International Journal of Medical Informatics 126 (2019) 72–81

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Privacy-enhancing ETL-processes for biomedical data

Fabian Prasser[1,*], Helmut Spengler[1], Raffael Bild, Johanna Eicher, Klaus A. Kuhn

Institute of Medical Informatics, Statistics and Epidemiology, University Hospital rechts der Isar, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany

ARTICLE INFO

Keywords:
Clinical data warehousing
Extract Transform Load
Privacy
Anonymization

ABSTRACT

Background: Modern data-driven approaches to medical research require patient-level information at comprehensive depth and breadth. To create the required big datasets, information from disparate sources can be integrated into clinical and translational warehouses. This is typically implemented with Extract, Transform, Load (ETL) processes, which access, harmonize and upload data into the analytics platform.

Objective: Privacy-protection needs careful consideration when data is pooled or re-used for secondary purposes, and data anonymization is an important protection mechanism. However, common ETL environments do not support anonymization, and common anonymization tools cannot easily be integrated into ETL workflows. The objective of the work described in this article was to bridge this gap.

Methods: Our main design goals were (1) to base the anonymization process on expert-level risk assessment methodologies, (2) to use transformation methods which preserve both the truthfulness of data and its schematic properties (e.g. data types), (3) to implement a method which is easy to understand and intuitive to configure, and (4) to provide high scalability.

Results: We designed a novel and efficient anonymization process and implemented a plugin for the Pentaho Data Integration (PDI) platform, which enables integrating data anonymization and re-identification risk analyses directly into ETL workflows. By combining different instances into a single ETL process, data can be protected from multiple threats. The plugin supports very large datasets by leveraging the streaming-based processing model of the underlying platform. We present results of an extensive experimental evaluation and discuss successful applications.

Conclusions: Our work shows that expert-level anonymization methodologies can be integrated into ETL workflows. Our implementation is available under a non-restrictive open source license and it overcomes several limitations of other data anonymization tools.

1. Introduction

Modern medical research requires data of comprehensive depth and breadth to improve our understanding of the development and course of diseases and to ultimately develop methods for prevention, targeted diagnosis and therapy. In a learning health system "every clinical encounter contributes to research and research is being applied in real time to clinical care" [1]. To implement this on a large scale, data must be made accessible, harmonized and integrated [2,3]. This also requires using data for secondary applications that go beyond the initial purpose of collection [4,5].

Data integration and in particular data warehouses are central to these efforts. In this context, database systems are set up that integrate disparate data into a common layout which efficiently supports complex analyses. The i2b2 platform [6] is a well-known example of a system that focuses on data generated by clinical and health services and by epidemiological studies [7]. A related platform is tranSMART, which has been developed for the analysis of integrated clinical and 'omics' data for translational research [8]. Some institutions, such as the Vanderbilt University Medical Center [5], have also developed custom solutions.

Data is typically replicated from routine systems into warehouses using ETL processes [9,10]: (1) data is *extracted* from source systems, (2) cleansed, harmonized and *transformed* into a form suitable for analyses, and (3) *loaded* into the analytics solution. To manage the complexity of such processes, they are often implemented using specific environments, which offer libraries of connectors to different types of sources, transformation operators and a graphical workbench for

* Corresponding author.
E-mail address: fabian.prasser@tum.de (F. Prasser).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ijmedinf.2019.03.006
Received 22 June 2018; Received in revised form 6 November 2018; Accepted 6 March 2019

combining them into complex workflows. Well-known solutions are Pentaho Data Integration (PDI, also known as Kettle) [11], which is the standard tool for loading data into tranSMART, and Talend Open Studio (TOS) [12], which is a central component of the Integrated Data Repository Toolkit (IDRT) [13] for creating i2b2-based warehouses.

When pooling medical data or when re-using it for secondary purposes, privacy concerns and legal requirements need careful consideration. Privacy protection involves ethical, legal and societal issues (ELSI) and several layers of technical and non-technical measures are typically required to implement it [14]. On the technical side, the privacy of patients and probands is often protected by data anonymization, which means that datasets are altered in a way that prevents successful re-identification. National and international privacy regulations address data anonymization. In the United States, the *Safe Harbor* method of the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) provides a catalog of attributes for which values need to be removed or modified [15]. In addition, the *Expert Determination* method permits the use of formal and statistical methods for assessing and managing re-identification risks, which is similar to the way in which data anonymization needs to be implemented in the European Union [16].

Data anonymization is a complex process in which the resulting reduction of re-identification risks needs to be balanced against a reduction of data utility [14,17]. A wide variety of different models and methods for data transformation, risk assessment and utility estimation have been proposed to address this trade-off. To manage this complex process, a number of tools have been developed, including sdcMicro [18], which focuses on official statistics, and ARX [19], which has specifically been designed for applications to biomedical data by implementing methods which have been recommended in the field [19–21]. Both tools offer a high level of maturity and they have been included into official guidelines, e.g. from the European Union Agency for Network and Information Security (ENISA) [22] and the European Medicines Agency (EMA) [23].

1.1. Objectives and outline

Performing data anonymization and re-identification risk analyses as part of ETL workflows is a common requirement (see Section 4.1). Typical application scenarios include the loading of data into clinical and translational warehouses, the extraction of data from cross-institutional research registries, and the sharing of data with external research groups. However, ETL platforms such as TOS or PDI do not provide modules which support formal methods of data anonymization and re-identification risk analysis. Although anonymization tools such as sdcMicro or ARX can be used for these purposes, they are based on their own working environments which cannot easily be integrated into ETL platforms (see Section 2).

To bridge this gap, we have developed a plugin for an ETL platform, which supports data anonymization and re-identification risk assessment. The most important design goals were (1) to utilize expert-level risk assessment methodologies, (2) to implement a data transformation method which preserves both the truthfulness of input data and its schematic properties (e.g. data types), (3) to utilize an anonymization process which is easy to understand and intuitive to configure, and (4) to achieve high scalability.

To meet these design goals we had to overcome various challenges. First, we needed to decide on a suitable design and execution environment for ETL processes. Second, we needed to select and integrate methodologies for risk assessment and anonymization which are well-known, flexible and easy to understand. This involved managing the complex interplay of methods for measuring and reducing privacy risks. Finally, we had to develop an efficient implementation.

The remainder of this paper is structured as follows: in Section 2 we describe the methods for risk assessment that we build upon, present a novel anonymization method, and describe how we have implemented and integrated it into an existing ETL platform. In Section 3 we describe how we designed our experiments and present the results. In Section 4 we discuss the principal results, applications in practice, and perform a conceptual comparison with prior work. In Section 5 we conclude and point out directions for future work.

2. Materials and methods

Our method for integrating data anonymization and risk assessment into ETL processes is based on established methods for estimating re-identification risks of medical data, which we present in the first part of this section. In the second part we present a novel anonymization algorithm which we have developed in order to facilitate an effective integration of these methods into ETL platforms. The last part of this section focuses on how we implemented these methods and how we integrated them into a concrete ETL platform.

2.1. Common models for risk assessment

Re-identification is the primary threat addressed by laws and regulations [15,16] and models for quantifying related risks are therefore central to data anonymization and privacy risk management. Re-identification can be understood as a *linkage process* [24]: the uniqueness of (combinations of) attributes is exploited to link records of datasets with additional data or background knowledge of the adversary. Attributes that can be used for establishing a link are termed *quasi-identifiers* [25]. Typical examples include demographic data and other information that is likely to be known to adversaries, such as educational or employment status [21]. Implementing protection requires to consider various factors, e.g. the objectives of likely attackers, the replicability and distinguishability of the data to be protected, and the availability of background knowledge [26,27].

Three different threat scenarios can be distinguished [28]. Under the *prosecutor* model, the adversary is assumed to target a specific individual and to know that data about this individual is contained in the dataset. The risk of a successful attack can be calculated, based on the distinguishability of records in the dataset regarding the quasi-identifiers [26]. It has been shown, however, that this method significantly over-estimates risks in most cases [29]. Under the *journalist* model, the adversary is assumed to target an arbitrary individual without prior knowledge about membership. Often, this background knowledge is much more realistic than in the prosecutor model, as the set of individuals represented in a dataset is just a sample of a larger population. However, the fact that knowledge about the population is typically not available makes it also difficult to reliably determine and manage the risk of successful journalist attacks. Finally, under the *marketer* model, the adversary is assumed to aim at re-identifying as many individuals as possible. Thus the risk of a successful attack can be expressed as the expected average number of re-identified individuals.

El Emam has proposed a methodology that combines estimates of risks under these established models [28]. As journalist risk cannot be quantified in most cases, the methodology makes use of the fact that prosecutor risk is always an upper bound for journalist and marketer risk. Prosecutor risk is quantified for all records and aggregated into three global measures. The first measure is the *Highest Risk* (R_h). It quantifies risks in the worst case scenario, i.e. a prosecutor attack against the record with the highest re-identification risk in the whole dataset. For each record r, the re-identification risk is calculated as $\frac{1}{f_r}$, where f_r is the number of records in the dataset that are indistinguishable from r regarding the quasi-identifiers (including r itself). As noted before, this is also an upper bound for risks in the other scenarios, i.e. for journalist or prosecutor attacks. Even when this risk is bound by a threshold, an attacker can expect to re-identify a certain number of individuals by random linkage to matching records. This is captured by the second measure, *Average Risk* (R_a), which provides a

more tight bound for marketer risks. To account for the fact that the prosecutor model is based on worst-case assumptions, a third measure, called *Records at risk (R_r)* can be used to slightly relax the protection requirements. It expresses the frequency of records that are associated with a re-identification risk higher than a given threshold θ. Formal definitions of these three risk measures are provided in Section A of the supplementary file.

With this methodology and just a single user-specified parameter (θ), three intuitive risk measures can be derived that quantify the susceptibility of data to all types of attacks considered. At the same time, the model facilitates a balancing of privacy protection and the usefulness of data, as it enables the user to permit that a fraction of records has a risk that is higher than the threshold θ. Given a sufficiently small θ and a sufficiently small fraction of records at risk, a high degree of protection can be assumed, as it is very unlikely that the record targeted in a (prosecutor or journalist) attack is one of the records that exceeds the threshold [28]. Thresholds τ_a for the average risk R_a and τ_h for the highest risk R_h can be introduced in addition to θ to specify protection levels which must be satisfied by a data anonymization procedure.

2.2. A novel anonymization method

Automatically altering data such that it meets user-specified risk thresholds is complex and requires integrating risk models with data transformation techniques and methods for measuring data utility. Producing truthful output data implies that input data is not perturbed and that no synthetic data is generated, which is particularly important in medical research where plausibility and correctness are central [30]. Therefore, we decided against transformation schemes which employ noise addition [31] or aggregation of data [32]. Moreover, we wanted to ensure that our method can be integrated into existing ETL workflows without the need to modify intermediate or target data representations. This implies that schematic properties of input data must be preserved, which means that data types must not be altered and that no additional attributes must be introduced into the tables and rows processed. Thus we could not use data generalization [25] or bucketization [33].

Based on these considerations, we decided to implement a cell suppression algorithm. With this model, risk thresholds are enforced by removing individual attribute values from individual records. The method requires zero configuration (apart from specifying risk thresholds), output data is truthful and schematic properties are being preserved. Moreover, the results are well suited for performing common statistical analyses, provided that the effect of cell suppression is considered (e.g. by imputation) [28,30,34].

Fig. 1 shows how cell suppression can be used to protect a dataset from two different threat scenarios. In this simplified example, a clinical dataset is protected from marketer attacks by *external attackers* using the demographic attributes {*Age*, *Sex*, *Region*} and from prosecutor attacks by *internal attackers* using the clinical attributes {*Weight*, *ICD-10*}. Suppressed values (which are denoted by *) are treated as an own category, which means that suppressed values are only considered to be equal to other suppressed values. Under this assumption all sets of rows containing the same quasi-identifying attribute values are pairwise disjoint and form so-called *equivalence classes*. Each equivalence class describes a set of records which are indistinguishable to the attacker and hence its size determines the risk of successful re-identification. In the example, equivalence classes are illustrated by dotted lines. By suppressing 20 of the 50 attribute values in the dataset (40%), the risk of a successful external attack dropped from 60% ($R_a = \frac{6}{10}$) to 30% ($R_a = \frac{3}{10}$) and the risk of a successful internal attack dropped from 100% ($R_h = \frac{1}{1}$) to 33% ($R_h = \frac{1}{3}$). The example also shows that cell suppression is challenging to implement efficiently, as the space of potential solutions for a given dataset consists of $O(2^{n \cdot m})$ transformations where n is the number of records and m is the number of attributes that could be used for linkage. This equates to 2^{50} potential solutions already in our simple example. Thus cell suppression is typically performed using heuristic algorithms.

Our implementation follows this approach by recursively enforcing the user-defined thresholds τ_a and τ_h for subsets of the input dataset. This is implemented with ARX, which is able to compute an optimal solution to a data anonymization problem that is specified as follows [35]: (1) all risk thresholds must be met, (2) each column that contains quasi-identifying values may either be kept as-is or suppressed entirely (attribute suppression), (3) a specified number of records may be entirely suppressed (called the *suppression limit*), (4) the overall number of suppressed cells must be minimal. Our method executes this process recursively for the records that have been suppressed, as is illustrated in Fig. 2. In each iteration, τ_h and τ_a are enforced on a set of the records; the others are suppressed. We use the *k*-anonymity privacy model to enforce τ_h [25] and enforce τ_a by specifying an upper bound on the arithmetic mean of the records' re-identification risks. An additional parameter l_s specifies the maximum number of recursive calls by defining a suppression limit for each iteration. Pseudocode illustrating the anonymization method in more detail and a discussion of implications for data quality is provided in Section B of the supplementary file. While this process is very efficient and effective, as we will show in the next section, it remains necessary to show that it is actually correct. It is easy to see that enforcing an overall threshold on the highest re-identification risk R_h can be performed by enforcing the same threshold on disjoint subsets of records. However, it is not trivial to see that this process can be used to implement a global threshold on the average re-identification risk R_a. A proof is provided in Section C of the supplementary file.

2.3. Implementation and integration

To make our solution accessible to a broad spectrum of users we

Age	Sex	Region	Weight	ICD-10
53	F	North	73	C18.7
68	F	North	73	C18.7
68	M	North	82	C18.7
68	M	North	77	C18.7
71	M	North	73	C18.2
71	M	North	67	C18.2
68	M	South	67	C18.2
68	F	South	67	C18.7
68	F	South	67	C18.7
68	F	South	67	C18.7

(a) Input dataset

Age	Sex	Region	Weight	ICD-10
*	*	North	*	C18.7
*	*	North	*	C18.7
*	*	North	*	C18.7
*	M	North	*	C18.7
*	M	North	*	C18.2
*	M	North	*	C18.2
68	*	South	*	C18.2
68	*	South	67	C18.7
68	*	South	67	C18.7
68	*	South	67	C18.7

(b) Output dataset

Fig. 1. Example dataset before (a) and after (b) it has been transformed using cell suppression. Dotted lines illustrate equivalence classes with respect to two different sets of quasi-identifiers: {*Age*, *Sex*, *Region*} and {*Weight*, *ICD-10*}.

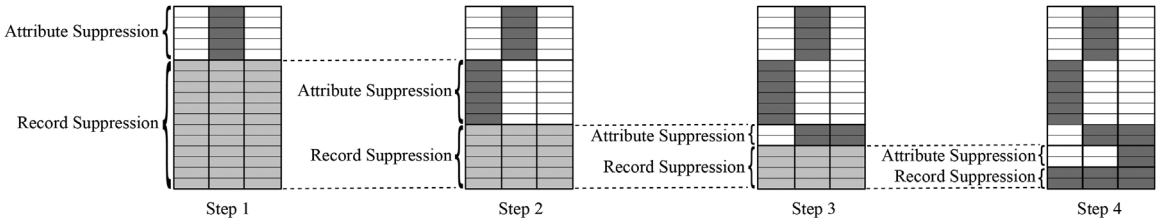


Fig. 2. Illustration of the recursive cell suppression algorithm. In each of the recursion steps the algorithm determines the optimal balance between attribute and record suppression.

decided on a two-step implementation strategy. In the first step, the described anonymization and risk assessment methodology was implemented into ARX. This allowed us to leverage its highly scalable anonymization framework [19] to create a risk assessment and anonymization operator which can then be integrated into ETL environments in the second step. In this context, we decided to develop a plugin for the PDI platform for several reasons. First, we frequently use PDI for loading data into tranSMART. Second, the interface provided by PDI is quite intuitive while the learning curve for TOS can be considered to be somewhat steeper. Third, PDI offers a broad set of features in its community edition (e.g. deployment to clusters) while most advanced features of TOS are only available through a commercial license. Moreover, with the recent release (version 8.0), the programming interfaces of the PDI platform have received significant modernization.

In the PDI workbench, ETL processes can be modeled as directed graphs, where data sources, transformations, and data sinks are represented as nodes called “steps”. Data flow between nodes is represented by edges. Data that could not be processed can be annotated with additional information and routed to a dedicated error output. By combining multiple steps, complex ETL processes integrating heterogeneous sources can be designed, executed and monitored. Fig. 3 shows a screenshot of an ETL process in which data from three different data sources (a CSV file, a relational database, and a HL7 message stream) are joined, validated, transformed, and finally loaded into a target database.

Data processing in PDI is stream-oriented with single rows of data constituting atomic and isolated units of a data stream. This means that data is passed through the ETL pipeline row by row. This enables *pipeline parallelism* across a chain of steps. However, it also implies that plugins that require a holistic view on the overall dataset, such as our plugin for assessing risks or anonymizing data, need to buffer the incoming rows. There are trade-offs involved in implementing this, as the latter breaks pipeline parallelism and high volume datasets can be too large to completely materialize them in main memory.

To solve this issue, we implemented a technique called *row blocking*. This means that our plugin materializes sets of records (i.e. blocks) of a user-defined size, which are then analyzed or anonymized. As soon as each block has been processed, the contained rows are passed on to the next plugin in the workflow. As a consequence, parallelism can be maintained and very large datasets can be processed. In terms of privacy protection, the approach is guaranteed to be correct (see Section C of the supplementary file).

We implemented all methods into a plugin for the PDI platform. Our implementation is available as open source software [36,37] which is compatible with the latest version 8.0 of PDI. The plugin provides methods for re-identification risk analyses and data anonymization. It is compatible with all other functionalities and plugins of PDI.

The tab *Risk thresholds*, which is shown in Fig. 4(a), enables users to specify quasi-identifiers and the thresholds described previously. Compatible to the relational model underlying the ETL environment, values that are suppressed are replaced with *NULL*. Thus the schema and data types of input data are preserved. When risks are assessed and any of them exceeds a user-defined threshold, the incoming data will not be transferred to the subsequent step and, if desired, it can be routed to an error exit. Risk measures are printed to the console for logging purposes. The tab *Runtime settings*, which is shown in Fig. 4(b), can be used to specify parameters affecting the runtime behavior of the anonymization algorithm.

To address multiple threat scenarios, data can be passed through different instances of the plugin configured to address different threat scenarios (cf. example in Fig. 1). This is possible because the plugin preserves the schematic properties of input data and because it makes

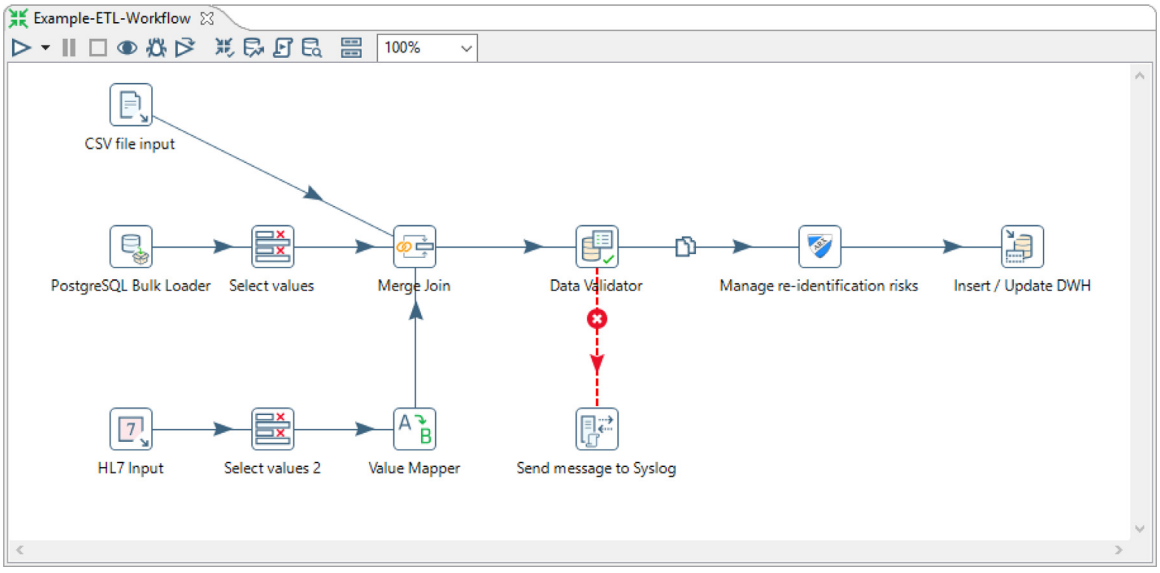
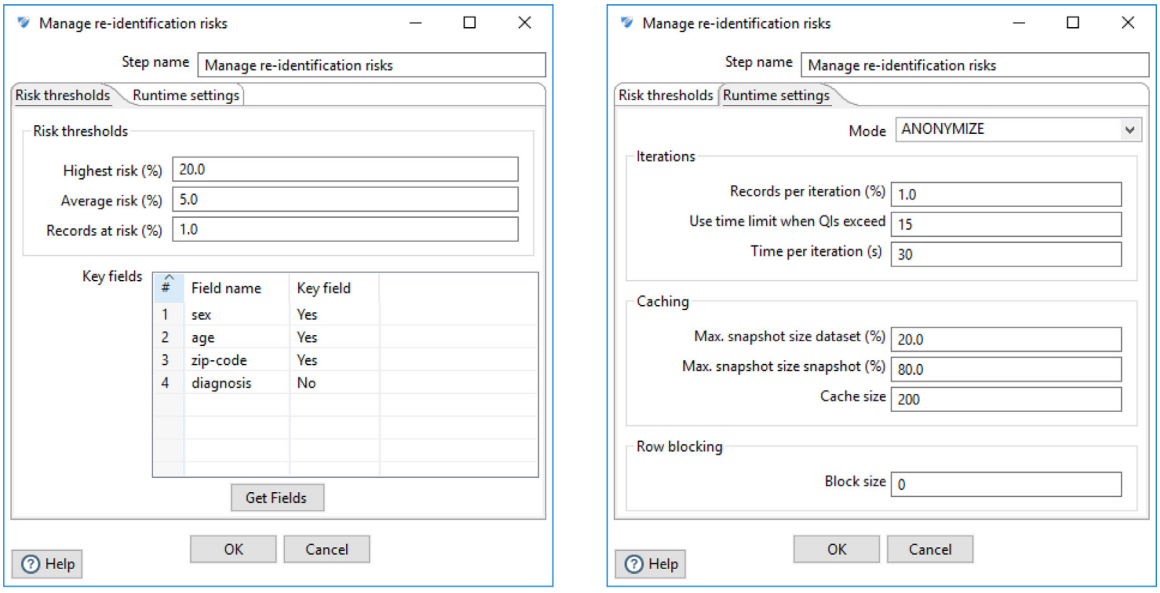


Fig. 3. A typical ETL process in PDI's design environment Spoon.



(a) Privacy related parameters. (b) Parameters determining runtime behavior.

Fig. 4. Screenshots of the plugin's configuration dialogs.

use of different ways of interpreting suppressed values. During anonymization, suppressed values are treated as an own category, meaning that *NULL* only matches *NULL* when calculating the distinguishability of records. However, in a chain of anonymization steps with *overlapping* quasi-identifiers, this can lead to situations, in which one anonymization operation invalidates the privacy guarantees that have been enforced in previous steps because new categories are introduced into quasi-identifying variables addressed previously (an example can be found in Section D of the supplementary file). For this reason, when *assessing* risks, our plugin interprets suppressed values as wild cards. This means that they can match any other (suppressed *or* unmodified) value, which avoids this problem. While it has been shown that this interpretation can provide adversaries with attack vectors under rare circumstances [38], we point out that this is the standard interpretation in the field of statistical disclosure control and also the default in sdcMicro.

3. Results

3.1. Experimental setup

In this section, we present results of evaluating the scalability of our solution as well as the quality of output data, including comparisons with prior work. We point out that a theoretical bound on the data quality provided by our approach cannot easily be obtained (for a discussion of optimality aspects we refer to Section B of the supplementary file). Hence, we focus on an experimental evaluation with real-world datasets to analyze how the method performs in practice. We performed four different sets of experiments:

- **Comparison with prior work**: We first compared the performance of our plugin to sdcMicro (version 5.0.3) [18], which features a cell suppression algorithm that has been implemented in C++ and linked into the software. Next, we studied the utility of output data produced by our cell suppression method in comparison to other data transformation methods using the concept of privacy-preserving data cubes proposed by Kim et al. [39].
- **Comparison using different threat scenarios**: sdcMicro and the work by Kim et al. focus on simple threat scenarios, while our approach supports combinations of several different risk thresholds. We performed additional experiments using various parameterizations and measured output data quality to study their effects.
- **Analysis of risk-utility trade-offs**: In the third set of experiments we constructed risk-utility frontiers, which are plots visualizing the trade-offs that an anonymization method provides between privacy protection and data quality [40].
- **Analysis of the effect of row blocking**: The parameter that specifies the block size has various influences on the quality of output data and the execution time of the anonymization process. In a final set of experiments we studied these effects to determine whether row blocking is an effective mechanism for processing large datasets with our plugin.

We used two datasets, which differ in scope and size and which have already been utilized for evaluating previous work on data anonymization: (1) *US Census*, an excerpt of 30,162 records from the 1994 census database, which serves as the de-facto standard for the evaluation of anonymization algorithms, and (2) *Health Interviews*, a set of 1,193,504 responses to a large health survey. For a detailed description we refer to [41]. For each dataset we selected up to nine quasi-identifiers, consisting of demographic data and further attributes, which are often considered to be associated with a high risk of re-identification [21]. All experiments were performed on a desktop machine equipped with a quad-core 3.2 GHz Intel Core i5 CPU running a 64-bit Windows 7 operating system. The PDI platform (version 8.0) was executed using a 64-bit Oracle JVM (1.8). The number of iterations performed by our algorithm (parameter l_s) was set to 100 in all experiments.

3.2. Experimental comparison with prior work

We first compared our plugin to sdcMicro [18]. The cell suppression algorithm of sdcMicro has been implemented in C++ and linked into the package to improve scalability. The software only supports cell suppression for enforcing a threshold on the highest risk. Hence we set τ_r (records at risk) to zero and used a threshold on the prosecutor re-identification risk (τ_h) of 20%, which is a common parameterization [21].

Fig. 5(a) shows the execution times measured while increasing the number of quasi-identifying attributes. It can be observed that our implementation is significantly more scalable than sdcMicro. While our method was able to easily handle the US Census dataset regardless of

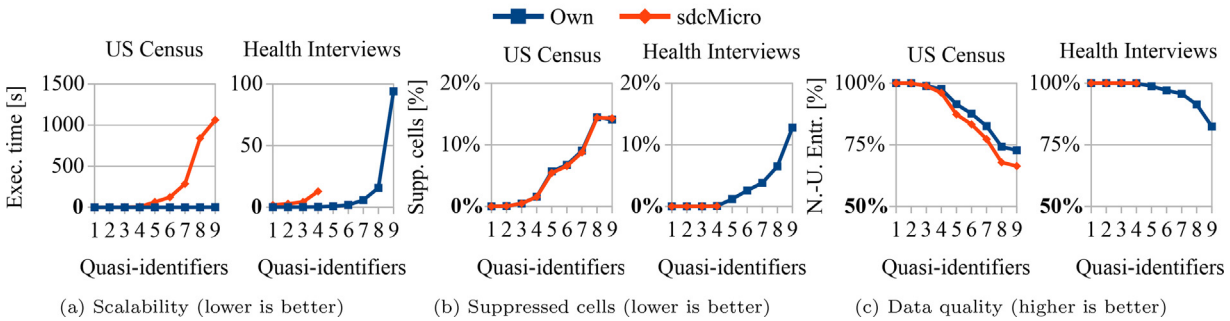


Fig. 5. Comparison of the results obtained with our plugin and the results obtained using sdcMicro. We report average execution times, the number of suppressed cells and data quality quantified with the Non-Uniform Entropy model.

the number of quasi-identifiers selected (≤2 s in all configurations), sdcMicro already needed more than 1000 s to process the dataset with nine quasi-identifiers configured. Furthermore, sdcMicro was not able to handle the Health Interviews dataset within 1800 s when more than four quasi-identifiers were specified. For practical reasons, we cancelled all experiments using sdcMicro that did not complete within this time frame. Our plugin generally processed this dataset in not more than 94 s. It can be seen that both implementations were affected by the exponential increase in the size of the solution space with an increasing number of quasi-identifiers [35]. However, our plugin can be configured to use an effective heuristic algorithm when the solution space becomes too large [42].

Regarding data quality, we measured comparable numbers of cells suppressed by our method and by sdcMicro (Fig. 5(b)). Finally, Fig. 5(c) shows how anonymization has impacted the distributions of attribute values. To measure this, we used the Non-Uniform Entropy model [43] which is often used to assess the quality of de-identified data and is based on the concept of mutual information [28]. We normalized the results obtained by this model in such a way that 100% represents the original input dataset while 0% represents a dataset from which all values have been removed. It can be seen that data quality decreased when the number of quasi-identifiers increased, especially for the smaller dataset US Census. It can also be observed that our method had less impact on the distribution of attribute values, implying a more balanced application of value suppression.

Recently, Kim et al. performed an experimental evaluation of the effects of different data anonymization methods when implementing privacy-preserving warehouses for medical data [39]. In their study, data was anonymized and then aggregated into data cubes, which is a model used in warehousing applications. The authors then measured the information loss induced by the anonymization methods and the precision of the results of two types of queries issued against the data cubes: *point queries*, which count the number of records matching a specific combination of attribute values and *range queries*, which count the number of records matching a combination of ranges over the domain of attribute values. They studied two generalization-based approaches and one bucketization algorithm.

We exactly reproduced their experimental setup, which also used the US Census dataset, and compared results obtained using our method with the results presented in [39]. For an exact specification of the algorithms and an in-depth discussion of the results we refer to Section E of the supplementary file. As can be seen in Table 1, our method outperformed both generalization-based approaches in terms of information loss, performed very well on point queries and provided reasonable performance on range queries. At the same time, our method is the only approach considered in the experiments that preserves the schematic properties of input data, and it is much easier to configure than generalization-based algorithms.

3.3. Experimental analysis using different threat scenarios

Our plugin supports thresholds on prosecutor re-identification risk (τ_h) and marketer re-identification risk (τ_a). *Strict-average risk* [21] is a common privacy model combining both risk thresholds. To analyze the improvements in data utility that can be obtained by using this model, we have performed a comparison of both approaches. As a risk threshold, we also used 20%. We used the same threshold once for controlling prosecutor risk and once for controlling marketer re-identification risk but combined the latter with a threshold of 50% on prosecutor risk, which ensures that no record is uniquely identifiable. We note that this comparison focused on our plugin only, as strict-average risk is to our knowledge not supported by any other tool.

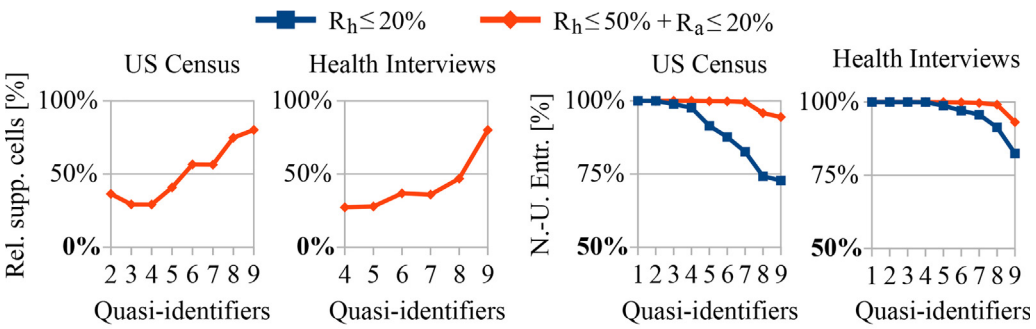
We measured no significant differences in execution times when using the two models. We did, however, observe notable improvements in data quality when using strict-average risk. Fig. 6(a) shows the number of suppressed cells when enforcing the thresholds on strict-average risk relative to the number of suppressed cells when enforcing the threshold on prosecutor risk. It can be seen that using strict-average risk resulted in significantly less suppressed cells, especially when configurations with fewer quasi-identifiers were being used. Effects on the distribution of attribute values are presented in Fig. 6(b). In contrast to the effect on the number of suppressed cells, the improvements obtained in terms of Non-Uniform Entropy increased with the number of quasi-identifiers. This implies that data quality can be more effectively increased by using less strict privacy models when it must be assumed that the adversary possess a lot of background knowledge.

3.4. Experimental analysis of the risk-utility trade-off provided

Our plugin provides a broad spectrum of anonymization options, ranging from very strict to very relaxed parameterizations. To analyze these different options in more detail, we constructed risk-utility

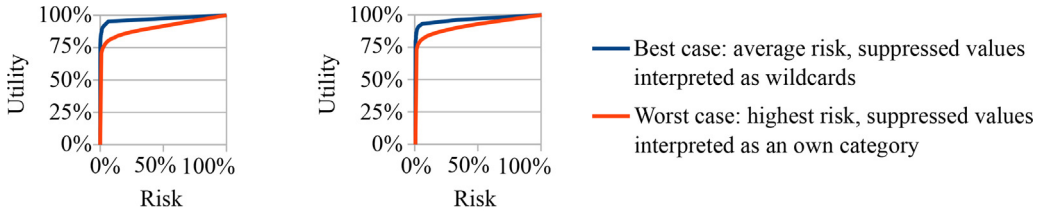
Table 1
Comparison of methods for creating privacy-preserving data cubes as proposed by Kim et al. [39].

	Global generalization	Local generalization	Bucketization	Cell suppression
Information loss	0.41	0.13	Not applicable	0.10
Median relative error for point queries (%)	18.3	9.79	0.02	0.00
Median relative error for range queries (%)	10.16	0.81	0.02	41.33



(a) Suppressed cells (lower is better) (b) Data quality (higher is better)

Fig. 6. Comparison of the results obtained when only enforcing a threshold on prosecutor risk with the results obtained enforcing a threshold on strict-average risk. We report the number of suppressed cells relative to the numbers obtained using the prosecutor model for cases in which at least one cell was suppressed. Data quality was quantified using the Non-Uniform Entropy model.



(a) US Census (b) Health Interviews

Fig. 7. Risk-utility frontiers for different risk models and different interpretations of missing values.

frontiers, which are plots visualizing the trade-offs that an anonymization method provides between privacy protection and data quality [40]. Each point in these plots represents a transformed dataset offering an optimal privacy/utility trade-off, which means that risk cannot be reduced further without reducing quality and vice versa. Fig. 7 shows the results of our method for both datasets using two extreme configurations addressing all quasi-identifiers. In the best case scenario, thresholds on the average risk R_a have been enforced while interpreting missing values as wild cards. In the worst case scenario, thresholds on the highest risk R_h have been enforced while treating missing values as an own category. Data utility was estimated with the relative number of cells that have *not* been suppressed.

As can be seen, we could not measure any significant differences between the results for the two datasets. In both cases, we observed that high data quality can be maintained at very low risk levels. The frontiers for the best case scenarios were almost optimal. Here, we measured an area under the curve (AUC, 1 optimal, 0 worst) of 0.971 for the US Census dataset and 0.966 for the Health Interviews dataset. In the worst case scenarios we measured AUCs of 0.901 and 0.912, respectively.

3.5. Experimental analysis of the effect of row blocking

Next, we investigated the effect of row blocking on execution times and on output data quality. The experiments were performed with nine quasi-identifying attributes and the same risk models and thresholds as in the previous experiments while varying the *block size*. Previously, we did not use row blocking and were thus able to only report the time needed to anonymize the data. In the results presented here, execution times include the time needed to read the data from disk, anonymize it and persist the results on disk.

As can be seen in Fig. 8(a), execution times decreased with increasing block sizes up to a block size of roughly 10^5, from where on they slowly increased again. This increase can be explained by the fact that much larger data volumes needed to be processed in each anonymization operation. For strict-average risk and block sizes between 10^2 and 10^3, we also observed an increase of execution times. This can be explained by the fact that this setup significantly increased the number of invocations of the underlying anonymization algorithm. Although each invocation had to handle a smaller number of records, the complexity of the anonymization problem with respect to the number of quasi-identifiers remained constant. Moreover, anonymizing fewer records tends to be more computationally expensive, as good solutions are harder to find [35]. Regarding the number of suppressed cells and effects on data quality, when increasing block sizes, we measured a logarithmic decrease (Fig. 8(b)) and increase (Fig. 8(c)), with values converging towards the baselines (dotted) obtained without row blocking. With block sizes of about 10^4 (US Census) and 10^5 (Health interviews) or bigger, the effects of row blocking on output data quality were almost negligible compared to anonymization without row blocking. This indicates that row blocking can be used to effectively balance data quality and execution times when processing large datasets.

4. Discussion

4.1. Principal results and applications in practice

In this article, we have presented a plugin supporting integrated data anonymization and re-identification risk analysis during ETL processes. Our implementation is based on the PDI platform, which is in widespread use within the biomedical field. The methods presented in this paper have also been implemented directly into ARX [19]. The risk assessment methodology described is robust, easy to configure and it provides a good balance between simple but strict approaches such as *k*-anonymity [25] and more flexible but complex models that provide higher degrees of output data quality (e.g. super-population models [44] or game theoretic approaches [45,46]). The proposed transformation method produces truthful datasets which are well suited for performing common statistical analyses [21,28,47,34]. Finally, the software overcomes several limitations of previous data anonymization solutions: data can easily be protected from multiple threats by combining different anonymization operations within a single ETL workflow and very large datasets can be processed by leveraging the streaming-based processing model of the underlying platform. Due to the fact that our approach can be used to process data which has been partitioned into independent subsets (see Section 2.3 and Section C of the supplementary file) it can also be used to incrementally add data to

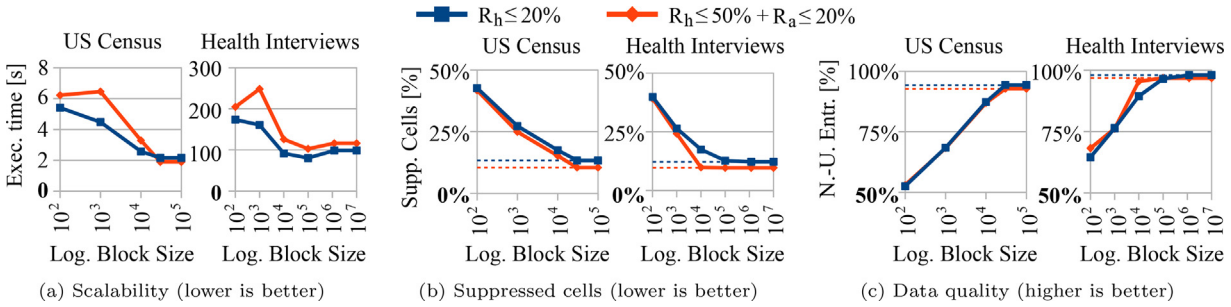


Fig. 8. Semi-log plots visualizing the results of row blocking experiments. We report average execution times, the number of suppressed cells and data quality as reported by the Non-Uniform Entropy model. Dotted lines represent baseline values obtained without row-blocking.

existing databases without violating the privacy guarantees provided.

The software described in this article has already been used in various projects. For example, it was used to anonymize demographic data for a research data warehouse at the Department of Cardiovascular Diseases of the German Heart Centre Munich. The warehouse integrated phenotypic and genotypic data of more than 70,000 patients with coronary artery disease to support data visualization, cohort discovery and hypothesis generation. We have also frequently used the methodology described here when protecting data extracts before sharing them with external partners, for example in the context of research registries for mitochondrial disorders [48] and for neurodegenerative diseases [49]. Finally, the described methods have also been used through ARX by other research groups, for example to create an open dataset for studies of learning behaviour [50] and for anonymizing data from a cancer screening program [51].

4.2. Conceptual comparison with prior work

On the conceptual level, prior work can be found in many areas, including data anonymization, synthetic data generation and data masking. We have already covered related environments for implementing ETL processes and other open source data anonymization solutions in the previous sections. Another software worth mentioning is Privacy Analytics Eclipse [52], which is a commercial data anonymization platform built on Apache Spark [53]. While the software implements formal methods that are quite similar to the ones implemented by our plugin, little has been published about the exact methodology and its implementation.

In the remainder of this section, we focus on further solutions that integrate data protection features into ETL processes. *Data masking* is a technique which has also been integrated into ETL platforms. Methods from this field are not based on formal risk assessment and data anonymization, but they implement simple rule-based transformation processes, e.g. for the removal of data. They are typically used to create data for software development and testing purposes. Examples of relevant implementations include Informatica's *Data Masking* [54], IBM's InfoSphere Optim Data Privacy [55], Oracle's Data Masking and Subsetting Pack [56], ProxySQL [57], and Hush Hush's Data Masking Components [58]. Also, TOS and PDI both offer modules providing basic data masking functionalities.

Synthetic data generation is also supported by the masking solutions presented in the previous paragraph. Most implementations are rather simple, but there are also sophisticated approaches, such as the algorithms supported by sdcMicro [18] which are able to preserve uni- and multivariate statistical properties of input data. Random data generation is also supported by plugins for TOS and PDI. Bijoux is another well-known example [59]. However, data generation plugins for ETL processes are typically too simple to be useful for more than test data generation.

5. Conclusion and future work

In this article, we have described a plugin for a common ETL platform which supports robust anonymization and risk assessment functionalities. The software is available under a non-restrictive open source license. Our method can be integrated into existing ETL workflows, and it supports typical warehousing solutions for biomedical data, such as i2b2 and tranSMART. Even in cases where it is not possible to significantly reduce risks without considerable impacts on data utility, our software can be used to perform quantitative re-identification risk assessments for documenting privacy threats. This is an important aspect of modern privacy laws, such as the European General Data Protection Regulation [16].

The methods and implementations presented in this article are particularly well suited for protecting data that is collected infrequently (e.g. demographics) or which remains rather stable over time (e.g. diagnoses or lab values of particular interest for a specific study) [14,27]. If longitudinal or frequently changing data needs to be protected from linkage attacks, specific measures must be implemented that can cope with higher dimensionality and changes to data [60]. While we plan to extend our software to cover such use cases in future work, we also emphasize that such data often poses much less risk, as it is unstable, difficult to replicate and it is therefore less likely that adequate background knowledge is available to adversaries [14,27]. An additional area of future work is improved support for incrementally adding new data. While this is supported already by the current version of our plugin, we plan to add functionalities for considering data that already exists within the database when measuring and reducing risks during the process of loading new data. This could help to further reduce the amount of suppression needed.

Cell suppression enables the anonymization of datasets with minimal configuration efforts, but further transformation methods can also be useful in certain scenarios. Data generalization and microaggregation are two techniques of specific interest. We plan to add support in future versions of the plugin. However, as these methods may have impacts on the schematic properties of data (e.g. changes in data types and scales of measure) integrating them with the processing environments of ETL solutions is challenging. An alternative anonymization approach to cell suppression is cell swapping (or data swapping) [61] which essentially works by exchanging attribute values between records. Analogously to our work, it preserves the schematic properties of data. In contrast to our approach, data swapping does not remove attribute values and hence it preserves statistical aggregates such as counts of attribute values. However, unlike cell suppression, data swapping is inherently perturbative. Hence, it does not satisfy truthfulness, which is an important requirement in our context (cf. Section 2.2). Moreover, data swapping is typically implemented based on simple risk models, which offer much lower degrees of protection than the methods used in our work. A potential direction for future work would be to investigate how data swapping could be integrated into the proposed anonymization framework, including the strong

protection models used, and to examine potential resulting increases of data utility. One possible approach for this would be to firstly perform anonymization using cell suppression, including risk assessment as described in this article. This step could then be followed by a post-processing step in which the original values of suppressed cells are being swapped and then re-inserted into the output dataset.

While our plugin supports one of the most widely adopted environments for implementing ETL processes, TOS is also frequently used in biomedical data warehousing projects. We have already started to port our plugin to this platform but, due to the differences between development environments and concepts for managing data and control flows, a complete integration will require more work.

Summary points

What was already known on the topic?

- Anonymization is important in biomedical research, especially when data is pooled or re-used for secondary purposes.
- Common ETL (Extract-Transform-Load) tools for integrating data into clinical and translational warehouses do not support anonymization. Moreover, common anonymization tools cannot easily be integrated into ETL workflows.
- Anonymization tools can be difficult to configure and they have scalability issues when processing very large datasets.

What has this study added to the body of knowledge?

- Expert-level anonymization methodologies can be integrated as intuitive plugins into ETL platforms.
- With these plugins, data can be protected from multiple threats within a single ETL workflow.
- Very large datasets can be anonymized efficiently by leveraging the streaming-based processing model of ETL platforms.
- High data utility and compatibility with existing databases and platforms can be achieved by using transformation methods that preserve both the truthfulness of data and its schematic properties.

;1;

Authors' contributions

FP, HS, and RB developed the algorithms. HS and FP designed and implemented the plugins. HS, JE, RB and FP designed, implemented and performed the experiments. RB, HS and FP developed the formal proofs of correctness of the approach. HS, RB, JE, KK and FP discussed the conception and design of the work as well as the manuscript at all stages. All authors have contributed to the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The work was, in parts, funded by the German Federal Ministry of Education and Research (BMBF) within the "Medical Informatics Funding Scheme" under reference number 01ZZ1804A (DIFUTURE).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijmedinf.2019.03.006.

References

[1] S.O. Dyke, A.A. Philippakis, J.R. De Argila, D.N. Paltoo, E.S. Luetkemeier, B.M. Knoppers, A.J. Brookes, J.D. Spalding, M. Thompson, M. Roos, et al., Consent codes: upholding standard data use conditions, PLoS Genet. 12 (1) (2016) e1005772, https://doi.org/10.1371/journal.pgen.1005772.
[2] S. Schneeweiss, Learning from big health care data, N. Engl. J. Med. 370 (23) (2014) 2161–2163, https://doi.org/10.1056/NEJMp1401111.
[3] A.J. McMurry, S.N. Murphy, D. MacFadden, G. Weber, W.W. Simons, J. Orechia, J. Bickel, N. Wattanasin, C. Gilbert, P. Trevvett, et al., SHRINE: enabling nationally scalable multi-site disease studies, PLoS One 8 (3) (2013) e55811, https://doi.org/10.1371/journal.pone.0055811.
[4] K. Shameer, M.A. Badgeley, R. Miotto, B.S. Glicksberg, J.W. Morgan, J.T. Dudley, Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams, Brief. Bioinform. 18 (1) (2017) 105–124, https://doi.org/10.1093/bib/bbv118.
[5] I. Danciu, J.D. Cowan, M. Basford, X. Wang, A. Saip, S. Osgood, J. Shirey-Rice, J. Kirby, P.A. Harris, Secondary use of clinical data: the Vanderbilt approach, J. Biomed. Inform. 52 (2014) 28–35, https://doi.org/10.1016/j.jbi.2014.02.003.
[6] A.-S. Jannot, E. Zapletal, P. Avillach, M.-F. Mamzer, A. Burgun, P. Degoulet, The Georges Pompidou University Hospital Clinical Data Warehouse: a 8-years follow-up experience, Int. J. Med. Inform. 102 (2017) 21–28, https://doi.org/10.1016/j.ijmedinf.2017.02.006.
[7] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inform. Assoc. 17 (2) (2010) 124–130, https://doi.org/10.1136/jamia.2009.000893.
[8] E. Scheufele, D. Aronzon, R. Coopersmith, M.T. McDuffie, M. Kapoor, C.A. Uhrich, J.E. Avitabile, J. Liu, D. Housman, M.B. Palchuk, tranSMART: an open source knowledge management and high content data analytics platform, AMIA Jt. Summits Transl. Sci. Proc. (2014) 96–101.
[9] W. Inmon, Building the Data Warehouse, John Wiley & Sons, 2005.
[10] M.J. Denney, D.M. Long, M.G. Armistead, J.L. Anderson, B.N. Conway, Validating the extract, transform, load process used to populate a large clinical research database, Int. J. Med. Inform. 94 (2016) 271–274.
[11] M. Casters, R. Bouman, J. Van Dongen, Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration, John Wiley & Sons, 2010.
[12] J. Bowen, Getting Started with Talend Open Studio for Data Integration, Packt Publishing Ltd, 2012.
[13] C. Bauer, T. Ganslandt, B. Baum, J. Christoph, I. Engel, M. Löbe, S. Mate, S. Stäubert, J. Drepper, H.-U. Prokosch, U. Sax, Integrated Data Repository Toolkit (IDRT). A suite of programs to facilitate health analytics on heterogeneous medical data, Methods Inf. Med. 55 (2) (2016) 125–153, https://doi.org/10.3414/ME15-01-0082.
[14] B.A. Malin, D. Karp, R.H. Scheuermann, Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research, J. Investig. Med. 58 (1) (2010) 11–18, https://doi.org/10.2310/JIM.0b013e3181c9b2ea.
[15] US Department of Health and Human Services, Standards for privacy of individually identifiable health information, Final Rule. 45 CFR, Parts 160–164, Federal Register 67 (157) (2002) 53182–53273.
[16] Regulation (EU) 2016/679 of the Eur. 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), Off. J. Eur. Union (May 2016) L119/59.
[17] F. Kohlmayer, F. Prasser, K.A. Kuhn, The cost of quality: implementing generalization and suppression for anonymizing biomedical data with minimal information loss, J. Biomed. Inform. 58 (2015) 37–48, https://doi.org/10.1016/j.jbi.2015.09.007.
[18] M. Templ, A. Kowarik, B. Meindl, Statistical disclosure control for microdata using the R-package sdcMicro, J. Stat. Softw. 67 (1) (2015) 1–36, https://doi.org/10.18637/jss.v067.i04.
[19] F. Prasser, F. Kohlmayer, Putting statistical disclosure control into practice: the ARX data anonymization tool, Medical Data Privacy Handbook, Springer, 2015, pp. 111–148.
[20] K. El Emam, L. Arbuckle, Anonymizing Health Data: Case Studies and Methods to Get You Started, 1st ed., O'Reilly, 2013.
[21] K. El Emam, B.A. Malin, 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 (Eds.), Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, National Academies Press (US), Washington (DC), 2015, pp. 1–290.
[22] European Union Agency for Network and Information Security (ENISA), Privacy and Data Protection by Design – from policy to engineering (2014), 1–79.
[23] European Medicines Agency (EMA), EMA/90915/2016 – External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use (2016), 1–99.
[24] B. Fung, K. Wang, R. Chen, P.S. Yu, Privacy-preserving data publishing: a survey of recent developments, ACM Comput. Surv. (CSUR) 42 (4) (2010) 14.
[25] L. Sweeney, k-anonymity: a model for protecting privacy, Int. J. Uncertainty Fuzziness Knowl.-Based Syst. 10 (05) (2002) 557–570.
[26] F. Prasser, F. Kohlmayer, K.A. Kuhn, et al., The importance of context: risk-based de-identification of biomedical data, Methods Inf. Med. 55 (4) (2016) 347–355, https://doi.org/10.3414/ME16-01-0012.
[27] B. Malin, G. Loukides, K. Benitez, E.W. Clayton, Identifiability in biobanks: models,

measures, and mitigation strategies, Hum. Genet. 130 (3) (2011) 383.
[28] K. El Emam, Guide to the De-Identification of Personal Health Information, CRC Press, 2013.
[29] 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, Available from SSRN: http://ssrn.com/abstract=2076397. Accessed 5 January 2018 (2012). doi:10.2139/ssrn.2076397.
[30] F.K. Dankar, K. El Emam, Practicing differential privacy in health care: a review, Trans. Data Privacy 6 (1) (2013) 35–67.
[31] C. Dwork, Differential privacy: a survey of results, International Conference on Theory and Applications of Models of Computation Springer (2008) 1–19.
[32] J. Domingo-Ferrer, J.M. Mateo-Sanz, Practical data-oriented microaggregation for statistical disclosure control, IEEE Trans. Knowl. Data Eng. 14 (1) (2002) 189–201.
[33] X. Xiao, Y. Tao, Anatomy: simple and effective privacy preservation, Proceedings of the 32nd International Conference on Very Large Data Bases, VLDB Endowment, 2006, pp. 139–150.
[34] L. Ohno-Machado, S. Vinterbo, S. Dreiseitl, Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance, J. Am. Med. Inform. Assoc. 9 (Supplement_6) (2002) S115–S119.
[35] F. Prasser, F. Kohlmayer, K.A. Kuhn, Efficient and effective pruning strategies for health data de-identification, BMC Med. Inform. Decis. Mak. 16 (1) (2016) 49, https://doi.org/10.1186/s12911-016-0287-2.
[36] arx-deidentifier/arx-pdi-plugins, Plugins for the Pentaho Data Integration platform. Available from https://github.com/arx-deidentifier/arx-pdi-plugins. Accessed 23 March 2018.
[37] arx-deidentifier/cell-suppression-benchmark, Benchmark of cell-suppression methods in ARX. Available from https://github.com/arx-deidentifier/cell-suppression-benchmark. Accessed 23 March 2018.
[38] M. Ciglic, J. Eder, C. Koncilia, k-Anonymity of microdata with NULL values, Int. Conf. Database Exp. Sys. Appl. Springer (2014) 328–342, https://doi.org/10.1007/978-3-319-10073-9_27.
[39] S. Kim, H. Lee, Y.D. Chung, Privacy-preserving data cube for electronic medical records: an experimental evaluation, Int. J. Med. Inform. 97 (2017) 33–42, https://doi.org/10.1016/j.ijmedinf.2016.09.008.
[40] L.H. Cox, A.F. Karr, S.K. Kinney, Risk-utility paradigms for statistical disclosure limitation: how to think, but not how to act, Int. Stat. Rev. 79 (2) (2011) 160–183, https://doi.org/10.1111/j.17515823.2011.00140.x.
[41] F. Prasser, F. Kohlmayer, K.A. Kuhn, Efficient and effective pruning strategies for health data de-identification, BMC Med. Inform. Decis. Mak. 16 (1) (2016) 49, https://doi.org/10.1186/s12911-016-0287-2.
[42] F. Prasser, R. Bild, J. Eicher, H. Spengler, F. Kohlmayer, K.A. Kuhn, Lightning: utility-driven anonymization of high-dimensional data, Trans. Data Privacy 9 (2) (2016) 161–185.
[43] A. De Waal, L. Willenborg, Information loss through global recoding and local suppression, Netherlands Off. Stat. 14 (1999) 17–20.
[44] F.K. Dankar, K. El Emam, A. Neisa, T. Roffey, Estimating the re-identification risk of clinical data sets, BMC Med. Inform. Decis. Mak. 12 (1) (2012) 66, https://doi.org/10.1186/1472-6947-12-66.
[45] Z. Wan, Y. Vorobeychik, W. Xia, E.W. Clayton, M. Kantarcioglu, R. Ganta, R. Heatherly, B.A. Malin, A game theoretic framework for analyzing re-identification risk, PLoS One 10 (3) (2015) e0120592, https://doi.org/10.1371/journal.pone.0120592.
[46] F. Prasser, J. Gaupp, Z. Wan, W. Xia, Y. Vorobeychik, M. Kantarcioglu, K.A. Kuhn, B.A. Malin, An open source tool for game theoretic health data de-identification, AMIA Annu. Symp. Proc. (2017).
[47] K. El Emam, F.K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, et al., A globally optimal k-anonymity method for the de-identification of health data, J. Am. Med. Inform. Assoc. 16 (5) (2009) 670–682, https://doi.org/10.1197/jamia.M3144.
[48] B. Büchner, C. Gallenmüller, R. Lautenschläger, K. Kuhn, I. Wittig, L. Schöls, D. Rapaport, D. Seelow, P. Freisinger, H. Prokisch, et al., The German Network for Mitochondrial Disorders (mitoNET), Med. Genet. 24 (3) (2012) 193–199, https://doi.org/10.1007/s11825-012-0338-8.
[49] B. Kalman, B. Büchner, F. Kohlmayer, K.A. Kuhn, R. Lautenschlaeger, T. Klopstock, T. Kmiec, An international registry for neurodegeneration with brain iron accumulation, Orphanet J. Rare Dis. 7 (2012) 66, https://doi.org/10.1186/1750-1172-7-66.
[50] J. Kuzilek, M. Hlosta, Z. Zdrahal, Open University Learning Analytics dataset, Sci. Data 4 (2017) 170171, https://doi.org/10.1038/sdata.2017.171.
[51] G. Ursin, S. Sen, J.-M. Mottu, M. Nygård, Protecting privacy in large datasets—first we assess the risk; then we fuzzy the data, Cancer Epidemiol. Biomarkers Prev. 26 (8) (2017) 1219–1224, https://doi.org/10.1158/1055-9965.EPI-17-0172.
[52] Privacy Analytics, Inc., Privacy Analytics Eclipse. Available from https://privacy-analytics.com/software/privacy-analytics-eclipse/. Accessed 5 January 2018.
[53] Apache Spark. Available from https://spark.apache.org/. Accessed 12 January 2018.
[54] Informatica Corporation, Data Masking. Available from https://www.informatica.com/gb/products/data-security/data-masking.html. Accessed 5 January 2018.
[55] IBM Corporation, IBM InfoSphere Optim Data Privacy. Available from https://www.ibm.com/ms-en/marketplace/infosphere-optim-data-privacy/details#product-header-top. Accessed 5 January 2018.
[56] Oracle Corporation, Oracle Data Masking and Subsetting Pack. Available from http://www.oracle.com/technetwork/database/options/data-masking-subsetting/overview/ds-security-dms-2245926.pdf. Accessed 12 January 2018 (2016).
[57] R. Cannaò, ProxySQL. Available from http://proxysql.com. Accessed 5 January 2018 (2018).
[58] Hush, Hush Information Technology and Services, Data Masking Components. Available from http://mask-me.net/. Accessed 5 January 2018 (2017).
[59] V. Theodorou, P. Jovanovic, A. Abelló, E. Nakuçi, Data generator for evaluating ETL process quality, Inform. Sys. 63 (Supplement C) (2017) 80–100, https://doi.org/10.1016/j.is.2016.04.005.
[60] M. Terrovitis, N. Mamoulis, P. Kalnis, Privacy-preserving anonymization of set-valued data, Proceedings of the VLDB Endowment 1 (1) (2008) 115–125.
[61] S.E. Fienberg, J. McIntyre, Data swapping: variations on a theme by Dalenius and Reiss, J. Off. Stat. 21 (2) (2005) 309.