The Importance of Context: Risk-based De-identification of Biomedical Data*

Fabian Prasser; Florian Kohlmayer**; Klaus A. Kuhn**

Technical University of Munich, University Hospital rechts der Isar, Institute of Medical Statistics and Epidemiology, Munich, Germany

Keywords

Information science, computer security, data protection, data anonymization, risk, data quality

Summary

Background: Data sharing is a central aspect of modern biomedical research. It is accompanied by significant privacy concerns and often data needs to be protected from re-identification. With methods of de-identification datasets can be transformed in such a way that it becomes extremely difficult to link their records to identified individuals. The most important challenge in this process is to find an adequate balance between an increase in privacy and a decrease in data quality.

Objectives: Accurately measuring the risk of re-identification in a specific data sharing scenario is an important aspect of data de-identification. Overestimation of risks will significantly deteriorate data quality, while underestimation will leave data prone to attacks on privacy. Several models have been proposed for measuring risks, but there is a lack of generic methods for risk-based data de-identification. The aim of the work described in this article was to bridge this gap and to show how the quality of de-identified datasets can be improved by using risk models to tailor the process of de-identification to a concrete context.

Methods: We implemented a generic de-identification process and several models for measuring re-identification risks into the ARX de-identification tool for biomedical data. By integrating the methods into an existing framework, we were able to automatically transform datasets in such a way that information loss is minimized while it is ensured that re-identification risks meet a user-defined threshold. We performed an extensive experimental evaluation to analyze the impact of using different risk models and assumptions about the goals and the background knowledge of an attacker on the quality of de-identified data.

Results: The results of our experiments show that data quality can be improved significantly by using risk models for data de-identification. On a scale where 100% represents the original input dataset and 0% represents a dataset from which all information has been removed, the loss of information content could be reduced by up to 10% when protecting datasets against strong adversaries and by up to 24% when protecting datasets against weaker adversaries.

Conclusions: The methods studied in this article are well suited for protecting sensitive biomedical data and our implementation is available as open-source software. Our results can be used by data custodians to increase the information content of de-identified data by tailoring the process to a specific data sharing scenario. Improving data quality is important for fostering the adoption of de-identification methods in biomedical research.

Correspondence to:
Dr. Fabian Prasser
Institute of Medical Statistics and Epidemiology
University Hospital rechts der Isar
Technical University of Munich
Grillparzerstr. 18
81675 Munich
Germany
E-mail: fabian.prasser@tum.de

Methods Inf Med 2016; 55: 347–355
http://dx.doi.org/10.3414/ME16-01-0012
received: February 5, 2016
accepted in revised form: April 12, 2016
epub ahead of print: June 20, 2016

* Supplementary material published on our website http://dx.doi.org/10.3414/ME16-01-0012

** These authors contributed equally to this work

1. Introduction

Recent developments in medicine, notably the trends towards precision medicine and to "learning health systems" are accompanied by significant needs of privacy protection. This includes methods for secure exchange of data among systems "without making patients identifiable" [1]. Secure data sharing is needed for various applications [2], including phenome-wide association studies [3] and the secure secondary use of operational clinical data [4]. In research environments, increasingly, the sharing of research data is a central aspect which is also required from public sponsors [5]. At the same time, the number of health data breaches is growing [6] and there is a significant public pressure to ensure the anonymity of data subjects [7]. In an era of large scale collection and processing of sensitive personal data this is a challenging task [8].

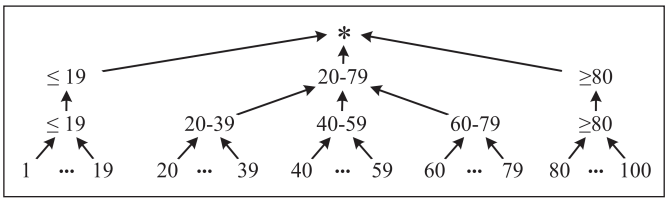

Figure 1
Domain generalization hierarchy for the attribute *age*.

A central aspect of anonymization is to protect data from *re-identification*. Protection against this privacy threat can be implemented by transforming datasets in order to ensure that it is extremely difficult to link its records to identified individuals [9]. This process is called *data de-identification* [10]. The most wide-spread transformation methods for de-identifying health data are *generalization* and *suppression* (i.e. removal) of attribute values [11]. The former is typically performed with user-defined domain generalization hierarchies. An example is shown in ▶Figure 1, where values of the attribute *age* are transformed into groups with decreasing precision. A well-known privacy model using such transformations is the *Safe Harbor* method, which is defined in the Privacy Rule of the US Health Insurance Portability and Accountability Act (HIPAA) [12]. This method specifies 18 rules for transforming attributes which are associated with a high-risk of re-identification. Among these rules are the suppression of names and the generalization of dates. Another privacy model which is often implemented with generalization and suppression is *k*-anonymity. It requires that a dataset is to be transformed in such a way that each record cannot be distinguished from k-1 other records [13].

Unfortunately, privacy risks can never be reduced to completely zero [11]. Therefore, national and international laws, such as HIPAA [14], European national laws, and the European Directive on Data Protection [15], recognize the importance of considering context when implementing measures for data protection. For example, the EU Directive states that "account should be taken of all the means *likely reasonably to be used* [...] to identify the said person [emphasis added]" [15]. The de-identification of data results in loss of information. Thus, finding an adequate balance between an increase in privacy and a decrease in data quality is a central challenge [16]. Different methods for de-identifying data result in datasets with different residual risks of re-identification and different degrees of utility for different applications. To consider this, methods must be tailored to the context. One important aspect is to accurately measure the re-identification risks of a dataset in relationship to a given data sharing scenario. This includes considering the characteristics and sensitivity of the data as well as the aim and possibilities of potential adversaries. Overestimation of risks will significantly deteriorate data quality, whereas underestimation will leave the data prone to attacks on privacy.

2. Objectives

Although several models for measuring the re-identification risks of biomedical datasets have been proposed, there is a lack of generic methods which can be used to automatically de-identify data with a wide variety of such models. This prevents data controllers from performing context-aware de-identification, which has a significant potential to improve data quality. The objective of the work described in this article was to bridge this gap by implementing models for managing re-identification risks into an open source anonymization tool for biomedical data. In this process, we developed a generic de-identification method and complemented an existing implementation of *k*-anonymity with three additional risk models. As all methods were realized within a common framework, we were able to perform an extensive experimental comparison which shows how the quality of de-identified data can be improved by considering the goals and the background knowledge of an attacker.

3. Re-identification Attacks

Typically, datasets with personal data represent a *sample* from a larger *population* of individuals. In a data sharing scenario, it is assumed that the dataset has already been stripped from any direct identifiers, such as names, but that the adversary has access to some background knowledge about the population which comprises identifying information. Typical examples of background knowledge are voter registration

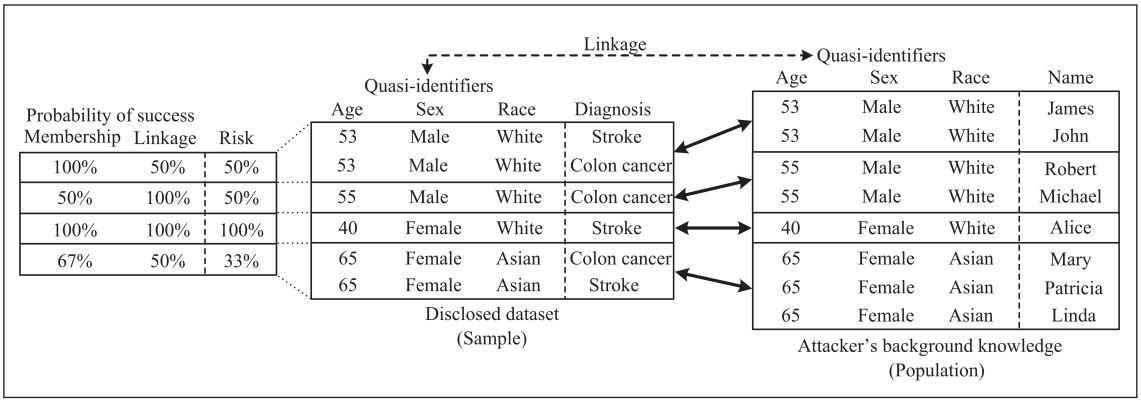

Figure 2
Examples of different re-identification risks. An arrow indicates that each element in a group of indistinguishable records in the population table matches all records in the according group in the sample. Risks can then be derived from the sizes of these groups.

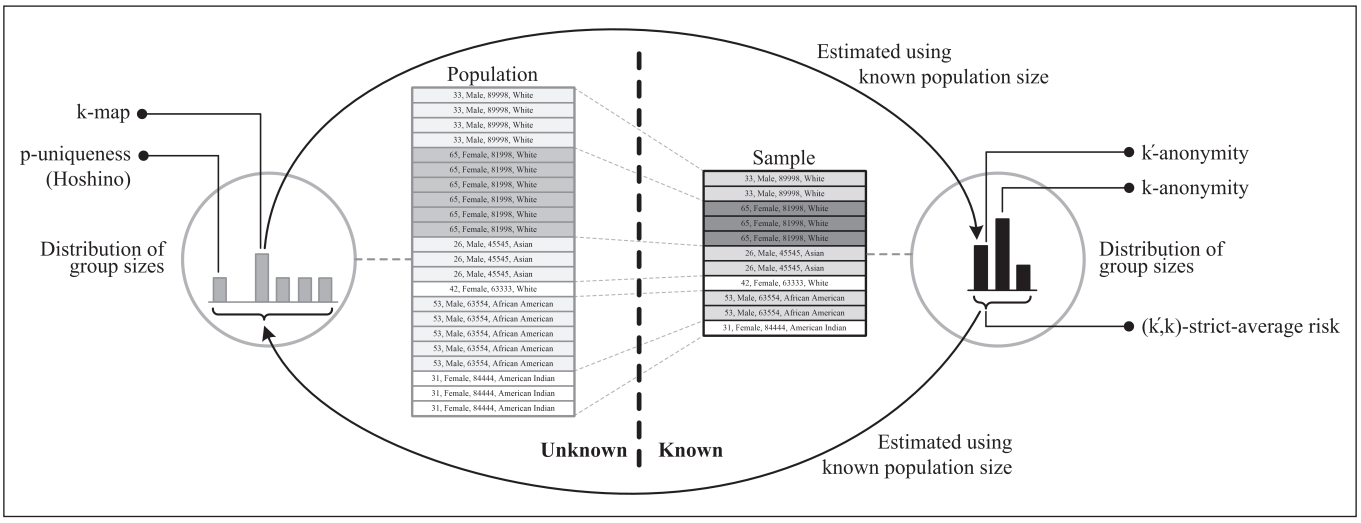

Figure 3 Principles underlying the risk models considered in this article.

lists, which can easily be obtained in many US states [17]. As is shown in ▶Figure 2, the adversary may use this information to link a disclosed dataset containing sensitive data (diagnoses in our example) with a dataset containing identifying information. The attributes which can be used to perform such attacks are called *quasi-identifiers*. While successful linkage attacks have been demonstrated by examples involving real-world data [17], they are complicated in general as the chance of success depends on a variety of parameters.

Most importantly, it has to be seen that a successful re-identification attack is actually a two-step process. First, the attacker needs to determine whether or not data about a specific individual is contained in the dataset (*membership*). Second, the individual must be *linked* with the corresponding record. For both process steps, the probability of success depends on the distinguishability (or *uniqueness*) of the individual's records in the sample *as well as* in the adversary's background knowledge.

The data about each individual is part of an *equivalence class* of records with identical values of those quasi-identifiers which are used by the adversary. Let us consider an individual I from the population. We denote the group containing data about I in the population with P and the group containing data about I in the sample with S. We note that S is empty if I is not contained in the sample. Three different types of attackers are typically considered [10]:

- In the *prosecutor attacker model* it is assumed that the adversary already knows that data about an individual I is contained in the dataset. Consequently, the probability for correctly linking the individual with an entry from the sample is $P(\text{Linkage} \mid \text{Membership}) = 1 / |S|$.
- In the *journalist attacker model* it is assumed that the adversary has no prior knowledge about membership. The probability that data about I is contained in the sample is $P(\text{Membership}) = |S| / |P|$ and therefore the probability of a successful linkage attack is $P(\text{Membership}) \times P(\text{Linkage} \mid \text{Membership}) = |S| / |P| \times 1 / |S| = 1 / |P|$.
- The basic assumptions of the *marketer attacker model* are that the adversary has no prior knowledge and that she aims at re-identifying a larger number of individuals (e.g. for marketing purposes, hence the name). As a consequence, the effort is only worth it if a significant fraction of records can be re-identified. Therefore marketer risk can be expressed as an average of the re-identification risks of all records [18].

4. Methods

4.1 Risk Models

Protecting a dataset against prosecutor attacks will also protect the dataset against journalist attacks. Moreover, protecting a dataset against journalist attacks will also protect the dataset from marketer attacks [10]. The most well-known privacy model, *k*-anonymity, controls the sizes of groups in the sample and therefore focuses on protecting datasets from prosecutor attacks [13]. However, *k*-anonymity is very strict and there is a potential to significantly improve the quality of de-identified data by considering less powerful adversaries and the relationships between the sample and the attacker's background knowledge.

The general principles underlying the models considered in this article are illustrated in ▶Figure 3. As can be seen, all models focus on the distribution of the sizes of groups in the dataset. In each histogram, the x-axis represents the sizes (starting from 1, increasing from left to right) and the height of a bar represents the number of groups of the according size. Information about the sample (right side) is known, information about the population (left side) must be estimated from the sample using a known population size. The *k*-anonymity model focuses on the distribution of group sizes in the sample as it requires that no groups exist which contain less than *k* records.

k-Map is a privacy model which also controls the sizes of the groups of indistinguishable records. However, in contrast to *k*-anonymity, the risk threshold is not enforced on the groups in the sample but on the groups in the population table [17]. The model requires the data custodian to have access to a database with detailed data

	Degree of Protection Against Attacker		
Risk Model	**Prosecutor**	**Journalist**	**Marketer**
k-Anonymity	High	High	High
k-Map	Low	High	High
(k',k)-Strict-average risk	Low	Low	High
p-Uniqueness (Hoshino)	-	Low	High

Table 1 Overview of risk models and the protection provided against different types of attackers.

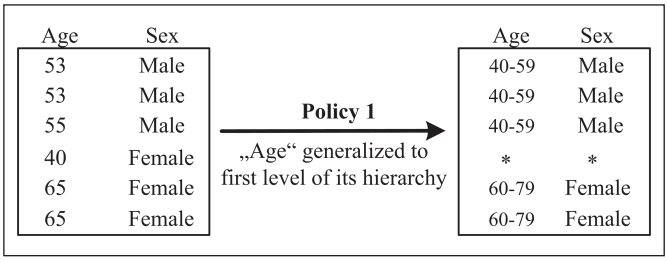

Figure 4 Example of applying a de-identification policy.

about the population, which is not a realistic assumption in practice. A solution to this problem was originally proposed in [19] and adopted for biomedical data in [20]. As is sketched in ▶Figure 3, the basic idea is to assume that group sizes are following a zero-truncated Poisson distribution and to use hypothesis testing to find a parameter $k' \leq k$, such that a dataset which is transformed to fulfill *k'*-anonymity is likely to fulfill *k*-map. For example, the probability that a randomly sampled subset of 30,000 records about individuals from the US which fulfills 2-anonymity also fulfills 5-map is at least 99%. As this model focuses on the size of groups in the population, it can be used to protect datasets from journalist attacks. The implementation is based on *k*-anonymity, and it therefore also offers weak protection against prosecutor attacks.

Measures against marketer attacks can be implemented by enforcing a limit on the average re-identification risk, which can be measured with the average group size in the sample. As a dataset with an average risk below a given threshold may still contain some records with rather high risks, the concept of *(k',k)*-strict-average risk has been proposed. It focuses on the distribution of group sizes in the sample (see ▶Figure 3) and combines a threshold on the average re-identification risk with *k*-anonymity. For example, *(5,3)*-strict-average risk requires an average group size of 5 and a minimal group size of 3. This model offers protection against marketer attacks and weak protection against prosecutor and journalist attacks.

Finally, *super-population models* estimate the distribution of group sizes in the population from the distribution of group sizes in the sample (see ▶Figure 3). As a basis, a distribution is used which is flexible enough to represent various populations. The parameters of this distribution are then estimated from properties of the sample [21–23]. Dankar et al. have experimentally validated a large number of such models with clinical datasets [24]. Their results show that the model by Hoshino [21], which uses Pitman's sampling formula as the underlying distribution, is the most accurate method for common data sharing scenarios. One important application of such models is to calculate the number of records from the sample that are *unique* within the population, as this measure is well suited for estimating marketer re-identification risks [24]. We call the according privacy model *p*-uniqueness. When the threshold on the number of population uniques is set to zero (i.e. *0*-uniqueness), the model also offers weak protection against journalist attacks.

▶Table 1 shows an overview of all methods that we have implemented as part of the work described in this article. Formal definitions can be found online in the supplementary ▶Appendix. It also shows which degree of protection the models offer against different types of attackers when they are used with typical parameters.

4.2 Implementation

We have integrated the above risk and privacy models into the open source data anonymization tool ARX [25]. From a set of possible *de-identification policies* which can be used to transform the input dataset, the tool automatically selects a solution which fulfills the predefined privacy requirements while minimizing the loss of information. For measuring data quality we used the method by Iyengar which calculates data precision by determining the extent to which the domain of an attribute is covered by the transformed values [26]. A formal definition can be found online in the supplementary ▶Appendix. A value of 100% represents the original input dataset and a value of 0% represents a dataset from which all information has been removed.

As is shown in ▶Figure 4, each de-identification policy defines a *generalization scheme*, which is constructed from the user-defined generalization hierarchies (see ▶Figure 1). The application of a generalization scheme is followed by record suppression. In the example, the privacy requirement is *2*-anonymity which defines an upper bound of 50% on prosecutor re-identification risks. The output dataset has a quality of 75%, which means that 25% of the information from the input dataset has been lost as a result of the transformation.

We note that when using 'traditional' de-identification methods it is easy to decide which records need to be suppressed, because the decision can be made by separately analyzing the individual groups. For example, when implementing *k*-anonymity all groups need to be suppressed which contain less than *k* records. In contrast, when using risk models for data de-identification this decision must be based on a holistic view of the whole generalized dataset. For example, it is difficult to decide which groups of records need to be suppressed to achieve a predefined average group size.

To solve this problem, we have designed and implemented a generic process

for risk-based data de-identification which is shown in ▶Figure 5. After a given generalization scheme has been used to transform the input dataset we first calculate the risk according to the given model. If it is not below the given threshold, our method suppresses the group of records with the least information content. This group is determined using the model by Iyengar which is also used by the tool to select the optimal solution [26]. Suppressing a group changes the distribution of group sizes indicated in ▶Figure 3. Next, the risk model is evaluated again, this time for the modified output dataset. This process is repeated until the risk is acceptable. Finally, the algorithm computes the total loss of information for the current de-identification policy and proceeds with the next solution candidate. When all policies have been evaluated it returns the solution with the highest data quality.

The described process is computationally complex, as it needs to evaluate multiple policies to determine a good solution for a given input dataset in terms of output quality. Even evaluating a single policy involves multiple evaluations of the risk model. We have implemented two important optimizations to make this process feasible. First, we used a pruning strategy proposed by Bayardo et al. to exclude policies from the search process by exploiting the fact that lower bounds on the quality of the result of a generalization scheme can be calculated without applying the scheme to the input dataset [27]. Second, we developed highly optimized implementations of the risk models. This is especially important for the *p*-uniqueness model by Hoshino, because evaluating it requires to numerically solve a bivariate non-linear equation system [21].

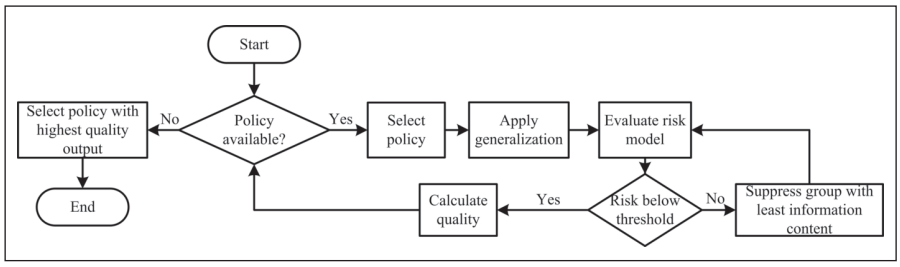

Figure 5 Overview of our risk-based data de-identification process.

5. Experimental Evaluation

5.1 Setup

In our experiments, we used five different datasets: *1)* an excerpt of 30,162 records from the 1994 US census database (ADULT), *2)* a dataset covering 63,441 individuals from the 1998 KDD data mining competition (CUP), *3)* NHTSA crash statistics containing 100,937 cases from their Fatality Analysis Reporting System (FARS), *4)* 539,253 responses from the American Time Use Survey (ATUS) and, *5)* 1,193,504 records from the Integrated Health Interview Series (IHIS). The 1994 census dataset serves as a de-facto standard for the evaluation of de-identification algorithms. All datasets included between eight and nine quasi-identifiers, such as demographics (e.g., age, marital status, sex), information about insurance coverage, social parameters (e.g., education) and health parameters (e.g., weight, health problems). For more information about the datasets the reader is referred to [28].

As a model for managing prosecutor re-identification risks, we used *k*-anonymity with a risk threshold of 20% which is a common parameter for biomedical data [20, 29]. As a model for managing journalist risks while also providing weak protection against prosecutor attacks, we used *k*-map with a risk threshold of 20% and a significance level of $\alpha = 0.1$. These values have been recommended in the literature [20]. As a model for managing marketer re-identification risks while also providing weak protection against prosecutor and journalist attacks, we used *(k',k)*-strict-average risk with thresholds of 50% and 33% for the maximal risk and a threshold of 20% for the average risk. These values have been recommended for health data [29]. To analyze how the improvements in data quality depend on the acceptable risk, we have also performed all experiments with twice as strict thresholds. As a model for managing marketer re-identification risks, we used *p*-uniqueness (Hoshino) with strict thresholds [24] on the proportion of population uniques of between 10^{-6}% and 1%.

Where applicable, we also considered the impact of different assumptions about how good the attacker's background knowledge is. The quality of this knowledge corresponds with the attacker's ability to narrow down the set of individuals that may be represented in the dataset. Therefore, we model this aspect by assuming different sizes of the underlying population. As the datasets are from the US, we chose the following populations with decreasing size: *1)* all US citizens (318.9 million), *2)* the population of the State of Texas (27.0 million) and *3)* the population of Texas' largest city, Houston (2.2 million).

5.2 Results

▶Table 2 and ▶Table 3 show an overview of the data quality and residual risks obtained when de-identifying the five datasets with *k*-anonymity, *k*-map and *(k',k)*-strict-average risk using the previously described parameters. *k*-Anonymity provides the highest degree of protection and it therefore defines the baseline in terms of data quality.

With *k*-map, data quality could be improved significantly while providing identical degrees of protection against journalist and marketer attacks. The decrease in information loss depended on the quality of the attacker's background knowledge. For the ADULT dataset an improvement of 5.4% could be achieved with a risk threshold of 20% and an improvement of 10% could be achieved with a threshold of 10%. No instance of background knowledge considered in the experiments was sufficient to significantly narrow down the set of individuals represented in this dataset. For the largest dataset, IHIS, an improvement of up to 5.2% could be achieved with

Table 2 Data quality and residual risks when de-identifying data using *k*-anonymity, *k*-map and *(k',k)*-strict-average risk with a threshold of 20%. Primary risk thresholds are highlighted. Risk thresholds in parenthesis are not defined by the user but result from the de-identification process.

	Risk Thresholds			Data Quality				
Privacy Model	**Prosecutor**	**Journalist**	**Marketer**	**ADULT**	**CUP**	**FARS**	**ATUS**	**IHIS**
k-Anonymity	**20%**	20%	20%	77.9%	86.0%	84.7%	91.3%	87.1%
k-Map (Houston)	(20%-50%)	**20%**	20%	83.3%	88.6%	86.8%	92.0%	87.1%
k-Map (Texas)	(33%-50%)	**20%**	20%	83.3%	88.6%	88.7%	94.5%	90.0%
k-Map (USA)	(50%)	**20%**	20%	83.3%	88.6%	88.7%	94.5%	92.3%
(k',k)-Strict-average risk	33%	33%	**20%**	80.8%	87.4%	86.8%	92.8%	90.0%
(k',k)-Strict-average risk	50%	50%	**20%**	83.8%	88.6%	88.7%	94.5%	92.3%

a risk threshold of 20% and an improvement of up to 7.4% could be achieved with a threshold of 10%. However, when assuming that the attacker is able to narrow down the individuals to the population of the city of Houston, no improvements could be measured. The degree of protection offered against prosecutor re-identification risks decreased with increasing data quality. For higher quality output data, residual prosecutor risks were 50%, which decreased to up to 10% with increasing loss of information.

Although *(k',k)*-strict-average risk offers weaker protection than *k*-map, we could only measure few significant improvements in data quality. A slight improvement could be achieved in cases where the focus lies on protection against marketer attacks and where it must be assumed that the attacker possesses high quality background knowledge (e.g. *(k',k)*-strict-average risk vs. *k*-map, *Houston* for IHIS). The residual risks measured for prosecutor attacks are comparable to the risks obtained with *k*-map, but risks for successful journalist attacks are increased.

The models evaluated previously offer protection against all three types of attackers, although to different degrees. With *p*-uniqueness protection can be focused on the marketer risk alone. ▶Figure 6 shows risk-utility frontiers obtained by de-identifying the datasets using the Hoshino model. Risk-utility frontiers are plots of re-identification risks versus data quality which illustrate the trade-offs a specific method provides between these two aspects [30]. Each point in such a plot represents a policy which offers an optimal trade-off. This means that risk cannot be reduced without reducing quality and that quality cannot be improved without increasing risk. As a baseline we use *2*-anonymity (50% risk threshold), which results in datasets without any unique records.

It can be seen that, compared to *k*-anonymity, significant improvements in data quality could be achieved. Using a risk threshold of 1%, which is still rather strict, quality improved by 4% – 17% compared to *2*-anonymity, by 7% – 20% compared to the baseline from ▶Table 2 and by 9% – 24% compared to the baseline from ▶Table 3. Even with stricter thresholds for re-identification risks significant improvements could be achieved. As expected, data quality decreased when more accurate background knowledge was assumed to be available to the attacker. The data also shows some irregularities. For example, at some measurement points data quality for the ATUS dataset was higher when more accurate background knowledge was considered. The reason for this behavior is that it is not always possible to automatically solve the equation system used by the Hoshino model [21]. However, this is a rare event which does not have any privacy implications but only leads to output data with non-optimal quality. The vertical lines in ▶Figure 6 indicate the point at which the fraction of population uniques drops to zero and the model thus offers weak protection against journalist attacks (50%

Table 3 Data quality and residual risks when de-identifying data using *k*-anonymity, *k*-map and *(k',k)*-strict-average risk with a threshold of 10%. Primary risk thresholds are highlighted. Risk thresholds in parenthesis are not defined by the user but result from the de-identification process.

	Risk Thresholds			Data Quality				
Privacy Model	**Prosecutor**	**Journalist**	**Marketer**	**ADULT**	**CUP**	**FARS**	**ATUS**	**IHIS**
k-Anonymity	**10%**	10%	10%	73.8%	82.4%	81.8%	89.0%	84.9%
k-Map (Houston)	(10%-50%)	**10%**	10%	83.8%	87.4%	86.8%	90.7%	84.9%
k-Map (Texas)	(33%-50%)	**10%**	10%	83.8%	88.6%	88.7%	94.5%	90.0%
k-Map (USA)	(50%)	**10%**	10%	83.8%	88.6%	88.7%	94.5%	92.3%
(k',k)-Strict-average risk	20%	20%	**10%**	77.9%	86.0%	84.7%	91.3%	87.1%
(k',k)-Strict-average risk	33%	33%	**10%**	80.8%	87.4%	86.8%	92.8%	90.0%

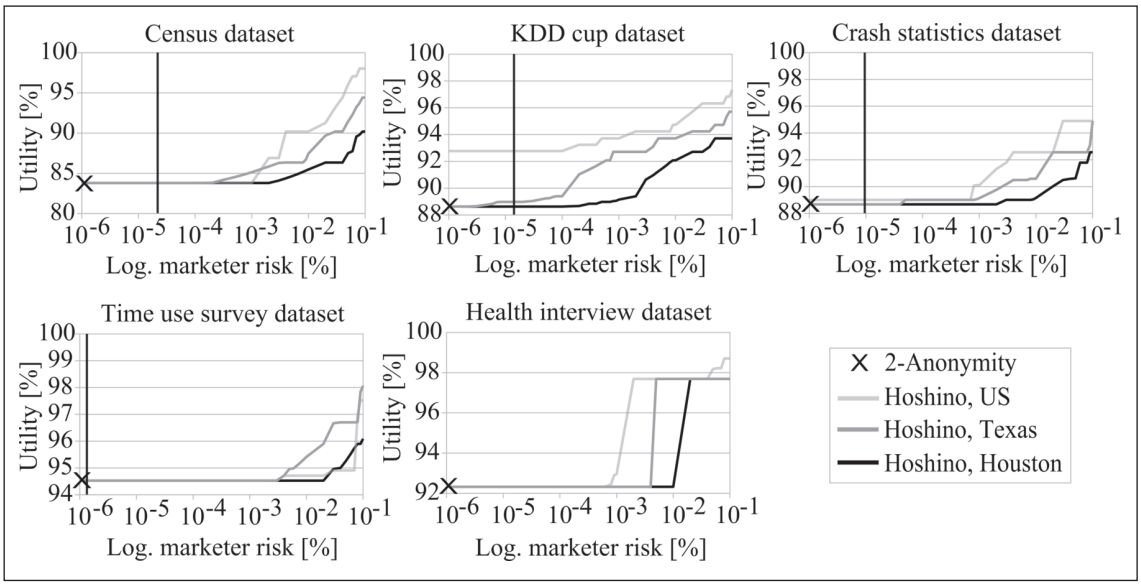

Figure 6
Semi-log plots comparing data utility measured for different de-identification data with p-uniqueness (Hoshino) and *2*-anonymity. The vertical lines indicate the point at which the fraction of population uniques drops to zero.

risk). In some cases small improvements in data quality could also be achieved with these strict thresholds. For example, de-identifying the CUP dataset with a threshold of 10^{-5} improved quality by 4% compared to *k*-anomymity, *k*-map and *(k',k)*-strict-average risk.

6. Discussion

In this article, we have presented a generic method for automatically de-identifying biomedical data with a broad spectrum of models for re-identification risks. We have implemented several such models and we have analyzed the degrees of protection they offer against different types of re-identification attacks. We have further performed extensive experiments which show the improvements in data quality that can be achieved by implementing a risk-based approach to data de-identification. Compared to *k*-anonymity, which is the model that provides the highest degree of protection, the loss of information content could be reduced by up to 10% when protecting datasets against strong adversaries and by up to 24% when protecting datasets against weaker adversaries.

The protection of personal health data against privacy threats is a challenging task involving multiple trade-offs. As mentioned previously, privacy risks can never be reduced to completely zero [11]. Therefore, data controllers must employ multiple layers of safeguards. A typical example are data use agreements, which are contracts that hold data recipients responsible to comply with relevant terms, conditions and regulations. Moreover, multiple fine-grained levels of access to sensitive data should be implemented and data access committees should be involved [31]. By controlling the context in which sensitive data is disclosed, reasonable assumptions can be made about possible attacks. In other words, depending on which additional safeguards are implemented, different options for data de-identification can be used [32]. For example, the HIPAA Privacy Rule permits the use of less strict methods of data de-identification if data use agreements are signed [14]. Our results can be used as a guideline for tailoring de-identification processes to a specific context in order to maximize data quality.

The risk estimates used in this article are based on worst-case assumptions [10] and there is evidence that re-identification risks are generally overestimated [33]. While the *k*-anonymity model has been recommended for public data disclosure [20], methods such as *k*-map and *(k',k)*-strict-average risk have been recommended for non-public data sharing [11, 29]. Super-population models are used by statistical agencies for the preparation of public use files [19] and population statistics have been used for defining the privacy requirements underlying the Safe Harbor method [34]. Models considering the relationships between a dataset and the underlying population can also be important for de-identifying data from small geographic regions, where datasets are often too sparse to be de-identified with strict models [35].

The basic steps of our risk-based de-identification algorithm (see Section 4.2) are similar to process steps suggested in other articles [24, 36]. However, previous work focused on abstract workflows for performing manual risk analyses. In contrast, we have presented the first software implementation which enables the use of risk models for automatically de-identifying data. Other de-identification tools, in particular sdcMicro [37] and μ-Argus [38], also implement models for estimating re-identification risks. However, in contrast to our work, these tools implement the models for performing accompanying risk analyses only and they do not provide means to directly use them for data de-identification. It has been shown that the model by Hoshino, which we have implemented into ARX, significantly outperforms the risk estimator implemented in μ-Argus [24].

All risk models described in this article have been recommended for de-identifying biomedical data [11, 29, 24, 20]. The transformation method used by our implementation produces datasets which are well

suited for analyses by epidemiologists and which is intuitive enough to enable non-IT experts to configure the de-identification process [39]. It has been demonstrated that modern analyses used in biomedical research, such as phenome-wide association studies, can effectively be performed with de-identified data [40]. The ARX data anonymization tool has been designed specifically for applications in the biomedical domain. While data can (and sometimes must) be protected from threats which go beyond re-identification (e.g. probabilistic inference of sensitive attribute values [41]), it is generally accepted that data de-identification is of central relevance [42]. However, the ARX system implements a wide variety of further privacy models and our implementation therefore supports combining risk-based data de-identification with other methods of data anonymization.

Our results show that the quality of anonymized data can significantly be improved by considering the context of data sharing for choosing appropriate privacy models. However, there are additional ways with which the quality of data can be optimized. First, further methods have been proposed for transforming data. The amount of information removed from a dataset can be reduced by not requiring all values of an attribute to be transformed to the same generalization level [43]. The most important methods in this context are local recoding, e.g. [44], and subtree generalization, which has been used to construct risk-utility frontiers for biomedical data in [16]. However, results obtained with these models are complicated to analyze [45]. Microaggregation is a pertubative transformation method which is specifically well suited for continuous variables, e.g. [43, 45]. However, it has been argued that pertubative methods cannot be used in biomedical research [46]. ARX supports local recoding as well as microaggregation. Second, it is also important to choose an appropriate model for measuring data quality. KL-Divergence [41] and Non-Uniform Entropy [39] are two important methods which have also been recommended for biomedical research [11, 16]. Both are implemented in ARX. We have performed our experiments with these models as well and we have measured improvements comparable to the numbers presented in Section 5. Finally, inherently different approaches to privacy-preserving data publishing can be used. With the Differential Privacy [47] method, privacy models are not applied to the data which is being released but to the *mechanism* with which it is being processed. The approach provides stronger privacy guarantees while requiring less assumptions to be made about likely attacks. However, Differential Privacy usually involves significant trade-offs in terms of supported workflows and it has been argued that it is not well suited for the biomedical domain due to its pertubative nature [46]. ARX also implements a differentially private data release mechanism.

7. Conclusion

The aim of this article was not to propose new privacy models, but to present a generic method that can be used to tailor de-identification processes to a concrete context. We have shown that our approach enables significant improvements in terms of data quality. The preconception that de-identification generally results in an unacceptable loss of information – and a lack of publicly available tools which demonstrate the opposite – is a major barrier for its broader adoption [48]. As a consequence, our open source implementation of the methods presented in this article is a valuable resource for fostering the use of de-identification methods in biomedical research.

References

1. Schneeweiss S. Learning from Big Health Care Data. N Engl J Med. 2014; 370(23): 2161–3. PubMed PMID: 24897079.
2. Murdoch T, Detsky A. The inevitable application of big data to health care. J Am Med Assoc. 2013; 309(13): 1351–2. PubMed PMID: 23549579.
3. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013; 31(12): 1102–10. PubMed PMID: 24270849.
4. Christoph J, Griebel L, Leb I, Engel I, Köpcke F, Toddenroth D, et al. Secure secondary use of clinical data with cloud-based NLP services. Methods Inf Med. 2015; 54(3): 276–82. PubMed PMID: 25377309.
5. US National Institutes of Health. NOT-OD-14–124: NIH Genomic Data Sharing Policy [Internet]. Genomic Data Sharing Policy Team; 2014 [cited 2016 Feb 04]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14–124.html.
6. Liu V, Musen M, Chou T. Data breaches of protected health information in the united states. J Am Med Assoc. 2015; 313(14): 1471–3. PubMed PMID: 25871675.
7. Hallinan D, Friedewald M, McCarthy P. Citizens' perceptions of data protection and privacy in Europe. Comp Law Sec Rev. 2012; 28(3): 263–72. doi: 10.1016/j.clsr.2012.03.005.
8. Schadt EE. The changing privacy landscape in the era of big data. Mol Syst Biol. 2012; 8: 612. PubMed PMID: 22968446.
9. Sweeney L. Computational disclosure control – A primer on data privacy protection [dissertation]. Cambridge (MA): Massachusetts Institute of Technology; 2001.
10. El Emam K. Guide to the de-identification of personal health information. 1st ed. Boca Raton: CRC Press; 2013.
11. El Emam K, Arbuckle L. Anonymizing health data: case studies and methods to get you started. 1st ed. Sebastopol: O'Reilly and Associates; 2014.
12. HIPAA administrative simplification statute and rules, 45 C.F.R. Parts 160, 162, and 164 (2013).
13. Samarati P. Protecting respondents' identities in microdata release. Trans Knowl Data Eng. 2001; 13(6): 1010–27. doi: 10.1109/69.971193.
14. US Health insurance portability and accountability act of 1996, Pub. L. 104–191, 110 Stat. 1936 (August 21, 1996).
15. 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 and on the free movement of such data, Official Journal L 281 , 23/11/1995 P. 0031 – 0050 (October 24, 1995).
16. Xia W, Heatherly R, Ding X, Li J, Malin BA. R-U policy frontiers for health data de-identification. J Am Med Inform Assoc. 2015; 22(5): 1029–41. PubMed PMID: 25911674.
17. Sweeney L. k-anonymity: A model for protecting privacy. Int J Uncertain Fuzz. 2002; 10(05): 557–70. doi: 10.1142/S0218488502001648.
18. El Emam K. Risk-based de-identification of health data. IEEE Security & Privacy. 2010; 8(3): 64–7. doi: 10.1109/MSP.2010.103.
19. Pannekoek J. Statistical methods for some simple disclosure limitation rules. Statistica Neerlandica. 1999; 53(1): 55–67. doi: 10.1111/1467–9574.00097.
20. El Emam K, Dankar FK. Protecting privacy using k-anonymity. J Am Med Inform Assoc. 2008; 15(5): 627–37. PubMed PMID: 18579830.
21. Hoshino N. Applying pitman's sampling formula to microdata disclosure risk assessment. J Off Stat. 2001; 17(4): 499–520.
22. Chen G, Keller-McNulty S. Estimation of identification disclosure risk in microdata. J Off Stat. 1998; 14(1): 79–95.
23. Rinott Y. On models for statistical disclosure risk estimation. In: Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality; 2003 Apr 7–9; Luxembourg; 2003.

24. Dankar FK, El Emam K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. BMC Med Inform Decis Mak. 2012; 12: 66. PubMed PMID: 22776564.
25. Prasser F, Kohlmayer F. Putting statistical disclosure control into practice: The ARX data anonymization tool. In: Gkoulalas-Divanis A, Loukides G, editors. Medical Data Privacy Handbook. New York: Springer; 2015. p. 111–48.
26. Iyengar V. Transforming data to satisfy privacy constraints. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2002 Jul 23–26; Edmonton, Canada. ACM; 2002. p. 279–88. doi: 10.1145/775047.775089.
27. Bayardo RJ, Agrawal R. Data privacy through optimal k-anonymization. In: Aberer K, Franklin MJ, Nishio S, editors: Proceedings of the 21st International Conference on Data Engineering; 2005 Apr 5–8; Tokyo, Japan. IEEE Computer Society; 2005. p. 217–28. doi: 10.1109/ICDE.2005.42.
28. Prasser F, Kohlmayer F, Lautenschlaeger R, Eckert C, Kuhn KA. ARX – A Comprehensive tool for anonymizing biomedical data. In: Proceedings of the AMIA 2014 Annual Symposium; 2014 Nov 15–19; Washington, DC, US. AMIA; 2014. p. 984–93. PubMed PMID: 25954407.
29. El Emam K, Malin BA. Appendix B: Concepts and methods for de-identifying clinical trial data. In: Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine, editor. Sharing clinical trial data: Maximizing benefits, minimizing risk. Washington (DC): National Academies Press (US); 2015. p. 1–290.
30. Cox LH, Karr AF, Kinney SK. Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act. Int Stat Rev. 2011; 79(2): 160–83. doi: 10.1111/j.1751–5823.2011.00140.x.
31. Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J Investig Med. 2010; 58(1): 11–8. PubMed PMID: 20051768.
32. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ. 2015; 350: h1139. PubMed PMID: 25794882
33. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PloS one. 2011; 6(12): e28071. Epub 2011 Dec 2. PubMed PMID: 22164229.
34. US Department of Health and Human Services – Office of the Assistant Secretary for Planning and Evaluation. Standards for Privacy of Individually Identifiable Health Information. Fed Regist. 2000; 65(250): 82462–829.
35. El Emam K, Brown A, AbdelMalik P, Neisa A, Walker M, Bottomley J, et al. A method for managing re-identification risk from small geographic areas in Canada. BMC Med Inform Decis Mak. 2010; 10: 18. PubMed PMID: 20361870.
36. El Emam K, Dankar FK, Vaillancourt R, Roffey T, Lysyk M. Evaluating the risk of re-identification of patients from hospital prescription records. Can J Hosp Pharm. 2009; 62(4). PubMed PMID: 22478909.
37. Templ M, Kowarik A, Meindl B. Statistical disclosure control for micro-data using the R package sdcMicro. J Stat Softw. 2015; 67(1): 1–36. doi: 10.18637/jss.v067.i04.
38. Hundepool A, Wetering A, Ramaswamy R, Franconi L, Polettini S, Capobianchi A, et al. Mu-Argus, Version 4.2 User's Manual [Internet]. The Hague, Netherlands: Statistics Netherlands; 2008 [cited 2016 Feb 04]. Available from: http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf.
39. El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E et al. A globally optimal k-anonymity method for the de-identification of health data. J Am Med Inform Assoc. 2009; 16(5): 670–82. PubMed PMID: 19567795.
40. Heatherly RD, Loukides G, Denny JC, Haines JL, Roden DM, Malin BA. Enabling genomic-phenomic association discovery without sacrificing anonymity. PloS one. 2013; 8(2): e53875. Epub 2013 Feb 6. PubMed PMID: 23405076.
41. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. Trans Knowl Discov Data. 2007; 1(1): 3. doi: 10.1145/1217299.1217302.
42. McGraw D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. J Am Med Inform Assoc. 2013; 20(1): 29–34. PubMed PMID: 22735615.
43. Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min. Knowl. Discov. 2005; 11(2): 195–212. doi: 10.1007/s10618–005–0007–5.
44. Goldberger J, Tassa T. Efficient anonymizations with enhanced utility. In: Saygin Y, Xu Yu J, Kargupta H, Wang W, Ranka S, Yu PS, Wu X, editors: Proceedings of the ICDMW'09 IEEE International Conference on Data Mining Workshops; 2009 Dec 6; Miami, USA. IEEE Computer Society; 2009. p. 106–13. doi: 10.1109/ICDMW.2009.15.
45. Soria-Comas J, Domingo-Ferrer J, Sanchez D, Martinez S. t-Closeness through microaggregation: strict privacy with enhanced utility preservation. Trans Knowl Data Eng. 2015; 27(11): 3098–110. doi: 10.1109/TKDE.2015.2435777
46. Dankar FK, El Emam K. Practicing differential privacy in health care: A Review. Trans Data Priv. 2013; 6(1): 35–67.
47. Dwork C. Differential privacy. In: Bugliesi M, Preneel B, Sassone V, Wegener I, editors: Proceedings of the 33rd International Colloquium; ICALP 2006 Jul 10–14; Venice, Italy. Berlin; Heidelberg: Springer; 2006. p. 1–12. doi: 10.1007/11787006_1.
48. El Emam K, Álvarez C. A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques. Int Data Priv Law. 2015; 5(1): 73–87. doi: 10.1093/idpl/ipu033.