2017 IEEE 30th International Symposium on Computer-Based Medical Systems

A Tool for Optimizing De-Identified Health Data for Use in Statistical Classification

Fabian Prasser, Johanna Eicher, Raffael Bild, Helmut Spengler and Klaus A. Kuhn
Technical University of Munich
University Hospital rechts der Isar
Institute of Medical Statistics and Epidemiology
Ismaninger Str. 22, 81675 Munich, Germany
firstname.lastname@tum.de

***Abstract*—When individual-level health data is shared in biomedical research the privacy of patients and probands must be protected. This is typically achieved with methods of data de-identification, which transform data in such a way that formal guarantees about the degree of protection from re-identification can be provided. In the process it is important to minimize loss of information to ensure that the resulting data is useful. A typical use case is the creation of predictive models for knowledge discovery and decision support, e.g. to infer diagnoses or to predict outcomes of therapies. A variety of methods have been developed which can be used to build robust statistical classifiers from de-identified data. However, they have not been tuned for practical use and they have not been implemented into mature software tools. To bridge this gap, we have extended ARX, an open source anonymization tool for health data, with several new features. We have implemented a method for optimizing the suitability of de-identified data for building statistical classifiers and a method for assessing the performance of classifiers built from de-identified data. All methods are accessible via a comprehensive graphical user interface. We have used our implementation to create logistic regression models from a patient discharge dataset for predicting the costs of hospital stays. The results show that our approach enables the creation of privacy-preserving classifiers with optimal prediction accuracy.**

***Keywords*-biomedical data; security; privacy; de-identification; data mining; classification;**

I. Introduction

Collaborative collection and sharing of sensitive individual-level health data have become vital aspects of modern biomedical research. Organizational and technical safeguards must be implemented to protect the privacy of patients and probands [1]. Data de-identification, i.e. the sanitization of datasets in ways that prevent attackers from breaching the subjects' privacy, is an important building block. However, simply removing all directly identifying information (e.g. names) is not sufficient [2], [3], making health data de-identification a non-trivial issue [4], [5].

Two conflicting objectives have to be balanced with each other: *minimizing privacy risks* and *maintaining data quality*. The conflict is typically resolved by letting a decision maker specify a *risk threshold*, which defines a preference on one of the optimization goals. The result is a simpler optimization problem in which the reduction of data quality must be minimized while data is transformed (e.g. by removing attribute values) to ensure that risk thresholds are met.

Data mining and predictive modelling are important use cases for health data. Typical application scenarios include knowledge discovery, i.e. the detection of unknown relationships between biomedical parameters, and decision support where the knowledge about such relationships is used to infer or predict parameters, e.g. diagnoses or outcomes. Specialized de-identification methods must be used to ensure that output data remains useful for such purposes.

II. Related Work

Early work on privacy-preserving data mining has suggested that de-identified data is not well suited for building statistical classifiers [6]. Since then, many methods for optimizing de-identified data for data mining tasks and statistical classification have been developed (see [7] for a survey). Moreover, measuring the performance of classifiers built from output data is now a standard method for evaluating the degree of data quality provided by de-identification methods [8] and increasingly more sophisticated de-identification methods are being developed which overcome the limitations of early approaches [9].

There is also a growing number of software solutions which aim at making these methods available to data controllers by providing easy to use graphical interfaces. For example, *μ-ARGUS* [10] and *sdcMicro* [11] are tools developed in the context of official statistics. *ARX* is the only open source de-identification tool which has specifically been designed for biomedical data [12]. The strengths and weaknesses of all three tools have been described in the privacy-by-design guideline from the *European Union Agency for Network and Information Security* (ENISA) [13]. ARX has been mentioned by the *European Medicines Agency* (EMA) in a guideline on the implementation of its policy on clinical trial data sharing [14].

An important usage scenario for these tools is to help users with assessing and optimizing the usefulness of de-identified data for a wide variety of applications. However, methods for building statistical classifiers from de-identified data have not been tuned for applications in

2372-9198/17 $31.00 © 2017 IEEE
DOI 10.1109/CBMS.2017.105

practice and they have not been implemented into mature tools, such as *sdcMicro*, *μ-ARGUS* or ARX.

III. Objective and Outline

Motivated by the increasing importance of data mining in biomedical research, we have extend ARX with methods for optimizing the suitability of de-identified data for building statistical classifiers and for assessing the performance of classifiers built from de-identified data. Our main objective was to make tools for building privacy-preserving statistical classifiers readily available to medical informatics professionals and to other experts, e.g., researchers responsible for the sharing of data. To achieve this, we had to integrate a wide variety of different methods (e.g. for transforming data, for measuring data usefulness and for measuring privacy risks) with each other.

Firstly, we have developed a data quality model which can be used to optimize output data for use in statistical classification. Secondly, we have implemented a method for assessing the performance of classifiers built from de-identified data. Thirdly, we have extended ARX's graphical user interface with views for visually analyzing classification performance. We have conducted experiments with a patient discharge dataset by de-identifying it and using the output to train logistic regression models to predict the cost of hospital stays. The results show that our method can be used to build classifiers with optimal prediction accuracy.

The remainder of this paper is structured as follows: In Section IV we provide background information on data de-identification and statistical classification. We describe the methods which we have implemented in Section V. Section VI provides a brief description of the setup of our experiments. The results of our work and the experiments are then presented in Section VII. Finally, we conclude with a summary of our findings in Section VIII.

IV. Background

A. Risk Models

The most important privacy risk from which datasets are typically protected is *re-identification*, which means that an individual can be linked to a specific data record [3]. If successful, such a privacy breach can have severe legal consequences in many jurisdictions around the world. The attack vector is *linkage* of a sensitive dataset with a dataset containing identifying information about individuals. The attributes that may be used for linkage are termed *quasi-identifiers*. These attributes are not directly identifying (such as names) but they may be combined to form a unique key which is also contained in other datasets. Also, they cannot simply be removed as they may be required for analyses.

The most well-known privacy model for protecting data from re-identification is *k-anonymity* [3]. A dataset is said to be *k-anonymous* if, regarding the quasi-identifiers, each record cannot be distinguished from at least $k-1$ other records. As a result, the probability of correctly linking an individual to the corresponding record is guaranteed to be $\leq \frac{1}{k}$ and k can therefore be used as a risk threshold. In Section VIII we will briefly cover further privacy threats and risk models.

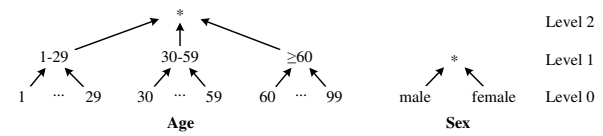

Figure 1. *Simple generalization hierarchies for attributes age and sex.*

B. Data Transformation

Health data is typically de-identified with user-defined generalization hierarchies [4], [5]. With these type of transformation rules the precision of attribute values can iteratively be reduced. As a consequence, the records' characteristics become less unique and privacy risks decrease. Two simple examples are shown in Fig. 1. As can be seen, each hierarchy contains a set of increasing levels, which specify values with increasing coverage of an attribute's domain. Generalization hierarchies are specifically well suited for transforming categorical attributes but they can also be used for continuous attributes by performing categorization [15].

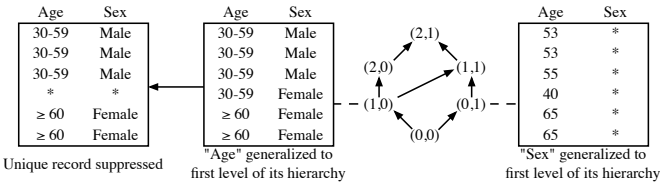

Figure 2. *Generalization lattice constructed with the example hierarchies and the results of two transformations.*

Different methods exist for transforming data with generalization. *Full-domain generalization* means that the same level of generalization is used to transform all values of an attribute. An example which uses the hierarchies from Fig. 1 is provided in Fig. 2. Each node represents a transformation which defines generalization levels for all quasi-identifiers. The transformation $(0,0)$ which does not generalize any value is at the bottom, whereas the transformation with maximal generalization $(2,1)$ is at the top.

Full-domain generalization is a rather restrictive transformation model. To improve the quality of output data it is therefore often combined with other transformation methods. *Record suppression* is a typical example. In the output of applying the transformation $(1,0)$ to the dataset, which is shown on the left in Fig. 2, one record has been suppressed to make the dataset 2-anonymous (50% risk threshold).

C. Data Quality

De-identification algorithms typically search through the set of available data transformations to find a combination of attribute generalization and record suppression which results

in optimal output data quality. For this purpose they employ mathematical models which quantify the loss of information or the quality or usefulness of data [16].

Measuring data quality is non-trivial as the nature of usefulness of data depends on the application [8]. It is important that a de-identified dataset is either *analytically valid* or *analytically interesting*. Analytical validity means the preservation of major statistical characteristics while data is said to be analytically interesting if at least some properties remain intact that are useful for further analyses [17].

Data quality models typically calculate a normalized score for the output of each transformation. A score of 0% represents the original input dataset while a score of 100% represents a dataset from which all information has been removed. The optimal solution to a de-identification problem is then defined as the output dataset with the lowest score.

D. Statistical Classification

Statistical classification is a common use case for individual-level data [18]. The goal is to predict the value of a predefined *class attribute* from a given set of values of *feature attributes* as accurately as possible. This is implemented with *supervised learning* where a model is created from a *training set* [19].

It is an inherent property of de-identification that information is removed from a dataset. However, when de-identified data is to be used as a training set for statistical models it is important to distinguish between the removal of *structure* and the removal of *noise*. While the former significantly impacts the suitability of data for this purpose, the latter does not [20]. For instance, if there is a strong correlation between the age of individuals and the prevalence of a certain disease, adequate age groups instead of more granular data are likely to be very well suited for capturing the relationship.

Unfortunately, it is typically not possible to optimize the output of a de-identification algorithm for statistical classification by training a classifier on each of the possible output datasets and comparing their performance. The reason is that the number of possible solutions can become very large and the training of a classifier is computationally expensive. However, the aspects covered in the previous paragraph can be implemented into data quality models, such that the output of a de-identification algorithm can be optimized to contain relevant information in adequate granularity. Related feature selection and preprocessing steps are often used to improve statistical classifiers [19].

V. Methods

A. Optimizing the Usefulness of Output Data

Iyengar has proposed a quality model which aims at reducing the loss of information regarding features that are most discriminating for the class attribute [21]. First, the model groups the records of a dataset D by the values of the defined feature attributes. As a result, each record $r \in D$ is contained in an *equivalence classes* $E(r)$ of records having the exact same features. The model then penalizes all records which contain a value for the class attribute that is different from the most frequent value of the class attribute in the according equivalence class.

We have extended this model by introducing more fine-grained penalties that capture more aspects of discrimination. Each record r is penalized with P_r if the record is suppressed, penalized with P_m if the records' class value $class(r)$ is different from the most frequent value of the class attribute $majority(E(r))$ within $E(r)$ or penalized with P_h if no unique most frequent value of the class label exists within $E(r)$.

For the individual parameters, we propose the following configuration: Analogously to [21] we define $P_m = 1$. In addition, we define $P_h = 1$, since an equivalence class of features without a most-frequent class label is not discriminating for the class attribute, either. Moreover, we define $P_r = \frac{1}{2}$, as a removed record implies loss of information but the contained information is likely to be noise because de-identification algorithms tend to extract patterns and remove outliers. The score of a dataset D is defined as the sum of the penalties of all records divided by the number of records:

$$\frac{1}{|D|}\sum_{r \in D} p(r) = \begin{cases} \frac{1}{2} & \text{if } r \text{ is suppressed,} \\ 1 & \text{if } class(r) \neq majority(E(r)), \\ 1 & \text{if } majority(E(r)) \text{ does not exist,} \\ 0 & \text{otherwise.} \end{cases}$$

During de-identification ARX traverses the solution space and finally returns the output dataset with the lowest score.

B. Assessing the Performance of Classifiers

In addition to providing methods for optimizing the output of data de-identification for statistical classification it is also important to provide methods for assessing the performance of classifiers built from such data. Firstly, this functionality may be used to review results obtained with the method presented in the previous section. Secondly, users may also

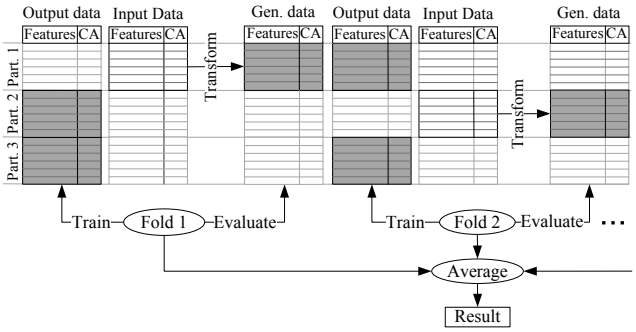

Figure 3. *Example for interwoven 3-fold cross-validation of a model built with de-identified data against the input dataset.*

want to determine whether data which has been optimized for other purposes is also suitable for statistical modelling.

A typical method for estimating the performance of statistical classifiers is to estimate their prediction *accuracy*, i.e. the expected number of correct predictions, using *k-fold cross-validation* [22]. In this process the records of a dataset are first divided randomly into k partitions of equal size. Each iteration $i \in \{1, 2, ..., k\}$ consists of four steps: (1) select the records in partition p_i as evaluation data, (2) train the model using all other records as training data, (3) evaluate the accuracy of the classifier by using it to predict the (known) values of the class attribute from the features of the records in partition p_i. The results obtained for all folds are then averaged to derive an overall estimate of the model's performance.

However, the accuracy of classifiers trained with de-identified data cannot be assessed by simply performing k-fold cross-validation on the de-identified data and comparing the results with the results of k-fold cross-validation performed on the input dataset [18]. Instead, a classifier must be built from transformed output data which is able to make predictions based on features which have not been transformed. This can be achieved by implementing a preprocessing step which transforms a given set of previously unknown features in the same manner in which the de-identified training data had been transformed before passing it to the classifier to make predictions [18]. The classifier can then be trained on de-identified data but evaluated on input data.

We have used this approach to efficiently implement *interwoven k-fold cross-validation* into ARX. As is illustrated in Fig. 3 classifiers are trained with the records from the de-identified output dataset (e.g. *Partition 2* and *Partition 3* for *Fold 1*) but they are evaluated using a transformed version of the records from the according partition of the *input* dataset (*Partition 1* for *Fold 1*). By using the exact same folds, results obtained with classifiers trained on de-identified data are comparable to the results obtained for classifiers trained on input data. We compute the following four basic parameters to assess prediction accuracies:

1) *Baseline accuracy:* We determine the baseline accuracy by using the *ZeroR* method, which is a trivial classifier that always returns the most frequent class of the training dataset, while using the original dataset as both training and evaluation data.
2) *Original accuracy:* We determine the performance of a non-trivial classifier trained with and evaluated on the original input dataset.
3) *Accuracy*: We determine the accuracy of a non-trivial classifier built from the de-identified dataset by using the interwoven k-fold cross-validation process described in the previous paragraph (see also Fig. 3).
4) *Relative prediction accuracy:* We determine the relative prediction accuracy of the classifier trained with de-identified data by normalizing its accuracy using the baseline accuracy and original accuracy.

Table I
DESCRIPTION OF THE PATIENT DISCHARGE DATASET.

Attribute	Type	Description
Hospital ID	Spatial	Unique identifier
Age	Demographic	Patient's age at admission in years
Sex	Demographic	Patient's sex
Ethnicity	Demographic	Patient's ethnicity
Race	Demographic	Patient's racial background
ZIP Code	Spatial	Patient's ZIP code of residence
County	Spatial	Patient's county of residence
Length of stay	Temporal	Total number of days from admission to discharge
Admission quarter	Temporal	The calendar quarter the patient was admitted
Charge	Monetary	Total charges for the stay

Our implementation supports a wide variety of classification methods. We have performed experiments with C4.5 decision trees and logistic regression models [19]. For licensing reasons, the open source version of ARX implements logistic regression using *Apache Mahout* [23].

VI. EXPERIMENTAL SETUP

We have used a publicly available patient discharge dataset to evaluate our method [24]. Although our implementation is highly scalable, we have taken a random sample of 39,676 records (1% of the original dataset) because our evaluation required hundreds of thousands of experiments to be performed (see below). The dataset consisted of ten attributes, which are described in more detail in Table I. We considered all spatial, demographic and temporal attributes to be quasi-identifying. We also used them as features for predicting whether the *charge* for a hospitalization will be *1)* between \$10,000 and \$50,000, *2)* lower, or *3)* higher.

We de-identified the dataset using the *k-anonymity* privacy model. We chose $k = 5$, which is a typical parameter in the biomedical domain that specifies a threshold of not more than 20% re-identification risk for each record [1]. As transformation methods, we used attribute generalization followed by record suppression. The generalization hierarchies for the quasi-identifiers comprised between two and four generalization levels. The number of transformations in the search space, which is defined by the product of the heights of the generalization hierarchies [15], was 36,864.

For each of the transformations available, we evaluated the performance of a classifier trained with the according output data using the method described in the previous section and interwoven k-fold cross-validation with $k = 3$, which is a typical parameter [18]. This means that three different types of classifiers (ZeroR and logistic regression models trained with input and output data) were created for each of the three folds of each of the 36,864 transformed datasets, resulting in well over 300,000 training and evaluation steps.

In addition, we have calculated the quality of each dataset with the model proposed in Section V-A and with three

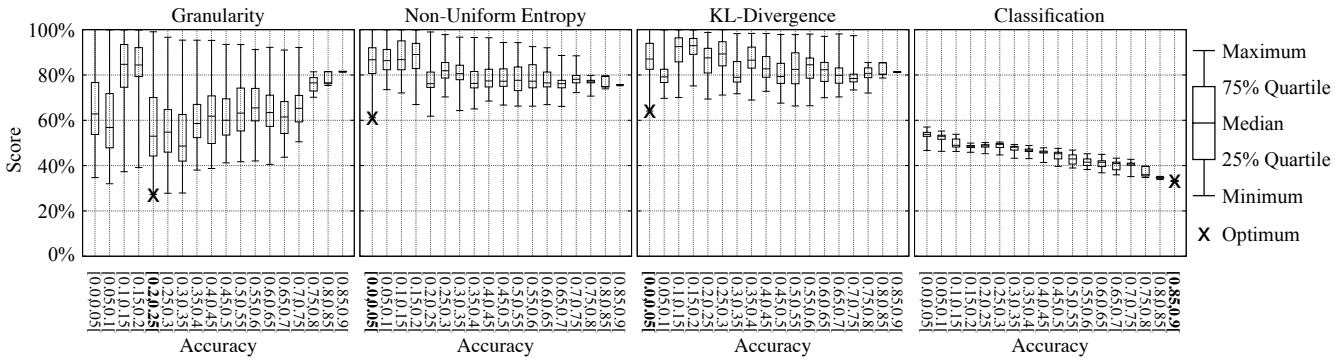

Figure 4. *Prediction accuracy (higher is better) vs. score (lower ist better) for various data quality models.*

other well-known models: 1) Granularity, which measures the precision of output data [21], 2) Kullback-Leiber (K.-L.) Divergence, which quantifies differences in the distributions of sizes of groups of indistinguishable records [25] and 3) Non-Uniform Entropy, which measures changes in the distributions of attribute values [26].

VII. Results

A. User Interface

Fig. 5 shows a screen shot of the graphical user interface which we have implemented into ARX for assessing the quality of statistical classifiers. As can be seen, a table shows the parameters described in Section V as well as additional information, such as the average classification error [19]. An additional plot displays precision and recall for a variety of confidence thresholds [19]. ARX presents this view for input data and output data to allow for visual comparisons.

Additional views, which are not shown in the figure, allow users to select the class attributes as well as feature attributes and to configure the parameters of the classification method. The time required to calculate all data displayed by the interface depends on the transformations applied to the dataset. For our evaluation dataset, execution times varied between 10s and 30s using the settings described in the previous section on a commodity desktop PC.

B. Experimental Results

Fig. 4 presents the results of our experimental evaluation. For each of the four quality models considered, we have correlated the score for each possible output dataset with the performance of the prediction model created from the output dataset. It can be seen that the models Granularity, Non-Uniform Entropy and KL-Divergence did not adequately capture the suitability of data for building classification models. Selecting the optimal solution according to their scores resulted in classifiers with relative prediction accuracies between not more than 5% and not more than 20%.

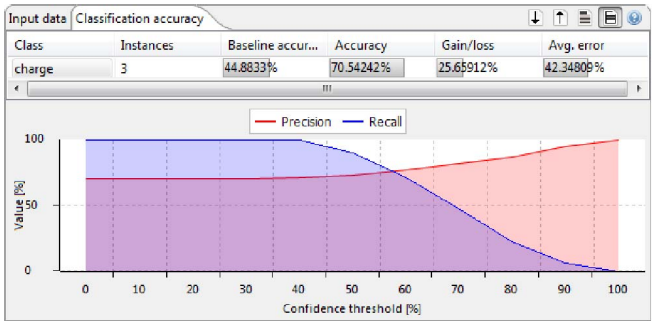

Figure 5. *View for assessing the performance of statistical classifiers.*

The quality model presented in this article, however, captured the resulting prediction accuracies very well. Choosing the optimal solution according to this model resulted in a classifier with a relative prediction accuracy of almost 90%. As can be seen, this is the optimal solution that is available for the transformation model and risk thresholds utilized.

VIII. Conclusion

We have extended ARX, an open source anonymization tool for biomedical data, with methods for optimizing its output for use in statistical classification. The most important building blocks of our implementation are a model for measuring the usefulness of de-identified data for building statistical classifiers and a method for assessing the performance of classifiers built from de-identified data. The results of our experimental analysis show that our approach can be used to create classifiers which are optimal in terms of prediction accuracy.

In our experiments we have used the simple k-anonymity privacy model and the restrictive transformation models full-domain generalization and record suppression. We emphasize that the described implementation is fully integrated into ARX and that it can therefore be used in combination with a wide variety of additional methods implemented by the software. For example, ARX supports privacy models for protecting medical data from sensitive attribute disclosure, e.g. ℓ-diversity and t-closeness [27]. Moreover, it supports the very strong privacy model *Differential Privacy* as well as several models for considering the adversary's background knowledge when protecting data from re-identification [28]. ARX also implements data transformation models which are more flexible than full-domain generalization, including multi-dimensional global recoding and microaggregation [8].

REFERENCES

[1] K. El Emam and B. Malin, "Appendix B: Concepts and methods for de-identifying clinical trial data," in *Sharing clinical trial data: Maximizing benefits, minimizing risk*. The National Academies Press, 2015.

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

[3] L. Sweeney, "Computational disclosure control - A primer on data privacy protection," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[4] K. El Emam and L. Arbuckle, *Anonymizing health data: Case studies and methods to get you started*, 1st ed. O'Reilly Media, Inc., December 2013.

[5] W. Xia, R. Heatherly, X. Ding, J. Li, and B. A. Malin, "R-U policy frontiers for health data de-identification," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 1029–1041, October 2015.

[6] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *14th International Conferene on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 2008, pp. 70–78.

[7] C. C. Aggarwal and S. Y. Philip, "A general survey of privacy-preserving data mining models and algorithms," in *Privacy-preserving data mining*. Springer, 2008, pp. 11–52.

[8] B. C. M. Fung, K. Wang, A. W.-C. Fu, and P. S. Yu, *Introduction to privacy-preserving data publishing: Concepts and techniques*, 1st ed. CRC Press, 2010.

[9] T. S. Gal, T. C. Tucker, A. Gangopadhyay, and Z. Chen, "A data recipient centered de-identification method to retain statistical attributes," *Journal of Biomedical Informatics*, vol. 50, pp. 32–45, 2014.

[10] A. Hundepool and L. Willenborg, "μ-and τ-argus: Software for statistical disclosure control," in *Third International Seminar on Statistical Confidentiality*, 1996.

[11] M. Templ, "Statistical disclosure control for microdata using the R-package sdcMicro," *Transactions on Data Privacy*, vol. 1, no. 2, pp. 67–85, 2008.

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

[13] European Union Agency for Network and Information Security (ENISA), "Privacy and data protection by design – from policy to engineering," pp. 1–79, 2014.

[14] 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," pp. 1–99, 2016.

[15] F. Prasser, F. Kohlmayer, and K. A. Kuhn, "Efficient and effective pruning strategies for health data de-identification," *BMC medical informatics and decision making*, vol. 16, no. 1, p. 49, 2016.

[16] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. A. Kuhn, "Highly efficient optimal k-anonymity for biomedical datasets," in *International Symposium on Computer-Based Medical Systems (CBMS), 2012*. IEEE, 2012, pp. 1–6.

[17] J. Domingo-Ferrer, Sánchez, and S. Hajian, "Database privacy," in *Privacy in a Digital, Networked World: Technologies, Implications and Solutions*, S. Zeadally and M. Badra, Eds. Springer, 2015.

[18] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in *25th International Conference on Data Engineering*. IEEE, 2009, pp. 429–440.

[19] I. H. Witten and F. Eibe, *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[20] B. C. M. Fung, K. Wang, and P. S. Yu, "Anonymizing classification data for privacy preservation," *Transactions on Knowledge and Data Engineering (IEEE)*, vol. 19, no. 5, pp. 711–725, 2007.

[21] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *International Conference on Knowledge Discovery and Data Mining*. ACM, 2002, pp. 279–288.

[22] T. L. Bailey and C. Elkan, "Estimating the accuracy of learned concepts," in *Proc. 13th International Joint Conference on Artifical Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, pp. 895–900.

[23] Apache Software Foundation, "Apache Mahout: Scalable machine-learning and data-mining library," 2011, accessed: January 10, 2017. [Online]. Available: http://mahout.apache.org/

[24] D. Sanchez, S. Martinez, and J. Domingo-Ferrer, "Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata"," *Science*, vol. 351, no. 6279, pp. 1274–1274, 2016.

[25] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," *Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, pp. 24–35, 2007.

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

[27] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," in *International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

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