat are *accidental disclosure* or *insider curiosity*, which often happen without the researcher actively and deliberately performing an attack [36].

B. Implementation and Mitigation Strategies

The main objective of controlled data sharing is to prevent the recipient from performing linkage attacks. For this purpose, attack vectors for remote and local linkage must be countered and several side-channels must be closed.

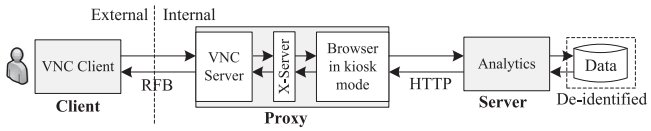

Fig. 4. Implementation of a controlled data sharing environment. Information security aspects, e.g., encrypted communication, have been omitted for brevity.

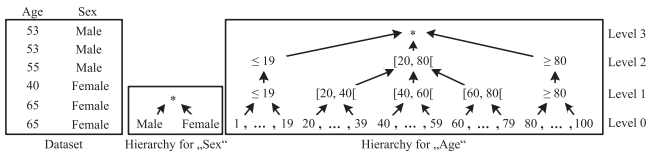

Fig. 5. Example dataset and generalization hierarchies.

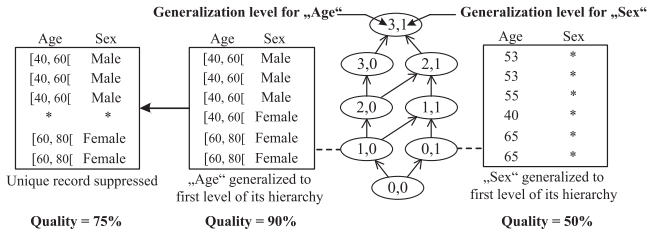

Fig. 6. Generalization lattice, de-identification policies and output datasets.

As is shown in Fig. 4, the proxy provides a way of controlling the interactions between the analyst and the analysis software by exposing its frontend via Virtual Network Computing (VNC) using the Remote Framebuffer (RFB) protocol [37]. The VNC server acts as terminal, which provides only the contents of its framebuffer to the VNC client. This means that to the data recipient all information displayed by the analytics software is only available as unstructured graphical data in form of a (compressed) stream of images. This makes it very hard to extract large amounts of information automatically, even with state-of-the-art OCR engines, see e.g., [38].

The VNC client accepts and transfers user input, i.e., mouse and keyboard events, to the VNC server. The VNC server redirects these events to an X-Server, which also provides the framebuffer that is exposed by the VNC server. As a bridge between the proxy and the analytics software, the X-Server is configured to execute exactly one predefined application: a web browser running in kiosk mode which is configured to show the interface of the analytics platform only.

Special care is needed to close a wide variety of potential side channels. Firstly, it must be ensured that the user is securely locked into the browser, that no other programs can be executed by the user, that the browser can only communicate with the server, and that the analytics software cannot be used to access external data. All of this requires special configuration. Secondly, the VNC server must be configured to not support copy-and-paste operations between the client and the server to prevent the user from transferring structured data. Finally, in order to prevent the user from uploading significant amounts of data via simulated mouse and keyboard interactions, rate limiting must be put in place for these operations. As a cross-sectional mitigation strategy, the user's interactions should be logged. We recommend to log all keyboard and mouse interactions, to record a video of the screen content during interactive sessions and to capture and store all network traffic.

The controlled data sharing environment prevents data recipients from performing linkage attacks. In order to also prevent direct disclosure, it must be ensured that an adequate degree of uncertainty is introduced regarding the identity of data subjects. For this purpose, we have developed a tailor-made data de-identification method.

VI. Data De-Identification

A. Background

Computational methods to health data de-identification typically use *generalization hierarchies* as a backbone for data transformation [7], [12], [13]. An example dataset and two hierarchies are shown in Fig. 5. Here, values of the attribute "age" are transformed into intervals with decreasing precision on increasing *levels* of generalization. Such hierarchies can be constructed for categorical and continuous variables, in the latter case by specifying functions for performing on-the-fly categorization.

The solution space consists of all possible combinations of generalization levels for each attribute. As is shown in Fig. 6, it forms a *generalization lattice*, in which two elements have a successor-predecessor relationship if their generalization degrees differ by exactly one level [13]. We call each element a *de-identification policy*, which is applied to the dataset by generalizing it to the specified degrees [7].

Generalization makes the records of a dataset less distinguishable, which reduces privacy risks. To prevent that a large amount of generalization must be applied to the overall dataset, further methods of data transformation, such as microaggregation or record suppression, are typically used. Microaggregation means that values are made indistinguishable by using aggregate functions, such as the arithmetic mean, to produce a common value. *Record suppression* means that outliers are removed from the dataset. Both methods can increase the quality of output data significantly [39].

Different de-identification policies result in datasets with different degrees of quality and different risks of re-identification [40]. Computational methods for health data de-identification utilize mathematical models for measuring both aspects. The overall objectives, i.e., minimizing re-identification risks while maximizing data quality, are conflicting. This contradiction is resolved by specifying a disclosure risk threshold for the privacy model. This reduces the de-identification process to a simpler optimization problem in which the objective is to make sure that risk thresholds are met while data quality is maximized [13]. This is also called *a priori* disclosure risk control.

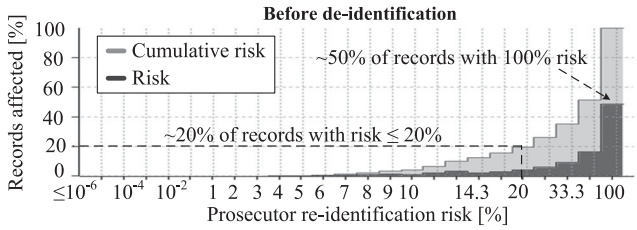

Fig. 7. Risk profile of the Health Interview Series dataset.

B. Measuring Data Quality

While our implementation supports a wide variety of methods for measuring data quality, including the models presented in [41]– [45], we focus on two particularly important approaches in the remainder of this article. Firstly, we will use *Loss* by Iyengar [46], because it is an intuitive and precise measure for the granularity of the resulting data. Fig. 6 shows the data quality of three different output datasets according to this model. The original dataset has a quality of 100%, while a transformed dataset in which all attribute values have been removed (either by generalization or suppression) has a quality of 0%. Secondly, we will use *Non-Uniform Entropy* by De Waal and Willenborg [47], because this model has been recommended for the de-identification of health data [48]. It measures the differences in the distributions of attribute values introduced by transformations.

C. Measuring Re-Identification Risks

The k-anonymity privacy model enforces an upper bound on the re-identification risk of each record in a dataset. Thereby, the risk of a record is estimated with its degree of *uniqueness*, i.e., the number of records k having similar characteristics [6]. The higher this number, the lower the uniqueness and, hence, the lower the risk. This scheme protects data from worst-case scenarios, called *prosecutor attacks*, where the adversary already knows that a record about the targeted individual is present [49], [50]. The basic idea is that the attacker will only be able to associate the individual with a set of records of size of at least k, which reduces the probability of correct linkage of the individual to its record to not more than $\frac{1}{k}$.

The same principles can be used to construct *risk profiles* which summarize the risks of all records in a dataset. An example is shown in Fig. 7 (details about the dataset will be given in Section VII-A). The parameter *risk* associates each possible prosecutor risk with the relative number of records which are affected by the respective value. About 50% of the records in the dataset are affected by a risk of 100%, which means that their characteristics are unique. The parameter *cumulative risk* associates each possible risk with the relative number of records which are affected by a risk that is not higher than the respective value. Only about 20% of the dataset's records are affected by a risk of not more than 20%, which means that they fulfill 5-anonymity, i.e., they cannot be distinguished from at least five other records.

D. Super-Population Models

While risk profiles are very helpful for developing an initial understanding of the privacy risks associated with a given dataset, prosecutor risks are a very conservative estimate of the actual risk [49]. The main reason is that each dataset represents information about a *sample* of individuals from an underlying *population*. Without exact knowledge of the population, a reliable re-identification of data subjects is complicated in practice. Firstly, it must be determined whether or not data about an individual is actually contained in a dataset. Next, the individual must be linked with the corresponding record. For both process steps, the probability of success depends on the uniqueness of the individual's characteristics in the dataset *as well as* in the underlying population [49].

Unfortunately, the exact characteristics of the population are typically not known. As a pragmatic solution, *super-population models* can be used to estimate *population uniqueness*, i.e., the degree to which combinations of specific characteristics are unique within the underlying population. Thereby, the uniqueness of characteristics is represented by the frequency with which groups of indistinguishable individuals of certain sizes exist. This is very similar to the risk profiles presented in the previous section, but absolute numbers are used instead of relative frequencies and risks. Super-population approaches use flexible distributions to model the profile of the population, which are then parameterized with the profile extracted from the sample.

Dankar *et al.* have used this approach to create and validate a risk model for clinical datasets, which estimates the number of records from a dataset which have characteristics that are unique within the underlying population [16]. The model by Dankar *et al.* chooses one of three different approaches, i.e., the model by *Chen and McNulty* [51], the model by *Hoshino* [52] or the model by *Zayatz* [53], depending on the size of the population.

Our novel de-identification method uses the model by Dankar *et al.* for measuring risks. In contrast to most previous approaches, the model tolerates unique records within a dataset and our approach can therefore be used to produce de-identified data with significantly improved quality (see Section VII-A). Moreover, we argue that the model provides a sufficient degree of privacy protection in our scenario. In order to prevent accidental disclosure and insider curiosity we need to introduce some degree of uncertainty about whether or not a record corresponds to a specific individual. Such uncertainty already exists for all records which are not unique within a de-identified dataset. When the number of *population uniques* (PU), i.e., records from a dataset with characteristics that are unique within the population, is controlled, uncertainty is also introduced for records that are *sample uniques* (SU), i.e., which are unique within the dataset. Only a subset of the unique records in a dataset will also be population uniques. The probability that a unique record in the sample corresponds to a given individual from the population is thus $\frac{|PU|}{|SU|}$. Assuming that $|PU| \ll |SU|$, a sufficient degree of protection is provided.

We will focus on the model by Hoshino in the remainder of this article, because it requires specific optimizations that are relevant for implementing all three models.

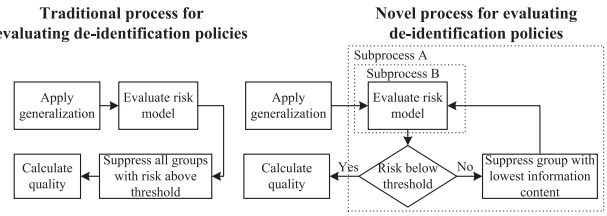

Fig. 8. Traditional and novel method for evaluating de-identification policies.

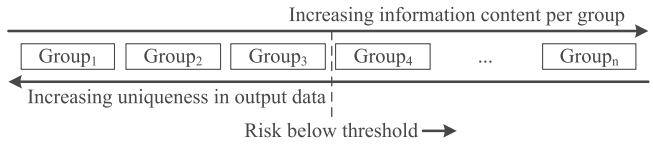

Fig. 9. Ordering groups to enable fast record suppression.

Moreover, it has been recommended by Dankar *et al.* for common data sharing scenarios [16]. Hoshino's approach uses Pitman's sampling formula as the base distribution for characterizing the population [52].

We will denote the number of records in the dataset that is to be de-identified with n and the size of the population from which it was extracted with N. Moreover, we will denote the number of groups of records in the dataset with u. Each group consists of a set of records with identical characteristics. We will denote the number of groups containing i records with s_i. Population uniqueness is estimated as follows:

$$PU = \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)} N^{\alpha}. \tag{1}$$

Here, Γ denotes the gamma function, which is an extension of the factorial function to real numbers, and θ as well as α are roots of the following bivariate non-linear equation system:

$$f_1(\alpha,\theta) = \sum_{i=1}^{u-1} \frac{1}{\theta+i\alpha} - \sum_{i=1}^{n-1} \frac{1}{\theta+i} = 0 \tag{2}$$

$$f_2(\alpha,\theta) = \sum_{i=1}^{u-1} \frac{i}{\theta+i\alpha} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{j-\alpha} = 0. \tag{3}$$

A common method for numerically solving such equation systems is the Newton-Raphson algorithm, which has also been recommended by Hoshino [52] and by Dankar *et al.* [16].

E. Basic Algorithm

Using the risk model presented in the previous sections for data de-identification requires a process that is very different from "traditional" de-identification algorithms. A comparison is shown in Fig. 8. When making sure that a dataset fulfills k-anonymity, it is sufficient to simply suppress all groups containing less than k records. In contrast, a holistic view of the whole output dataset is needed to measure risks according to the models considered in this work.

Consequently, the algorithm proposed in this article is much more complex. When evaluating a given de-identification policy, the defined generalization scheme is first applied to the dataset. As a result of this process the dataset is transformed into a set of groups of indistinguishable records. Next, a risk profile is computed and used as input for the super-population model. If the number of population uniques is not below the given threshold, the group of records with the lowest information content is suppressed. This group is determined using the same model for data quality which is used to select the overall solution. Next, the privacy model is evaluated again, this time for the modified dataset. This process is repeated until thresholds are met. Finally, the algorithm computes the quality of the output dataset and proceeds with the next policy. When all solution candidates have been processed, it returns the policy which results in output with maximal quality.

Several optimizations are needed to make this process feasible in real-world settings. Super-population models are complex and computationally expensive to evaluate. Multiple policies need to be assessed to determine a high-quality solution for a given input dataset. Assessing a single policy requires to calculate risks multiple times.

F. Reducing the Number of Candidate Policies

The first optimization reduces the number of candidate policies by implementing a generic pruning strategy based on the data quality model. The general idea is to consider the fact that data is transformed with attribute generalization and record suppression by measuring the reduction in data quality caused by both types of transformations independently.

The reduction in quality induced by only generalizing the data increases monotonically with increasing level of generalization. This can be used to exclude all direct and indirect successors of policies for which the quality of the dataset produced by only using generalization is already lower than the quality of the best solution that is currently known [40].

Let us consider the example from Fig. 6. Assuming that the dataset with a quality of 75% is a known solution to the de-identification problem, all successors of the policy $(0, 1)$ can be excluded, because they can never result in datasets with a quality of more than 50%.

G. Reducing the Number of Risk Calculations

For each candidate policy, our method first applies the defined generalization scheme and then iteratively suppresses the group of records with the lowest information content until risks fall below the given threshold (*Subprocess A*).

A method for speeding up this expensive process is sketched in Fig. 9. Firstly, the groups are ordered according to their information content from left to right. Next, it is assumed that each group $Group_i$ is labeled with the risk which would be associated with the dataset if all records from groups $Group_1$, $Group_2$, ..., $Group_i$ have been suppressed. Re-identification risks can only decrease when groups of records are removed from the dataset (cf. [30], [51]–[53]). Therefore, risk increases in the opposite direction of data quality, i.e., from right to left.

As a consequence, there is no need to actually calculate risks for all possible sets of groups that can be suppressed. Instead, we can exploit the fact that re-identification risk defines a total order on the groups by searching for the closest point at which it is lower than the given threshold using a binary search. During this process we can dynamically evaluate risks only for each cut-off point considered by the search algorithm. Suppressing all groups from $Group_1$ to the cut-off point will also maximize data quality.

H. Reducing the Complexity of Risk Calculations

Calculating risks (*Subprocess B*) requires solving a bi-variate non-linear equation system with the Newton-Raphson method. This iterative solver starts with an initial guess and then works its way towards the roots by repeatedly evaluating the object functions as well as their four partial derivatives.

We first modified the solver to evaluate the object functions as well as the derivatives within a single method call. This allowed us to decompose the functions into commonly used blocks which could then be fused and reused in different calculations. For example, let us consider the following two functions:

$$p_1(\alpha,\theta)=\sum_{i=1}^{u-1}\frac{1}{\theta+i\alpha} \qquad p_2(\alpha,\theta)=\sum_{i=1}^{u-1}\frac{i}{\theta+i\alpha}.$$

Here, p_1 is the first summand of the first object function f_1 and p_2 is the first summand of the second object function f_2 (see (2) and (3)). It is easy to see that these functions can be evaluated in a single loop (loop fusion), where the summands of p_2 can be obtained by multiplying the summands from p_1 with i.

However, evaluating the functions still required several iterations, e.g., over the range $[1,u]$, which is expensive. As both functions are finite sums of rational functions, they can be substituted with the digamma function ψ, which is the logarithmic derivative of the gamma function. The following recurrence formula is a well-known example [54]:

$$\sum_{i=0}^{N-1}\frac{1}{x+i}=\psi(N+x)-\psi(x) \tag{4}$$

It follows that p_1 can be written as:

$$\begin{aligned}
p_1(\alpha,\theta) &= \sum_{i=1}^{u-1}\frac{1}{\theta+i\alpha}\overset{(1)}{=}\frac{1}{\alpha}\sum_{i=1}^{u-1}\frac{1}{\frac{\theta}{\alpha}+i}= \\
&\overset{(2)}{=}\frac{1}{\alpha}\left(\sum_{i=0}^{u-1}\frac{1}{\frac{\theta}{\alpha}+i}-\sum_{i=0}^{0}\frac{1}{\frac{\theta}{\alpha}+i}\right)= \\
&\overset{(3)}{=}\frac{1}{\alpha}\left(\psi\left(u+\frac{\theta}{\alpha}\right)-\psi\left(\frac{\theta}{\alpha}\right)\right. \\
&\quad\left.-\psi\left(1+\frac{\theta}{\alpha}\right)+\psi\left(\frac{\theta}{\alpha}\right)\right)= \\
&= \frac{1}{\alpha}\left(\psi\left(u+\frac{\theta}{\alpha}\right)-\psi\left(1+\frac{\theta}{\alpha}\right)\right).
\end{aligned}$$

As is indicated in the equation, there are three important steps. Firstly, we make sure that the index variable i does not have a factor. This is achieved by factoring out α^{-1} from the sum. Secondly, we shift the index variable to start from 0 instead of 1. Finally, we use the equality described in Equation 4 to insert the digamma function ψ and simplify the formula.

Using the same process, we can obtain a closed form of p_2:

$$\begin{aligned}
p_2(\alpha,\theta) &= \sum_{i=1}^{u-1}\frac{i}{\theta+i\alpha}= \\
&= \frac{1}{\alpha^2}\left(-\theta\cdot\psi\left(u+\frac{\theta}{\alpha}\right)+\theta\cdot\psi\left(1+\frac{\theta}{\alpha}\right)+\alpha(u-1)\right).
\end{aligned}$$

As can be seen, p_1 and p_2 share evaluations of digamma for identical inputs and they can thus be fused with each other.

The motivation for performing these transformations is that the closed forms can be evaluated using efficient numerical approximations of digamma. We have used an implementation developed at Microsoft Research [55]. Because the objective functions and derivatives are only approximated, we need to make sure that no errors are introduced. For this purpose, we employ a three-step process:

Step 1: Try to solve the equation system with the approximations of all required functions.
Step 2: If a result has been found, *verify* it with the original iterative forms of the functions.
Step 3: If the verification fails or no solution has been found by step 1, solve the system using the iterative forms of all functions.

VII. Evaluation

A. Experimental Setup

We have integrated our method into ARX, which is an open source de-identification tool for health data [12], and we have performed experiments with two parameterizations: a uniqueness threshold of 1% and a threshold of 5%. We have compared our approach with the k-anonymity model using two parameterizations: a risk threshold of 20% ($k=5$), which has been recommended for biomedical data [56], and a threshold of 50% ($k=2$), which is the weakest possible parameterization. Unless noted otherwise, we will report results for risk thresholds of 1% and $k=5$ using the Loss quality model only, because we have observed comparable results with other configurations. All experiments were performed on a desktop PC with a quad-core 3.1 GHz Intel Core i5 CPU running a 64-bit Linux 3.2.0 kernel and a 64-bit JVM (1.7.0).

We have used the following data from registries and health surveys from the U.S. *($N=318.9$ M)*: 100,937 records about traffic accidents from the NHTSA Fatality Analysis Reporting System (FARS), 539,253 records from the American Time Use Survey (ATUS) and 1,193,504 records from the Integrated Health Interview Series (IHIS). We have also included two de-facto standard datasets for benchmarking de-identification methods: 30,162 records from the 1994 U.S. Census (ADULT) and 63,441 records from the 1998 KDD competition (CUP). For a detailed specification of the datasets we refer to [57].

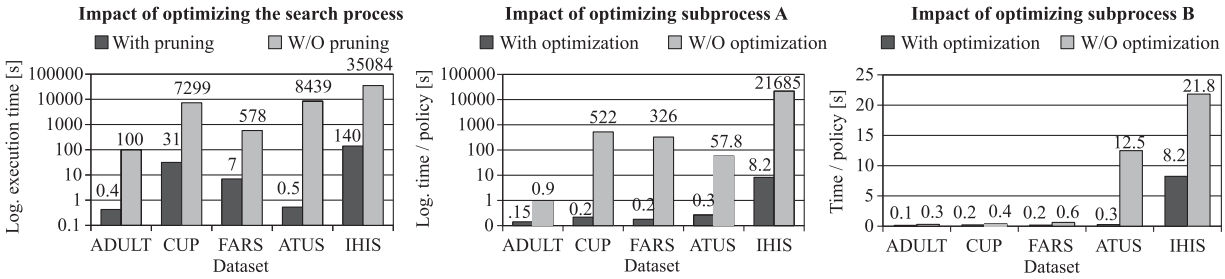

Fig. 10. Analysis of the impact of the developed optimizations on data de-identification with the model by Dankar *et al.* with 1% threshold.

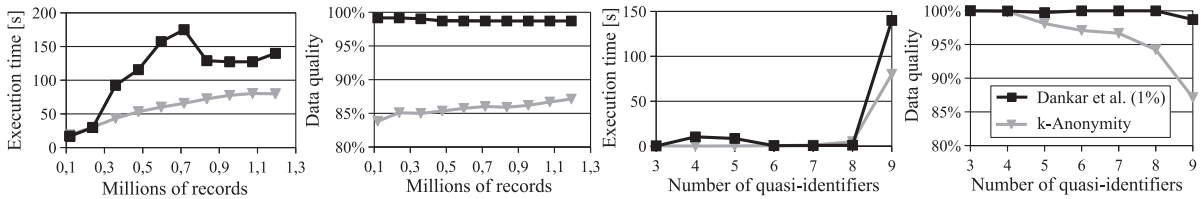

Fig. 11. Scalability of our approach for the largest evaluation dataset (Health Interview Series).

TABLE I
SUMMARY OF RESULTS

	Method	ADULT	CUP	FARS	ATUS	IHIS
Loss	5-Anonymity	77.9%	86.0%	84.7%	91.3%	87.1%
	2-Anonymity	83.8%	88.6%	88.7%	94.5%	92.3%
	Dankar *et al.* (1%)	98.1%	97.3%	96.2%	99.8%	98.7%
	Dankar *et al.* (5%)	100%	97.7%	97.3%	100%	99.6%
Entropy	5-Anonymity	60.5%	45.6%	70.9%	79.4%	71.9%
	2-Anonymity	69.3%	50.8%	76.7%	84.9%	77.4%
	Dankar *et al.* (1%)	92.5%	77.0%	87.5%	96.9%	89.8%
	Dankar *et al.* (5%)	100%	80.5%	93.2%	100%	96.0%
Time	k-Anonymity	2.4 s	**17.7 s**	**6.2 s**	5.1 s	**81.7 s**
	Dankar *et al.*	**0.4 s**	31.4 s	7.1 s	**0.6 s**	139.9 s

(Entropy = Non-Uniform Entropy).

For all datasets we have selected between eight and nine quasi-identifiers. They contained traditional demographics (e.g., age, sex), which have often been used in re-identification attacks, as well as further attributes, such as marital status and education, which may lead to accidental re-identification. Our implementation of the experiments is available online [58].

B. Summary of Results

Table I summarizes the results. It shows the execution times and data quality obtained for the five datasets, two privacy models and two quality models. As can be seen, choosing 2-anonymity over 5-anonymity only resulted in minor improvements (up to $\sim 9\%$). Using the super-population model, in contrast, consistently resulted in data with a significantly higher degree of quality than when using k-anonymity. The effect was particularly strong when information loss was measured with Non-Uniform Entropy (improvements by up to $\sim 40\%$). With our approach very little information content needed to be removed to protect the datasets in most cases.

We observed comparable execution times when using k-anonymity and population uniqueness. In some cases using population uniqueness was slower (CUP, FARS, IHIS), while in other cases it was faster (ADULT, ATUS). The reason for the absence of a general trend lies in the complex interplay of the different optimizations involved. For example, the effectiveness of the method for reducing the number of candidate policies (see Section VI-F) increases with the quality of the optimal solution. As a consequence, restrictions on population uniqueness can sometimes be achieved more efficiently than k-anonymity, although the process of evaluating individual de-identification policies is much more complex.

C. Impact of Optimizations

The results from Table I already show that the optimizations which we have developed are highly effective. Fig. 10 shows a detailed analysis. As can be seen, large parts of the search spaces were pruned by our method for excluding candidate policies. This resulted in significant reductions of execution times by factors of up to 15,929 (ATUS). Our method for fast record suppression (i.e., the optimization of subprocess A, cf. Section VI-G), further improved execution times significantly, i.e., by factors of up to 2,631 (IHIS). Finally, the efficient implementations of the formulas required to calculate risks (i.e., the optimization of subprocess B, cf. Section VI-H) provided additional speed-ups by factors of up to 47 (ATUS).

We emphasize that the optimizations are independent of each other and that the individual speed-up factors can therefore be multiplied to estimate the overall speed-up achieved by our implementation. For the largest dataset, IHIS, a conservative estimation yields a total speed-up factor of more than one million. This shows that our method would not be computationally feasible without implementing the optimizations presented in this article.

D. Analysis of Scalability

Fig. 11 shows the results of an evaluation of the scalability of our method. We measured a linear increase in execution times and data quality with increasing data volume. We observed a

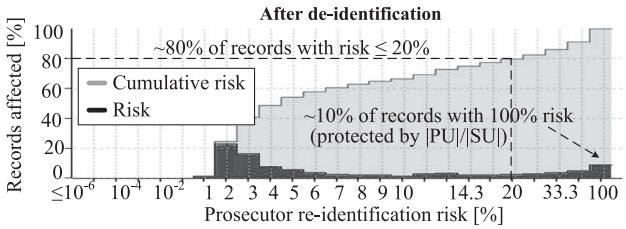

Fig. 12. Risk profile of the Health Interview Series dataset after de-identification with the model by Dankar *et al.* with a 1% threshold.

few minor irregularities when using our approach. The reason is that it is not always possible to automatically solve the equation system used by the privacy model [52]. However, this is a rare event which does not have any privacy implications and, as can be seen, only negligible effects on data quality. It can also be seen in Fig. 11 (and in Table I) that for small datasets the quality improvements achieved over k-anonymity tend to be higher than for large datasets.

The execution times of our approach are roughly exponential in the dimensionality of the data (apart from negligible variation as a result of our optimizations). This is because the size of the solution space also increases exponentially. For cases where this is a problem, ARX also provides an effective heuristic search algorithm [59].

VIII. Assessing Residual Risks

Our approach is scalable and it produces high-quality output. However, the privacy guarantees provided are weak compared to k-anonymity and they have complex semantics, which rely on estimates of population characteristics. It is therefore recommended that data custodians employ a semi-automated process which combines the a priori and the a posteriori methodology. For example, the ARX tool offers various visualizations for performing risk analyses and methods for optimizing data quality.

Firstly, changes in the risk profiles of datasets can be analyzed. As can be seen in Fig. 12, using a population-oriented de-identification model significantly changed the profile of the Health Interview Series dataset. In the input (see Fig. 7) only 20% of the records were affected by risks of not more than 20% and 50% of the records were affected by a risk of 100%. In the de-identified output dataset, however, almost 80% of the records were affected by re-identification risks of not more than 20% and only about 10% of the records had a risk of 100%. These are conservative estimates and risks are likely to be even lower in practice [17].

Secondly, it should be checked whether a sufficient degree of uncertainty has been introduced for records which are unique within the dataset (10% in our example), i.e., whether the quotient $\frac{|PU|}{|SU|}$ is small enough. While we have de-identified the datasets focusing on uniqueness within the U.S. population, we have also analyzed the uncertainty introduced for much smaller population sizes. As examples we chose California *($N = 39.1$ M)* and Los Angeles *($N = 4.0$ M)*.

As can be seen in Table II, a sufficient degree of uncertainty was introduced for all datasets. We recommend a threshold of

TABLE II
Degree of Certainty Regarding the Identity of Sample Uniques (Model by Dankar *et al.* with 1% Threshold)

		ADULT	CUP	FARS	ATUS	IHIS
PU/SU	USA	4.0%	1.1%	4.0%	12.2%	6.7%
	California	8.3%	4.3%	9.4%	24.5%	18.7%
	Los Angeles	18.4%	18.9%	23.6%	**52.0%**	**56.6%**

(Protection = High, Medium, **Low**)

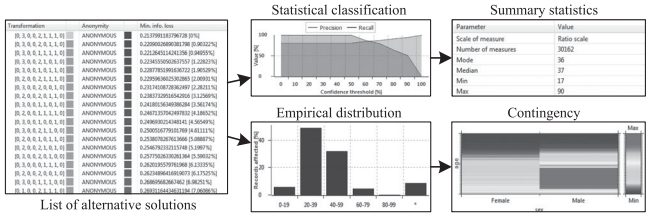

Fig. 13. Example of methods provided for tuning the quality of output data.

25% to account for the fact that the method relies on estimates which may be slightly off. This provides a large margin, as already a probability of 50% would be acceptable. It has also been shown that super-population models tend to over-estimate population uniqueness, which means that the degree of uncertainty is likely to be even higher in reality [16].

The probability that a unique record corresponds to a specific individual from the U.S. population was estimated to be between 1.1% and 12.2%. Even when assuming that the recipient has enough background knowledge to narrow the population down to much smaller sizes a sufficient degree of protection was often measured, especially for the smaller datasets. For the larger datasets, the model must be parameterized with correct population sizes to achieve adequate protection. This is reasonable, as e.g., the Health Interview Series dataset consists of records about 1.2 M individuals, which equals more than 25% of the population of Los Angeles.

Given that the datasets are to be provided to trusted researchers in controlled environments, they can be considered safe for sharing. This was achieved with reduction in data granularity (Loss) of only between 0.2% and 2.7%, resulting in changes to the distributions of data values (Non-Uniform Entropy) between 3.1% and 23%.

Finally, we also recommend that data custodians validate the quality of output data. As is shown in Fig. 13, the ARX tool provides various methods for evaluating the impact of de-identification on the properties of a dataset. This includes analyzing changes to distributions of attribute values, e.g., deviations in measures of tendency, dispersion or shape, and to dependencies between attributes. Moreover, users can analyze how de-identification affects the performance of statistical classification models. ARX also uses the information collected during data de-identification to provide users with alternative solutions which might fit their usage scenario. Finally, constraints can be specified prior to data de-identification. Examples include requiring that certain variables need to be represented with a specific granularity or that scales of measures of variables must be preserved [12].

IX. Discussion

The approach presented in this article follows the *"onion-skin" principle* (see Fig. 1) to protect the privacy of patients and probands when sharing data for biomedical research. At the outer layer, we assume that the data custodian has implemented adequate processes and organizational structures to ensure that the recipients of data are trustworthy. These organizational measures are then combined with two further technical measures. Firstly, a controlled data sharing environment restricts the ways recipients can interact with the data. As a result, it becomes extremely difficult to perform linkage attacks. Secondly, direct disclosure is mitigated by utilizing a population-oriented data de-identification method. It introduces an adequate degree of uncertainty about the identity of data subjects while still providing high quality data.

The presented method is pragmatic on multiple levels. Firstly, it accepts that there is no silver bullet in data privacy but that a variety of safeguards must be installed to achieve a high degree of protection. Secondly, it employs estimates of population characteristics, albeit with very strict thresholds. A related approach has also been used in the process of developing the HIPAA Safe Harbor methodology [60]. Thirdly, it trades data quality off against privacy risks to introduce uncertainty about the identity of data subjects. This approach has also been implemented into other interactive systems, e.g., into i2b2 which can be configured to add noise to query results [61]. Finally, controlled data sharing environments provide a relatively high degree of flexibility, e.g., support for explorative analyses, hypotheses generation and feasibility studies, but the approach must be extended to cover further use cases. For example, the basic design described in this article does not allow data recipients to retrieve results of analyses in structured form, as this may also pose privacy risks [18]. A possible solution is to employ methods of *secondary control*, where users are provided with secure remote workspaces in which they can store the results of analyses. These are then checked and potentially sanitized by the data custodian before they can be downloaded [19].

The proposed approach is also scalable on multiple levels. Firstly, as discussed previously, it can support a wide variety of use cases. Secondly, it can also be set up efficiently in large-scale data sharing environments. For example, the implementation of the proxy can be wrapped into Docker containers, which can then be deployed very quickly for providing specific recipients with secure access to specific datasets [62]. Finally, the tailor-made method of data de-identification is computationally efficient and it can handle very large datasets.

X. Future Work

There are several ways in which the described approach can be further enhanced. For example, our de-identification method is well suited for preventing attacks which use typical quasi-identifiers, e.g., demographic attributes. However, it cannot easily be used to protect data with complex high-dimensional identifiers. If such background knowledge may be available to the data recipient, our method must be extended. Loukides *et al.* [63] and Heatherly *et al.* [64] have proposed an approach for anonymizing diagnosis codes for association studies between phenotypic and genotypic data. They investigated a scenario in which a subset of the data from an Electronic Medical Record (EMR) system is de-identified for research purposes and where the adversary tries to use diagnosis codes to correlate the de-identified records with the original records from the EMR. Although the controlled environments used by our approach render such complex linkage attacks infeasible, the methods developed to counter them are also important in our context. The reason is that when the number of potentially identifying attributes becomes high, complex inter-attribute relationships may significantly reduce the quality of de-identified data [65]. As proposed by Loukides *et al.* several attributes with the same semantics can then be combined into a single set-valued attribute to remove irrelevant inter-attribute relationships. We plan to integrate this method into our system in future work.

Recently, models for analyzing privacy risks have also been proposed which can help to improve the quality of de-identified data with few and many potentially identifying attributes. For example, Wan *et al.* have developed a game-theoretic approach for reasoning about re-identification [15]. Here, the de-identification problem is modeled as a game between the data publisher and the data recipient, both of which try to maximize their monetary gain. Instead of directly reducing re-identification risks, the data is de-identified in a way which maximizes the publisher's payout. We plan to extend our work by implementing a variant of the game which is optimized for use in controlled data sharing environments.

XI. Conclusion

We have described a concept for the safe sharing of high-quality health data. We have focused on processes, methods and properties which we consider important in our environment. Depending on site-specific, e.g., regulatory, requirements and the sensitivity of the data which is to be shared, some safeguards implemented by our approach may also be optional. We also note that we have focused on methods for protecting data privacy and that it is also important to make sure that relevant measures of information security are employed to protect data from unauthorized access. We emphasize that our approach uses transformation methods which have frequently been recommended in guidelines for health data de-identification (see [13], [30], and [50]).

References

[1] J. C. Denny *et al.*, "Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data," *Nature Biotechnol.*, vol. 31, no. 12, pp. 1102–1111, Nov. 2013.

[2] S. Schneeweiss, "Learning from big health care data," *New England J. Med.*, vol. 370, no. 23, pp. 2161–2163, 2014.

[3] U.S. National Institutes of Health, "NIH genomic data sharing policy—NOT-OD-14-124," Bethesda, MD, USA, 2014. [Online]. Availabe: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html. Accessed on: Aug. 1, 2016.

[4] V. Liu, M. Musen, and T. Chou, "Data breaches of protected health information in the united states," *J. Amer. Med. Assoc.*, vol. 313, no. 14, pp. 1471–1473, 2015.

[5] D. Hallinan, M. Friedewald, and P. McCarthy, "Citizens' perceptions of data protection and privacy in europe," *Comput. Law Security Rev.*, vol. 28, no. 3, pp. 263–272, 2012.

[6] L. Sweeney, "Computational disclosure control—A primer on data privacy protection," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci. Massachusetts Inst. Technol., Cambridge, MA, USA, 2001.

[7] W. Xia, R. Heatherly, X. Ding, J. Li, and B. A. Malin, "R-U policy frontiers for health data de-identification," *J. Amer. Med. Informat. Assoc.*, vol. 22, no. 5, pp. 1029–1041, 2015.

[8] U.S. Department of Health and Human Services–Office for Civil Rights, "HIPAA administrative simplification statute and rules, 45 CFR Parts 160, 162, and 164," Washington, DC, USA, 2013.

[9] "U.S. Health insurance portability and accountability act of 1996," 1996, Public Law. 1996:1–349.

[10] "Directive 95/46/EC of the European parliament and of the council of 24 October 1995 on the protection of individuals with regard to the processing of personal data," *Official J. Eur. Union*, vol. L281/38, Nov. 1995, pp. 31–50.

[11] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in *Proc. Symp. Principles Database Syst.*, 1998, p. 188.

[12] F. Prasser and F. Kohlmayer, "Putting statistical disclosure control into practice: The ARX data anonymization tool," in *Medical Data Privacy Handbook*, New York, NY, USA: Springer, 2015, pp. 111–148.

[13] K. E. Emam *et al.*, "A globally optimal k-anonymity method for the de-identification of health data," *J. Amer. Med. Informat. Assoc.*, vol. 16, no. 5, pp. 670–682, 2009.

[14] K. El Emam and C. Álvarez, "A critical appraisal of the article 29 working party opinion 05/2014 on data anonymization techniques," *Int. Data Privacy Law*, vol. 5, no. 1, pp. 73–87, 2014.

[15] Z. Wan *et al.*, "A game theoretic framework for analyzing re-identification risk," *PloS one*, vol. 10, no. 3, 2015, Art. no. e0120592.

[16] F. K. Dankar *et al.*, "Estimating the re-identification risk of clinical data sets," *BMC Med. Informat. Decision Making*, vol. 12, no. 1, pp. 1–15, 2012.

[17] F. Prasser, F. Kohlmayer, and K. Kuhn, "The importance of context: Risk-based de-identification of biomedical data," *Methods Inf. Med.*, vol. 55, no. 4, pp. 347–355, 2016.

[18] G. Duncan, M. Elliot, and J. Salazar-González, *Statistical Confidentiality: Principles and Practice*. New York, NY, USA: Springer, 2011.

[19] F. Ritchie and M. Elliott, "Principles- versus rules- based output statistical disclosure control in remote access environments," *IASSIST Quart.*, vol. 39, no. 2, pp. 5–13, 2015.

[20] G. T. Duncan, M. Elliot, and J.-J. Salazar-González, "Restrictions on data access," in *Statistical Confidentiality*, New York, NY, USA: Springer, 2011, pp. 137–146.

[21] J. W. Bos, K. Lauter, and M. Naehrig, "Private predictive analysis on encrypted medical data," *J. Biomed. Informat.*, vol. 50, pp. 234–243, 2014.

[22] W.-J. Lu *et al.*, "Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption," *BMC Med. Informat. Decision Making*, vol. 15, no. Suppl 5, p. S1, 2015.

[23] C. Dwork *et al.*, "Calibrating noise to sensitivity in private data analysis," in *Proc. Conf. Theory Cryptography*, 2006, pp. 265–284.

[24] F. Yu and Z. Ji, "Scalable privacy-preserving data sharing methodology for genome-wide association studies: An application to iDASH healthcare privacy protection challenge," *BMC Med. Informat. Decision Making*, vol. 14, no. 1, pp. 1–8, 2014.

[25] F. K. Dankar and K. El Emam, "Practicing differential privacy in health care: A review." *Trans. Data Privacy*, vol. 6, no. 1, pp. 35–67, 2013.

[26] N. R. Council, *Expanding Access to Research Data: Reconciling Risks and Opportunities*. Washington, DC, USA: The National Academies Press, 2005.

[27] F. Ritchie, "Secure access to confidential microdata: Four years of the virtual microdata laboratory," *Labour Gazette*, vol. 2, no. 5, pp. 29–34, 2008.

[28] B. Malin, D. Karp, and R. H. Scheuermann, "Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research," *J. Investigative Med.*, vol. 58, no. 1, pp. 11–18, 2010.

[29] J. Eder, H. Gottweis, and K. Zatloukal, "It solutions for privacy protection in biobanking," *Public Health Genomics*, vol. 15, no. 5, pp. 254–262, 2012.

[30] K. E. Emam and B. A. Malin, "Appendix B: Concepts and methods for de-identifying clinical trial data," in *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, Committee on Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences Policy, and Institute of Medicine, Eds. Washington, DC, USA: National Academies Press, 2015, pp. 1–290.

[31] "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)," *Official J. Eur. Union*, vol. L119/59, May 2016.

[32] S. N. Murphy *et al.*, "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)," *J. Amer. Med. Informat. Assoc.*, vol. 17, no. 2, pp. 124–130, 2010.

[33] B. D. Athey, M. Braxenthaler, M. Haas, and Y. Guo, "Transmart: An open source and community-driven informatics and data sharing platform for clinical and translational research," *Proc AMIA Summits Translational Sci.*, vol. 2013, pp. 6-8, 2013.

[34] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen, "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements," *Requirements Eng.*, vol. 16, no. 1, pp. 3–32, 2011.

[35] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. 2008 IEEE Symp Security Privacy*, 2008, pp. 111–125.

[36] T. C. Rindfleisch, "Privacy, information technology, and health care," *Commun. ACM*, vol. 40, no. 8, pp. 92–100, 1997.

[37] T. Richardson and J. Levine, "The remote framebuffer protocol," RFC 6143 (Informational), Internet Engineering Task Force, Mar. 2011. [Online]. Available: http://www.ietf.org/rfc/rfc6143.txt

[38] R. Smith, "An overview of the tesseract OCR engine," in *Proc. Int. Conf. Document Anal. Recognit.*, 2007, pp. 629–633.

[39] F. Kohlmayer, F. Prasser, and K. Kuhn, "The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal loss of information," *J. Biomed. Informat.*, vol. 58, pp. 37–48, 2015.

[40] F. Prasser, F. Kohlmayer, and K. A. Kuhn, "Efficient and effective pruning strategies for health data de-identification," *BMC Med. Informat. Decision Making*, vol. 16, no. 1, pp. 1–14, 2016.

[41] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *Int. J. Uncertainty Fuzz*, vol. 10, no. 5, pp. 571–588, 2002.

[42] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. IEEE Int. Conf. Data Eng.*, 2006, pp. 25–25.

[43] M. E. Nergiz and C. Clifton, "Thoughts on k-anonymization," in *Proc. IEEE Int. Conf. Data Eng. Workshops*, 2006, pp. 96–96.

[44] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *Trans. Knowl Discovery Data*, vol. 1, no. 1, 2007, Art. no. 3.

[45] R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. Int. Conf. Data Eng.*, 2005, pp. 217–228.

[46] V. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 279–288.

[47] A. De Waal and L. Willenborg, "Information loss through global recoding and local suppression," *Netherlands Official Stat.*, vol. 14, pp. 17–20, 1999.

[48] K. El Emam and L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get you Started*, 1st ed. Sebastopol, CA, USA: O'Reilly and Associates, 2014.

[49] D. C. Barth-Jones, "The're-identification'of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now," *SSRN*, 2012. [Online]. Available: http://dx.doi.org/10.2139/ssrn.2076397

[50] K. El Emam, *Guide to the De-Identification of Personal Health Information*. 1st ed., Boca Raton, FL, USA: CRC Press, 2013.

[51] G. Chen and S. Keller-McNulty, "Estimation of identification disclosure risk in microdata," *J. Official Stat.*, vol. 14, pp. 79–95, 1998.

[52] N. Hoshino, "Applying pitman's sampling formula to microdata disclosure risk assessment," *J. Official Stat.*, vol. 17, no. 4, pp. 499–520, 2001.

[53] L. V. Zayatz, "Estimation of the percent of unique population elements on a microdata file using the sample," Statistical Research Division, U.S. Bureau Census, Washington, DC, USA, Tech. Rep. Census/SRD/RR-91/08, 1991.

[54] R. Mickens, *Difference Equations: Theory, Applications and Advanced Topics*, 3rd ed., (Monographs and Research Notes in Mathematics). Boca Raton, FL, USA: CRC Press, 2015.

[55] "Lightspeed Matlab Toolbox." [Online]. Available: http://research.microsoft.com/en-us/um/people/minka/software/lightspeed/. Accessed on: Aug. 1, 2016.
[56] K. El Emam and F. K. Dankar, "Protecting privacy using k-anonymity," *J. Amer. Med. Inf. Assoc.*, vol. 15, no. 5, pp. 627–637, 2008.
[57] F. Prasser, F. Kohlmayer, and K. A. Kuhn, "A benchmark of globally-optimal anonymization methods for biomedical data," in *Proc. 2014 IEEE Int. Symp. Comput., Based Med. Syst.*, 2014, pp. 66–71.
[58] "Uniqueness benchmark." [Online]. Available: https://github.com/arx-deidentifier/uniqueness-benchmark. Last accessed on: Aug. 9, 2016.
[59] F. Prasser, R. Bild, J. Eicher, H. Spengler, F. Kohlmayer, and K. A. Kuhn, "Lightning: Utility-driven anonymization of high-dimensional data," *Trans. Data Privacy*, vol. 9, no. 2, pp. 161–185, 2016.
[60] U.S. Department of Health and Human Services—Office of the Assistant Secretary for Planning and Evaluation, "Standards for privacy of individually identifiable health information," *Federal Register*, vol. 65, no. 250, pp. 82462–82829, 2000.
[61] S. N. Murphy, V. Gainer, M. Mendis, S. Churchill, and I. Kohane, "Strategies for maintaining patient privacy in i2b2," *J. Amer. Med. Informat. Assoc.*, vol. 18, no. Supplement 1, pp. i103–i108, 2011.
[62] D. Merkel, "Docker: Lightweight linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, 2014, Art. no. 2.
[63] G. Loukides and A. Gkoulalas-Divanis, "Utility-aware anonymization of diagnosis codes," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 60–70, Jan. 2013.
[64] R. D. Heatherly, G. Loukides, J. C. Denny, J. L. Haines, D. M. Roden, and B. A. Malin, "Enabling genomic-phenomic association discovery without sacrificing anonymity," *PloS one*, vol. 8, no. 2, 2013, Art. no. e53875 .
[65] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proc. Int. Conf. Very Large Databases*, 2005, pp. 901–909.