Technische Universität München
TUM School of Computation, Information and Technology

Measuring and Optimizing the Quality of Anonymized Health Data

Johanna Eicher, M.Sc.

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Daniel Rückert

Prüfer der Dissertation:

1. Prof. Dr. Martin Boeker
2. Prof. Dr. Oliver Kohlbacher

Die Dissertation wurde am 19.02.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 28.11.2024 angenommen.

Für Mama und Papa

Danksagung

Als erstes möchte ich mich bei meinem Betreuer Prof. Klaus A. Kuhn und meinem Mentor Prof. Fabian Prasser dafür bedanken, dass sie diese Dissertation ermöglicht und mich dabei unterstützt haben.

Mein besonderer Dank gilt meinen Eltern Tine und Berni. Eure Unterstützung hat mir den Mut gegeben, all die Entscheidungen zu treffen, die mich bis hier hin geführt haben. Euer Rückhalt ist unersetzlich.

Danke sagen möchte ich auch zu meinem besten Freund Luke. Du hast nicht nur die Höhen miterlebt, sondern warst vor allem auch bei den Tiefen an meiner Seite und hast mich mit ermutigenden Worten aufgebaut.

Schließlich möchte ich mich bei meinen aktuellen und ehemaligen Kolleg*innen am Institut für KI und Informatik in der Medizin bedanken. Die Zusammenarbeit mit euch hat mir immer großen Spaß gemacht. Insbesondere danke ich Ingrid, Florian, Helmut und Martin für anregende Diskussionen und motivierende Worte. Ein ganz besonderer Dank gilt meinem Mitstreiter Raffael. Ohne dich hätte ich mit ziemlicher Sicherheit auf der Zielgerade schlapp gemacht. Danke.

Abstract

Modern biomedical research heavily relies on data-driven approaches, holding immense potential for personalized medicine, more profound insights into diseases, and novel clinical decision support methods. In order to harness this potential, extensive data collections need to be established. However, this necessitates collaborative data gathering involving sensitive individual-level information shared with third parties or repurposed for secondary objectives. While leveraging these data-driven approaches in biomedical research, safeguarding the privacy of involved individuals is crucial. An essential technical aspect is data anonymization, transforming data to minimize privacy risks while optimizing data quality.

Quantifying data quality poses a complex challenge, dependent on the specific use context. For example, when predetermined statistical properties are to be analyzed in the anonymized data, transformation rules preserving these properties can be established. In this thesis, we propose guidelines for selecting quality models tailored to diverse usage scenarios. We accomplished this by integrating various general-purpose quality models into the widely recognized data anonymization tool ARX. Additionally, we conducted extensive experimental comparisons to explore different pertinent aspects. Our findings suggest that specific quality models are best suited for specific usage scenarios. Notably, the Non-Uniform Entropy quality model is particularly well-suited for general-purpose applications.

Even with prior knowledge of the usage scenario, modelling data quality is intricate. For instance, in privacy-preserving statistical classification, where prediction models are built from anonymized data to predict the class attribute value based on a set of feature attributes, minimizing information loss for features is critical, as they are highly discriminative for a specific class attribute.

Even when the usage scenario is known, modelling data quality is a non-trivial issue. For instance, in the context of privacy-preserving statistical classification, where prediction models are built from anonymized data to predict the value of a class attribute from a set of feature attributes, it is vital to minimize the loss of information for the features as these are most discriminating for a specific class attribute. We present a highly adaptable solution for developing and evaluating privacy-preserving prediction models, accommodating various prediction models alongside a range of privacy-preserving techniques. Three case studies are presented to exemplify the practical usability of our solution.

While formalizing the quantification of data quality addresses the challenge, efficient software support is imperative to leverage these quality models within data anonymization. Data anonymization is a multifaceted challenge involving models for defining transformation rules, quantifying privacy risks and data quality, and implementing anonymization algorithms. We introduce an innovative approach that enables ARX to accommodate nearly any combination of anonymization techniques. Through extensive experimental comparisons with existing solutions, we demonstrate superior scalability and data quality, offering support for a broader array of methods and techniques.

Zusammenfassung

Die moderne biomedizinische Forschung stützt sich stark auf datengetriebene Ansätze und birgt ein enormes Potenzial für personalisierte Medizin, tiefere Einblicke in Krankheiten und neuartige Methoden der klinischen Entscheidungsunterstützung. Um dieses Potenzial zu nutzen, müssen umfangreiche Datensammlungen aufgebaut werden. Dies erfordert jedoch eine gemeinschaftliche Datenerfassung, bei der sensible individuelle Informationen auf Ebene der Einzelpersonen gemeinsam gesammelt, mit Dritten geteilt oder für sekundäre Zwecke verwendet werden müssen. Bei der Nutzung dieser datengetriebenen Ansätze in der biomedizinischen Forschung ist der Schutz der Privatsphäre der involvierten Personen von entscheidender Bedeutung. Ein wesentlicher technischer Aspekt ist die Datenanonymisierung, bei der Daten transformiert werden, um die Datenschutzrisiken zu minimieren und gleichzeitig die Datenqualität zu optimieren.

Die Quantifizierung der Datenqualität stellt eine komplexe Herausforderung dar, die vom spezifischen Anwendungskontext abhängt. Zum Beispiel können bei der Analyse der anonymisierten Daten vordefinierte statistische Eigenschaften beibehalten werden, indem Transformationsregeln für diese Eigenschaften festgelegt werden. In dieser Arbeit schlagen wir Richtlinien zur Auswahl von Qualitätsmodellen vor, die auf verschiedene Nutzungsszenarien zugeschnitten sind. Dies haben wir erreicht, indem wir verschiedene allgemeine Qualitätsmodelle in das weit verbreitete Datenanonymisierungswerkzeug ARX integriert haben. Zusätzlich führten wir umfangreiche experimentelle Vergleiche durch, um verschiedene relevante Aspekte zu untersuchen. Unsere Ergebnisse deuten darauf hin, dass verschiedene Qualitätsmodelle am besten für bestimmte Nutzungsszenarien geeignet sind. Insbesondere das Qualitätsmodell "Non-Uniform Entropy" eignet sich besonders gut für allgemeine Anwendungsfälle.

Auch wenn das Nutzungsszenario im Voraus bekannt ist, ist die effektive Modellierung der Datenqualität komplex. Zum Beispiel ist es im Kontext der datenschutzerhaltenden statistischen Klassifikation, bei der Vorhersagemodelle aus anonymisierten Daten erstellt werden, um den Wert eines Klassenattributs basierend auf einer Reihe von Merkmalen vorherzusagen, wichtig, den Informationsverlust für die Merkmale zu minimieren, da diese für ein bestimmtes Klassenattribut äußerst diskriminierend sind.

Wir präsentieren eine flexible Lösung zur Entwicklung und Bewertung von datenschutzerhaltenden Vorhersagemodellen, die verschiedene Vorhersagemodelle in Kombination mit einer Vielzahl von datenschutzerhaltenden Techniken unterstützt. Drei Fallstudien werden vorgestellt, um die praktische Anwendbarkeit unserer Lösung zu veranschaulichen.

Während die Formalisierung der Quantifizierung der Datenqualität die Herausforderung angeht, ist eine effiziente Softwareunterstützung unerlässlich, um diese Qualitätsmodelle im Kontext der Datenanonymisierung effizient zu nutzen. Die Datenanonymisierung ist eine komplexe Problemstellung, die Modelle zur Definition von Transformationsregeln, zur Quantifizierung von Datenschutzrisiken und Datenqualität sowie zur Implementierung von Anonymisierungsalgorithmen umfasst. Wir stellen einen innovativen Ansatz vor, der es ARX ermöglicht, nahezu beliebige Kombinationen von Anonymisierungstechniken zu unterstützen. Durch umfangreiche experimentelle Vergleiche mit bestehenden Lösungen zeigen wir eine überlegene Skalierbarkeit und Datenqualität auf, die eine Unterstützung für eine breitere Palette von Methoden und Techniken bieten.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction and Objectives

Modern data-driven biomedical research approaches based on big data processing and artificial intelligence provide enormous potential for advances in personalized medicine, new insights into the development and course of diseases, and novel clinical decision support methods [HF11]. Extensive data collections must be established to unlock this potential, requiring sensitive individual-level data to be collected collaboratively, shared with third parties or used for secondary purposes.

Individuals' privacy must be protected when implementing such data-driven approaches to biomedical research. This becomes increasingly challenging as data privacy involves ethical, societal, and legal aspects, which require consideration of privacy concerns and restrictions imposed by national and international data protection laws [JJB12]. Examples are the US Health Insurance Portability and Accountability Act (HIPAA) [HIP], the European General Data Protection Regulation (GDPR) [Eur16], and the Chinese national standard on the protection of personal information [Sta18]. To help make decisions about the use of sensitive or confidential data, the *Five Safes* framework addresses data privacy on multiple levels [DRW16]: (1) *Safe projects* is concerned with organizational measures that ensure that data use is appropriate. (2) *Safe people* means that people working with the data are safe and trustworthy. (3) *Safe data* implies that re-identification risks are reduced to an acceptable minimum. (4) *Safe settings* means that risks of privacy breaches during data processing are reduced. (5) *Safe outputs* requires that disclosure risk of output data is controlled such that results do not leak sensitive personal information.

Data anonymization is an important concept for achieving safe input (3) and output (5) data. Here, data is transformed in such a way that privacy risks are minimized while data quality is maximized at the same time. While the term data anonymization has been established in many European countries, in other areas such as the U.S., *de-identification* may be used instead. Data anonymization involves the following concepts: (1) **Transformation models**, which specify data transformation rules,

(2) **Privacy models**, which formally specify and quantify privacy risks, (3) **Quality models**, which formally quantify data quality, and (4) **Anonymization algorithms**, which typically search a given solution space for (semi)-optimal solutions. Here, the solution space consists of transformations resulting from applying transformation rules to input data. In this process, the algorithm guarantees that a predefined threshold of privacy risks is met according to a privacy model and that maximal data quality is achieved according to a quality model.

Quantifying data quality is a complex issue, as the usefulness of data heavily depends on the context. For instance, when it is known that specific statistical properties shall be analyzed in the anonymized data, it is possible to establish transformation rules to keep these properties intact. This thesis addresses three problem areas involving data quality in the context of data anonymization. Solution proposals to overcome these challenges have been published by this thesis' author (with a first or shared first authorship) as full papers in international, peer-reviewed journals and conference proceedings.

Outline

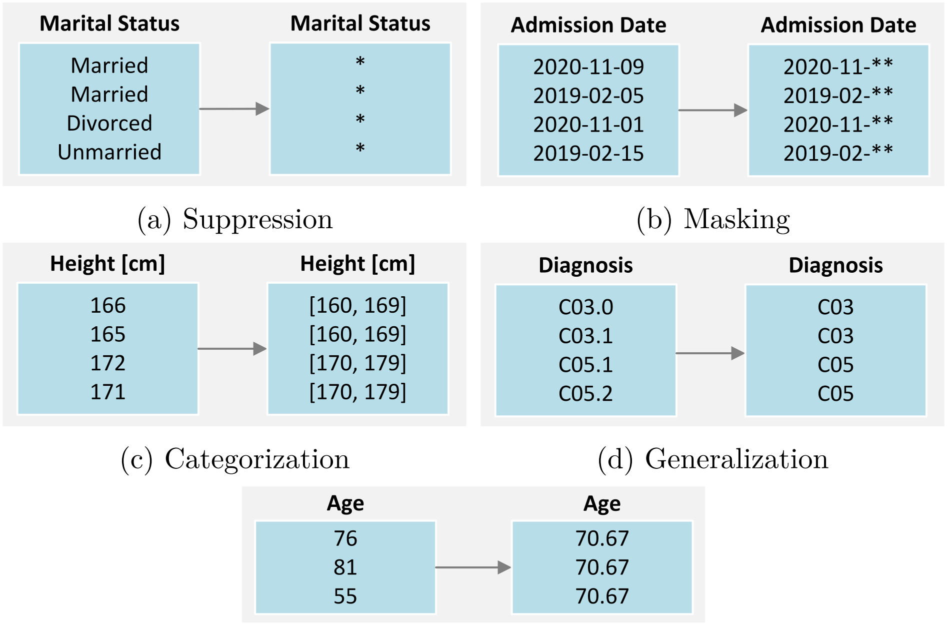
This thesis is publication-based and structured as follows: Chapter 1 introduces the topic as well as all relevant concepts and definitions. The chapter also describes the challenges this thesis tried to overcome. Chapter 2 provides an overview of the contributions, which have been published as journal articles, as responses to these challenges. Chapter 3 concludes the thesis by discussing the presented material and prior as well as future work. Appendix A provides detailed information about the publications. Finally, Appendix B lists all further publications to which this thesis author has contributed as co-author and which have been published during the period of this doctoral thesis.

1.1 Background

This section provides an extensive overview of the concepts and techniques used in this thesis. First, the concepts of **data transformation** and **privacy models** are introduced. Next, special attention is given to the concept of **quality models**, as these are the core of this thesis. Finally, in the last section, all three concepts are combined by a rather extensive introduction to **data anonymization algorithms**.

Data Transformation

The obvious first step in data anonymization is to remove all attributes from the dataset that are directly identifying, e.g. social security numbers [FWFY10]. However, the remaining attributes are potentially identifying if used in combination [Swe02a]. Thus,



(a) Suppression

(b) Masking

(c) Categorization

(d) Generalization

(e) Aggregation

Figure 1.1: Example transformations for different attributes using the different transformation methods.

the next step, which is more challenging, is to transform these so-called *quasi-identifiers* in such a way that it becomes very difficult for an adversary to link an identified or identifiable individual to the dataset and thus disclose sensitive information about this individual [NS08, Swe02a]. This process typically involves different *transformation methods*, which can also be used in combination.

Figure 1.1 shows example transformations for the transformation methods described as follows. **Aggregation** is a method where attribute values across multiple records are transformed into a common aggregate. In the example, the values of the attribute *age* were aggregated using the arithmetic mean. **Suppression** is a method where complete attribute values or even whole records are removed from the dataset. In the example, the values of the attribute *marital status* were suppressed (indicated by a semantic-free placeholder "*"). **Masking** is a method where attribute values or certain characters of these values are replaced by modified or even new values. In the example, the last two digits of the admission date, i.e. the day of the month, were masked by replacing them with a semantic-free placeholder. **Categorization** is a method where continuous attribute values are mapped to categories. In the example, the values of the attribute *height* were replaced by the categories $[160, 169]$ and $[170, 179]$, respectively. **Generalization** is a method where attribute values are iteratively replaced by less specific or less precise values. In the example, values of the attribute *Diagnosis* were replaced based on the International Classification of Diseases (ICD) [Wor].

Instead of using classifications, the replacement values can also be chosen based on user-defined generalization hierarchies, which is a typical strategy when anonymizing health data [EEDI+09, EEA13, XHD+15]. Generalization hierarchies are well-suited specifically for categorical attributes but can also be used for continuous attributes, e.g. by performing categorization [PKK16a].

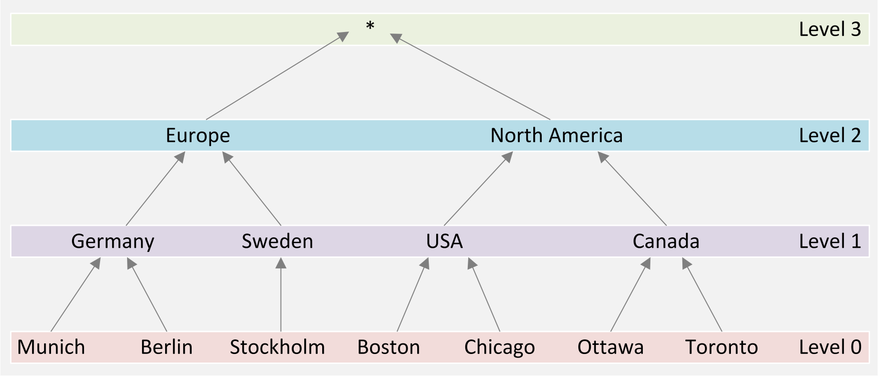
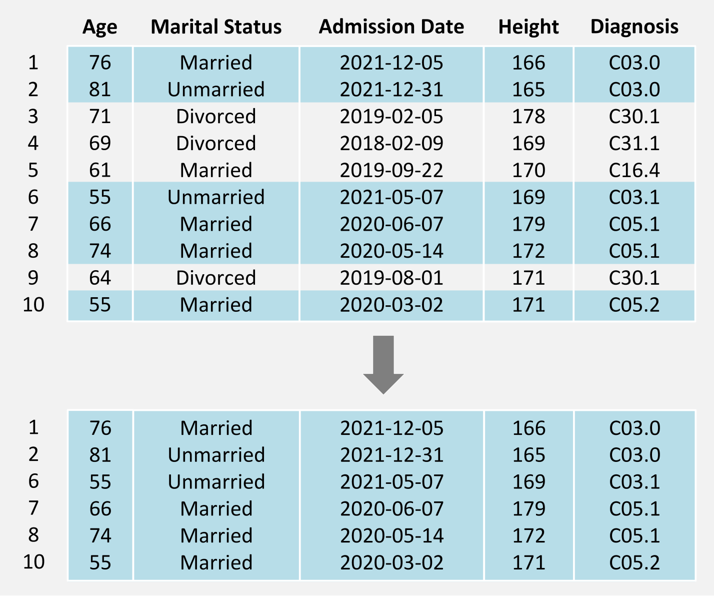


Figure 1.2: Example user-defined hierarchy for the attribute *city*.

A simple example of such a user-defined generalization hierarchy for the attribute *city* is shown in Figure 1.2. The hierarchy consists of a set of increasing levels, which specify attribute values with decreasing precision. The values are first transformed into countries, then continents, and finally, suppressed (indicated by the character ”*“).



	Age	Marital Status	Admission Date	Height	Diagnosis
1	76	Married	2021-12-05	166	C03.0
2	81	Unmarried	2021-12-31	165	C03.0
3	71	Divorced	2019-02-05	178	C30.1
4	69	Divorced	2018-02-09	169	C31.1
5	61	Married	2019-09-22	170	C16.4
6	55	Unmarried	2021-05-07	169	C03.1
7	66	Married	2020-06-07	179	C05.1
8	74	Married	2020-05-14	172	C05.1
9	64	Divorced	2019-08-01	171	C30.1
10	55	Married	2020-03-02	171	C05.2

1	76	Married	2021-12-05	166	C03.0
2	81	Unmarried	2021-12-31	165	C03.0
6	55	Unmarried	2021-05-07	169	C03.1
7	66	Married	2020-06-07	179	C05.1
8	74	Married	2020-05-14	172	C05.1
10	55	Married	2020-03-02	171	C05.2

Figure 1.3: Example transformation using Random Sampling.

Random Sampling is a method where the dataset is sampled, meaning records are randomly selected. This method can decrease an adversary's confidence about the success of a re-identification attack. In the example in Figure 1.3, records 1, 2, 6, 7, 8 and 10 were sampled.

Privacy Models

When sensitive individual-level data is collected collaboratively, shared with third parties or used outside of their original purpose, the involved individual's privacy must be protected against adversarial attacks. Three main types of *privacy disclosure* scenarios are commonly addressed by privacy models [LLZM10]: **Membership disclosure** means an adversary can learn whether an individual's information is contained in a specific dataset [NAC07]. In this case, the adversary is not able to link an individual to a specific data entry (row or column). However, they can infer information. For instance, if the dataset only contains information about diabetes patients, the adversary can infer that the found individual has diabetes. **Attribute disclosure** means that an adversary can learn whether an individual's information is in a specific set of rows in a dataset [MKGV07]. Like membership disclosure, the adversary can infer information about the individual without linking them to a specific data entry. For instance, if the attacker can learn that the individual's information is contained in a set of records that share a particular sensitive attribute value, then the attacker can learn this information. **Identity disclosure**, also termed *re-identification*, means an adversary can link an individual to a specific record in a dataset [Swe01]. This implies that the adversary can learn all sensitive information about the individual in the dataset.

In addition to privacy disclosure scenarios, privacy models also address *privacy threat* scenarios where factors such as objectives and intent, existing background knowledge of adversaries, replicability, and distinguishability of the data to be protected play an important role [Ema13]: In the **Prosecutor model**, it is assumed that the adversary already knows that the data about a targeted individual is contained in a data set (i.e. membership disclosure has been achieved pre-attack). The goal is to re-identify this specific individual to learn which exact record belongs to the individual. To this end, the distinguishability of records in the dataset regarding the quasi-identifiers can be used to calculate the re-identification risk [PKK16b]. This is a worst-case scenario. However, it has been shown that typically this method overestimates the risks [BJ12]. In the **Journalist model**, the adversary aims to re-identify an arbitrary individual without prior knowledge about membership. Since individuals represented in a dataset are typically a sample of a larger population, the assumption about background knowledge of the journalist model is believed to be more realistic than that of the prosecutor model. According to the journalist model, the re-identification risk can be calculated using a population table. However, information about the whole population, meaning all individuals in the population, is typically not available and thus, determining the risk of successful attacks according to the journalist model is difficult. In the **Marketer model**, in this scenario, the adversary aims to re-identify as many arbitrary individuals as possible without prior knowledge about membership. This attack can only be worthwhile if a non-trivial number of records can be re-identified. Therefore, according to the marketer model, the risk of successful attacks can be expressed as an average of the re-identification risks of all records.

The primary threat from which datasets are typically protected is *re-identification* [US 02,Eur16]. If successful, this type of attack can have significant legal consequences for data owners in many jurisdictions worldwide. Several high-profile attacks aiming at re-identification have shown that protecting against this type of attack is a complex issue [Leo12,EJAM11]. For example, re-identification attacks can typically not be prevented successfully by simply removing directly identifying attributes, such as social security numbers or names [DESG11,NS08,Swe01]. For this reason, more formal privacy models are required, which typically employ mathematical and statistical models to quantify the privacy risks.

k-anonymity is the most well-known privacy model for protecting data from re-identification or identity disclosure. A dataset is said to be k-anonymous if, regarding the quasi-identifiers, each record is indistinguishable from at least $k-1$ other records [Swe01]. To this end, the model forms equivalence classes by grouping records according to the quasi-identifier values. Records within the same equivalence class cannot be distinguished from each other. Consequently, an adversary can only link an individual to a group of records; thus, the probability of correct linkage is no more than $1/k$. k-anonymity belongs to the class of privacy models, which account for the distributions of attribute values within equivalence classes (also called groups) in a dataset (also called sample). k-anonymity enforces the risk threshold on the groups of the given sample and thus aims to protect from prosecutor attacks. k-anonymity and other similar models addressing re-identification risks constrain the quasi-identifiers but ignore sensitive attributes. If all records of one group share the same sensitive attribute value, an attacker can learn this information about an individual without linking the individual to a specific record. In order to counteract this problem, various extensions of the k-anonymity privacy model exist. The most well-known privacy models for protecting data from attribute disclosure are **ℓ-diversity** and **t-closeness**. ℓ-diversity requires that each equivalence class contains at least ℓ "well-represented" distinct values for each sensitive attribute [MKGV07]. The idea behind t-closeness is that equivalence classes must not "stand out". Therefore, the model requires that the distance between the distribution of sensitive attribute values in each equivalence class and the distribution of the sensitive attribute values in the original dataset is less than t [LLV07].

δ-presence is one of the most well-known privacy models to protect against membership disclosure. The model requires that the original dataset be modelled as a subset of a larger dataset that represents the adversary's background knowledge. Then, an anonymized dataset is said to be $(\delta_{min}, \delta_{max})$-present if the probability that an individual from the larger dataset is contained in the smaller dataset is between δ_{min} and δ_{max} [NAC07]. According to the δ-presence model, successful prevention of membership disclosure can indirectly lead to the prevention of identity and attribute disclosure. The reason is that if the probability of an individual being present in the dataset is at most δ %, then the probability of linking this individual to their record (re-identification) and sensitive attribute is also δ %. **ϵ-differential privacy(ϵ-DP)** is a strong privacy model that protects data from re-identification, attribute, and membership disclosure. Unlike most other privacy models, ϵ-DP does not apply privacy constraints to the dataset but to the mechanisms with which it is processed. The idea of ϵ-DP is that participating in a statistical database should not substantially affect an individual's privacy. To this end, the model guarantees that the anonymized data is independent from the contribution of individual records [Dwo11]. All previously described privacy models aim at minimizing privacy disclosure risks by guaranteeing specific levels of privacy protection. However, they do not account for the likeliness of an attack. The **game-theoretic** approach allows reasoning about re-identification, meaning how likely an attack will occur. To this end, the re-identification problem is analyzed from an economic perspective, assuming that an adversary will only launch an attack if tangible economic benefit is involved [WVX+15, PGW+17].

Privacy model	**Disclosure model**	**Threat model**
k-anonymity	Identity	All
ℓ-diversity	Attribute	All
t-closeness	Attribute	All
δ-presence	Membership	Journalist, Marketer
ϵ-DP	All	All
Game-theoretic model (prosecutor)	Identity	All
Game-theoretic model (journalist)	Identity	Journalist, Marketer

Table 1.1: Categorization of privacy models.

Table 1.1 shows an overview of all previously described privacy models and their categorization into disclosure as well as threat models. We note that if a privacy model protects a dataset against prosecutor attacks, that directly implies protection against journalist attacks. Moreover, the same applies when a privacy model protects against journalist attacks, i.e., it also protects against marketer attacks. Many more privacy models exist in the literature, as shown in a systematic survey by Wagner et al. [WE18]. However, this section focused on the privacy models integrated in ARX, an anonymization tool specifically tailored towards biomedical data [ARX]. The reason is that this thesis exists in the context of ARX as it is embedded in the development thereof.

Data Quality Models

Data anonymization inherently leads to the removal of information. Here, one optimization goal is to keep the data as *useful* as possible. However, formally defining the usefulness of data is a complex issue, as the nature of usefulness heavily depends on the use case. In the literature, two different classes of data quality models have been proposed, i.e. *general-purpose quality models* and *special-purpose quality models.* In the typical scenario where the purpose of the data, i.e., how the data will be analyzed, is unknown in advance, the former class can be used. In contrast, the latter class may be suitable if the usage scenario is known pre anonymization.

In the context of general-purpose quality models, Fung et al. propose measuring the *similarity* between the original and the anonymous data, where similarity can be defined differently [FWCY10]. Domingo-Ferrer et al. suggest a proper definition should capture the amount of information loss for a reasonable range of data uses [ZB15]. They define a dataset as having *little* loss of information if it is *analytically valid* and *analytically interesting.* The former property requires that specific statistical characteristics be preserved. On the other hand, a dataset is analytically interesting if specific attributes that are useful for further analyses remain intact [ZB15].

In contrast, if the usage scenario is known in advance, this knowledge can be taken into account during data anonymization to retain critical information. For example, consider the case where the data is used for statistical classification, a common use case for individual-level data, where the value of a predefined class attribute is predicted from a given set of values of feature attributes. In this context, it is essential to minimize the information loss of features that are most discriminating for the labels in the target attribute [FWCY10]. To this end, it is essential to distinguish between the removal of *noise* and the removal of *structure.* While removing noise is uncritical, removing structure may impact the suitability of data for this use case, i.e. statistical classification [FWY07].

Formal quality models either calculate *data quality* or a *reduction of data quality.* Reduction of quality quantifies how much information has been lost during the anonymization process, i.e., the original data set has a reduced quality value of 0 %, while this value is 100 % for a completely generalized or suppressed data set. Data quality, also termed *utility* or *Precision*, indicates how much information remains in the anonymized data set. Both measures can be used interchangeably by defining reduction of quality as the opposite of data quality and vice versa.

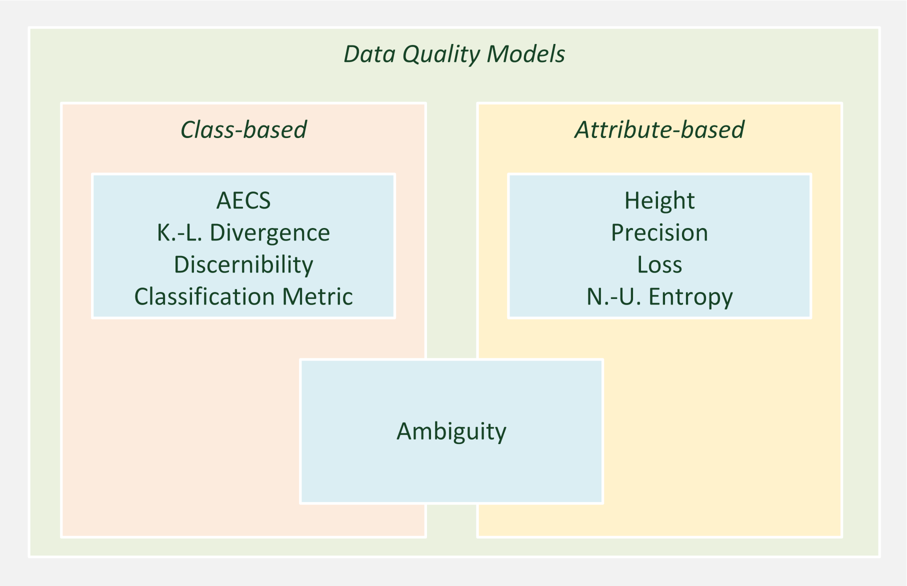


Figure 1.4: Categorization of quality models

Quality models can be interpreted in different ways. For example, there is a distinction between *class-based* and *attribute-based* models. While the former calculate the reduction of quality based on the sizes of the equivalence classes resulting from the anonymization process, the latter are based on the individual reduced quality values of each attribute of the data set. Then, as the name implies, attribute-based quality models retrieve one value for the reduced quality value of each attribute. As these values are possibly mapped into different intervals, they may be normalized to impose an equal weight on them. Lastly, the values obtained for each attribute in the data set might be aggregated differently to compile a quality value for the overall data set.

Figure 1.4 shows how seven general-purpose and one special-purpose quality models are categorized into the classes described above. It can be seen that classes may overlap. For instance, the Ambiguity metric is considered class-based as well as attribute-based. The remainder of this section comprises a detailed description of each of the eight quality models. Again, the focus lies on models which were integrated in ARX.

Average Equivalence Class Size (AECS) is a class-based and syntactic quality model proposed by LeFevre et al. It measures information loss based on the size of the equivalence classes resulting from a transformation. Thus, it only considers cardinalities and does not account for the actual values of the quasi-identifying attributes in the input data set [LDR06a]. **Discernibility** is a class-based and syntactic quality model introduced by Bayardo et al. Discernibility also measures information loss based on the size of the equivalence classes resulting from a transformation. It introduces a penalty for suppressed entries based on how many tuples in the transformed data set are indistinguishable from it [BA05]. **Height** is an attribute-based and syntactic quality model that measures information loss based on the overall distortion resulting from a transformation. It introduces a penalty for each instance of a value transformed, and then the overall distortion is the sum of all penalties. [Sam01]. **Precision** is an attribute-based and syntactic quality model proposed by Sweeney [Swe02a]. It measures data quality by reporting the amount of distortion in a transformed data set. The Precision metric is an extension of the Height metric. For each entry of the generalized data set, the ratio of the generalization level of an attribute to the height of the attribute's generalization hierarchy denotes this entry's distortion. The minimum generalization level is always 0 and represents the original data values. We note that as Precision measures data quality, the sign of the results can be inverted (by multiplying them with -1) to obtain a value for information loss. **Loss** is an attribute-based and syntactic quality model introduced by Iyengar [Iye02], which calculates the information loss. This measure considers the transformed data set and the given generalization hierarchies, where the idea is to quantify the loss induced by a transformation when an original value cannot be disambiguated from another value. For this purpose, each value of the transformed data set is penalized with a factor between 0 and 1, where the higher the factor, the more values cannot be distinguished from each other, i.e., are equal.

De Waal and Willenborg proposed the use of entropy as a measure for information loss in [DWW99]. The authors consider the entropy measure with global recoding and local suppression. In [GT09], Gionis and Tassa introduce a slight variation called **Non-Uniform Entropy (N.-U. Entropy)**, which, in contrast to the initially proposed measure, increases monotonically with increasing generalization. This measure is frequently used in scientific works, e.g., in [DFT01], [EEDI$^+$09], and it has also recently been recommended in a guideline for anonymizing health data [EEA13]. However, the measure proposed by [GT09] is unsuitable for generalization and suppression, as the latter type of recoding may increase Non-Uniform Entropy and thus reduce information loss. As a consequence, a variation of Non-Uniform Entropy can be used as an attribute-based and semantic quality model. The model adopts the principles from [DWW99] by separately analyzing tuples from the data set that have been suppressed and tuples from the data set that have been generalized. **Kullback-Leibler Divergence (K.-L. Divergence)** is a class-based and semantic quality model that is based on information theory and is a measure of the difference between two probability distributions. Precisely, it measures the information lost when one distribution is used to approximate the other. Li et al. ([LLV07]), Machanavajjhala et al. ([MKGV07]) and most recently, Xia et al. [XHD$^+$15] used the KL-divergence in order to measure the data utility when anonymizing data via generalization. **Ambiguity Metric**, a semantic quality model that is both attribute- and class-based, was introduced by Nergiz et al. The metric defines the information loss as the average size of the Cartesian products of all generalized entries for each tuple in the data set [NC06]. In other words, it calculates the number of possible combinations of input tuples that a generalized tuple stands for.

Classification Metric (CM), a syntactic and class-based quality model, was introduced by Iyengar specifically for statistical classification. The idea is to minimize the loss of information of feature attributes best suitable for discriminating the target attribute [Iye02]. To this end, first equivalence classes are built by grouping records with the same feature attribute values. Then, records are penalized if they are suppressed or have a class label different from the majority class label of this class.

Anonymization Algorithms

As described in Section 1.1, the first step in data anonymization typically involves removing all directly identifying attributes from the dataset, followed by a more complex step where quasi-identifiers are transformed in such a way that it becomes very hard to successfully utilize these attributes for attacks. This step typically involves quantifying privacy risks and data quality by mathematical privacy and quality models. Then, data anonymization algorithms implement a **search procedure**, where all possible outputs are traversed in order to find a solution which satisfies privacy risks according to a predefined **privacy model** while at the same time providing “optimal” data quality according to a **quality model**. Figure 1.5 illustrates this abstract mechanism.

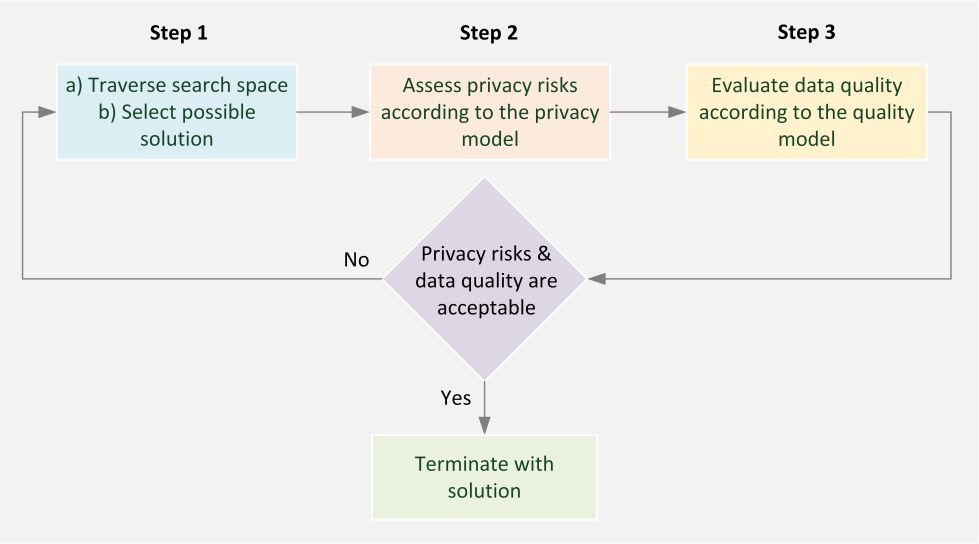

Step 1
Step 2
Step 3
a) Traverse search space b) Select possible solution
Assess privacy risks according to the privacy model
Evaluate data quality according to the quality model
Privacy risks & data quality are acceptable
No
Yes
Terminate with solution

Figure 1.5: Abstract mechanism used by anonymization algorithms

As described in Section 1.1, different types of transformation methods for transforming the quasi-identifiers exist. However, health data anonymization algorithms typically utilize generalization hierarchies. Here, the search space is typically modelled as a *generalization lattice*, which is a set of all possible combinations of generalization levels for all quasi-identifiers where each element represents a specific combination. The elements of the generalization lattice and, thus, the size of the search space and the amount of data distortion depend on the *generalization scheme*. The most widespread scheme is the so-called **full-domain generalization scheme**, where all values of a quasi-identifier are transformed to the same level of its hierarchy [LDR05, Sam01, Swe02b]. For instance, in Figure 1.2, if an occurrence of the city `Munich` is transformed to `Germany`, then it also requires transforming occurrences of `Chicago` to `USA`. Lesser known and used generalization schemes are **subtree generalization** and **sibling generalization**. In **subtree generalization**, all values represented by the same subtree of the hierarchy are either transformed to the same level or left unchanged [BA05, FWY05, FWY07, Iye02, LDR05]. For example, if an occurrence of the city `Munich` is transformed to `Germany`, then it also requires transforming `Berlin`, i.e. all other values represented by this subtree, to the same level. However, it does not require transforming occurrences of the city of `Chicago` to the same level, as this value belongs to a different subtree. **Sibling generalization** is similar to subtree generalization as it considers values subtree by subtree. However, while subtree generalization requires all values represented by a subtree to be treated the same, sibling generalization additionally allows for single values to be left unchanged [LDR05]. For example, suppose an occurrence of the city of `Munich` is transformed to `Germany`. In that case, an occurrence of the city of `Berlin` may or may not be transformed to the same interval. All three generalization schemes require that all instances of a value are treated the same. For example, if the city of `Munich` occurs in multiple rows of a dataset, all instances are transformed in the same manner or left unchanged. This scheme is called *global recoding*. In contrast, *local recoding*, also called **cell generalization**, allows for different instances of the same value to be transformed to different levels of the hierarchy or even left unchanged [LDR05, WLFW06, XWP$^+$06]. For example, the first instance of `Munich` may be transformed to level 1 of its hierarchy (`Germany`), while the second instance may be transformed to level 2 (`Europe`). Full-domain generalization, as well as subtree generalization, are *single-dimensional* generalization schemes, which means each quasi-identifier is transformed separately. In contrast, quasi-identifiers are generalized in groups in the **multidimensional generalization** scheme. For example, instead of transforming values of the attribute *country* first, followed by transforming values of the attribute *sex*, both attributes are transformed in one step [LDR06a, LDR06b]. This scheme allows for different instances of the same attribute value to be transformed to different generalization levels. For instance, `<Munich, Female>` might be transformed to `<Germany, Female>` while `<Munich, Male>` might be transformed to `<Europe, Male>`.

Generally, global recoding schemes produce smaller search spaces but higher data distortion than local recoding schemes. In particular, full-domain generalization has the smallest search space but the largest distortion. On the other hand, cell suppression has the largest search space but the least distortion. In order to decrease data distortion, generalization schemes are often combined with *suppression schemes*. The most common is **record suppression**, where complete records are removed from a dataset [BA05, Iye02, LDR05, Sam01]. **Value suppression** means that **all** instances of a specific value are removed (e.g. all occurrences of the age value `20`) [WFP07], while during **cell suppression** (also called *local suppression*), only **some** instances of a given value are removed [Cox80, MW04].

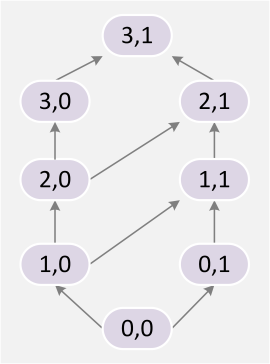
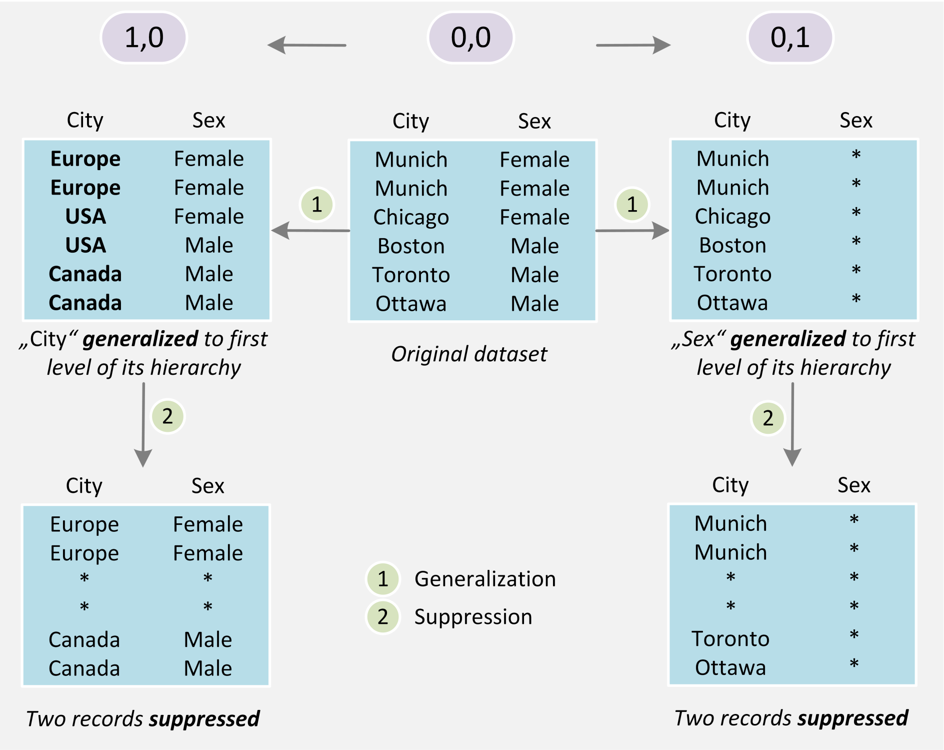


Figure 1.6: Example generalization lattice for the attributes *city* and *sex*. The first digit of the tuple refers to the former attribute and the second digit to the latter.

Figure 1.6 shows the generalization lattice constructed from the hierarchies for the attributes *city* and *sex*. While the hierarchy for the attribute city consists of 4 levels (0 - 3), as shown in Figure 1.2, the hierarchy for the attribute *sex* comprises two levels since values can only be suppressed. Each node represents a single transformation, which defines generalization levels for all quasi-identifiers in the dataset. In the above example, quasi-identifiers consist of the attributes *city* and *sex*. An arrow is drawn between a transformation and its successor to indicate that the successor represents a direct generalization of this transformation and that the two transformations differ by precisely one generalization level. The transformation `(0,0)` represents the original dataset, meaning neither values of the attribute *city* nor values of the attribute *sex* are generalized. In contrast, the transformation `(3,1)` represents the maximal generalization, which means that both attributes are generalized to the maximal level of their respective hierarchies.


1,0
0,0
0,1
1
1
2
2
„City“ **generalized** to first level of its hierarchy
Original dataset
„Sex“ **generalized** to first level of its hierarchy
1 Generalization
2 Suppression
Two records **suppressed**
Two records **suppressed**

City	Sex
Europe	Female
Europe	Female
USA	Female
USA	Male
Canada	Male
Canada	Male

City	Sex
Munich	Female
Munich	Female
Chicago	Female
Boston	Male
Toronto	Male
Ottawa	Male

City	Sex
Munich	*
Munich	*
Chicago	*
Boston	*
Toronto	*
Ottawa	*

City	Sex
Europe	Female
Europe	Female
*	*
*	*
Canada	Male
Canada	Male

City	Sex
Munich	*
Munich	*
*	*
*	*
Toronto	*
Ottawa	*

Figure 1.7: Example dataset including output datasets for attributes *city* and *sex* when applying generalization and suppression.

Figure 1.7 shows output datasets for the transformations (1,0) and (0,1) resulting from applying full-domain generalization to the input dataset, followed by record suppression. Referring to the abstract depiction of data anonymization algorithms in Figure 1.5, the “optimal” transformation can be found by traversing the lattice node by node. For each node, the transformation is applied to the dataset (Step 1). All records that violate the privacy constraint are suppressed (Step 2), and eventually, the quality is calculated according to a given privacy model. Finally, the transformation with the highest quality is returned as the solution.

1.2 Challenges

Secondary use of data in the biomedical domain, which means using data for research when it was initially collected for clinical care, is gaining support from businesses and governments alike [McK11, WDA+16, Com14]. Using data for biomedical research bears significant potential in areas such as epidemiology, public health, quality improvement or data privacy [FAWJ18]. However, secondary use of health data poses quite challenging in practice, mainly due to a lack of data quality. Reasons for data quality issues are complex. First, the incentive for data collection in the first place may have an impact on the accuracy. For instance, financial or contractual obligations mandate consistent data entry; the resulting data may be more accurate than when data is only used for internal purposes [ASS$^+$11]. Typically, priorities regarding data collection in the clinical vs. research setting may differ significantly. Consequently, it is not surprising and thus generally accepted that clinical data are not collected with the same attention as research data [WW13]. Further, data quality issues may arise from inadequate data entry, which can stem from both software and human deficiencies. For instance, the software might exude a flawed design where multiple fields are provided for the same data point. Moreover, a lack of documentation relating to data entry might lead to a difference in documentation practices among different personnel, e.g. the same field could be used for documenting a scheduled as well as the actual date of an event. Similarly, if the used vocabulary is not standardized, artificial differences in data entries could arise [ASS$^+$11].

As described above, data quality issues are very common when data is used for secondary purposes. When this data is used in a setting where data anonymization is necessary, it is easy to see that these quality issues become amplified because data anonymization inherently leads to loss of information. Therefore, minimizing the decrease in data quality during anonymization is crucial. As described in section 1.1, quantifying and adequately utilizing the concept of data quality in the data anonymization context is a complex task. Therefore, this thesis addresses three different problem areas relating to data quality:

C1. Modelling data quality for general purposes The usefulness of data often depends on the use case. Therefore, modelling data quality becomes a complex issue when the usage scenario is unknown. To overcome this, measuring data quality requires formal models which quantify data quality for general purposes. A multitude of such data quality models have been proposed in the literature, which typically define an increase in information loss as well as a decrease in data quality in order to be able to distinguish interesting from uninteresting data. However, prior to the work presented in this thesis, it was unclear which quality model is suited best for which usage scenario because a systematic evaluation was missing.

C2. Modelling data quality for specific purposes Even when the usage scenario is known in advance, modelling data quality is a non-trivial issue. For instance, in the context of privacy-preserving statistical classification, prediction models are built from anonymized data to predict a class attribute's value from a set of feature attributes. Here, it is important to minimize the loss of information for the features as these are most discriminating for a specific class attribute.

C3. Enabling anonymization tools to optimize their output regarding different quality models Assuming that the two problems mentioned above can be solved by finding the proper formalization of data quality in a specific context, the next problem that arises is that in order to properly utilize these quality models, software support is needed. As described in section 1.1, data anonymization is a complex task involving *transformation* and *privacy models* as well as *anonymization algorithms* and *quality models*. Since data quality heavily depends on the usage scenario, it is crucial that data anonymization software supports a wide range of models and combinations thereof. In order to achieve this goal, the software has to be very flexible in its design and provide scalability when handling high-volume and high-dimensional data.

Solution proposals to overcome the challenges addressed in this doctoral thesis have been published by this thesis' author (with a first or shared first authorship) as full papers in international, peer-reviewed journals and conference proceedings.

CHAPTER 2

Methods and Solutions

As described in section 1.2, this doctoral thesis addresses challenges in three different problem areas related to data quality in the context of data anonymization. For each of these problem areas, a solution proposal has been published as a full paper in an international, peer-reviewed journal or conference proceeding. We refer to these solution proposals as contributions **Ref.A.1**, **Ref.A.2** and **Ref.A.3**.

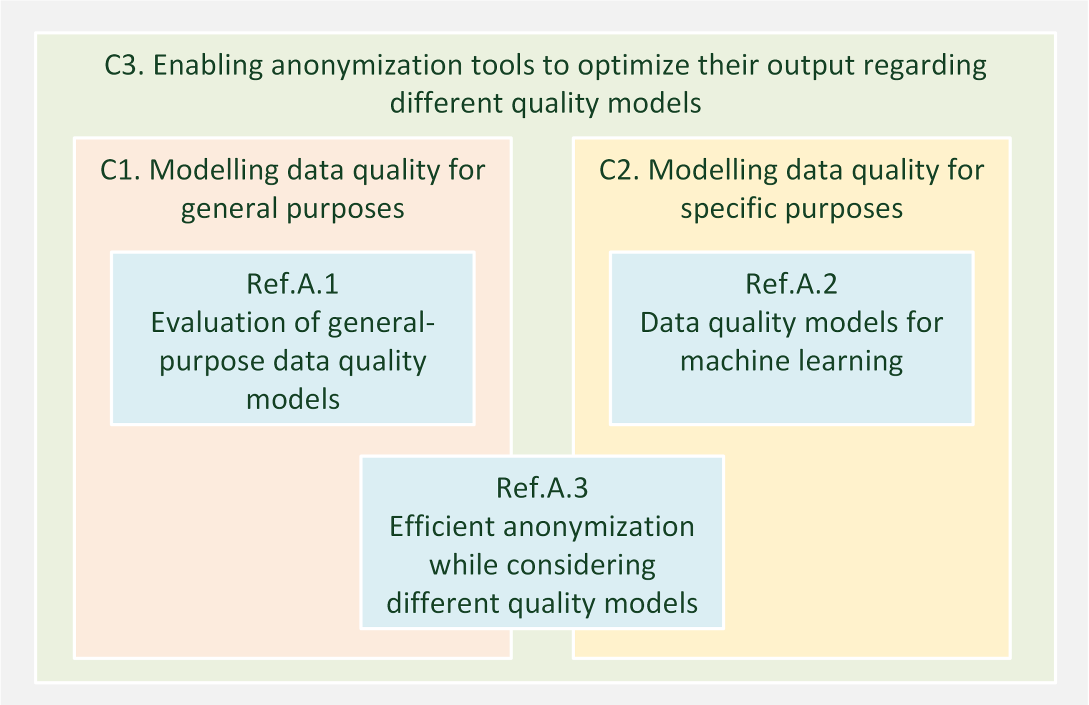

C3. Enabling anonymization tools to optimize their output regarding different quality models
C1. Modelling data quality for general purposes
C2. Modelling data quality for specific purposes
Ref.A.1 Evaluation of general-purpose data quality models
Ref.A.2 Data quality models for machine learning
Ref.A.3 Efficient anonymization while considering different quality models

Figure 2.1: Visual representation of the three problem areas, or challenges, presented in the previous section in relation to the solutions Ref.A.1 to Ref.A.3, described in this section.

The solutions of all three contributions have been integrated into and released with the anonymization tool ARX [ARX]. ARX has been mentioned in national and international guidelines addressing various aspects in the context of privacy protection and data anonymization, such as quantitative risk assessment [Eur14], data sharing in general [Eur18], privacy-preserving implementations [Eur15], best practices [EMOT16] and many more [Off17, Min15, BK17, Per18, Pol18, Dut18, Min16]. Furthermore, ARX has been integrated in a multitude of different software collections [Fin18, Res19, Uni18, LMU19, Uni19, Kor17, TMF16, Tem17, ZB15, Tor17, RL19, KHF19] and has been used in several research projects [CCZYM17, KHC$^+$16, ABMS15, JKH$^+$17, SKH16, LZJ$^+$16, XJC$^+$15, PMG$^+$18].

Figure 2.1 shows a visual representation of the three problem areas, or challenges C1 - C3, introduced in the previous chapter and their relation to the three solution proposals described in this chapter. Since all three solution proposals have been integrated into the software ARX, C3 is addressed by all three proposals. Furthermore, it can be seen that Ref.A.1 and Ref.A.2. mainly address C1 and C2, respectively, while Ref.A.3 addresses both C1 and C2.

2.1 Evaluation of General-Purpose Data Quality Models

As described previously, measuring data quality is particularly challenging when the usage scenario is unknown at the time of anonymization. Various general-purpose data quality models have been proposed in the literature, but it is unclear which model is best suited for which usage scenario.

In **Ref.A.1**, we have taken the first step towards a guideline for selecting an appropriate quality model. For this purpose, we have implemented seven general-purpose quality models, i.e. AECS, Discernibility, Precision, Loss, Ambiguity, Kullback-Leibler Divergence, and Non-Uniform Entropy, into ARX. We used each quality model to anonymize a publicly available dataset containing patient discharge data according to the k-anonymity privacy model.

The overall goal of this work was to conduct an extensive experimental comparison in order to investigate (1) how the quality models influence the transformation of data, (2) how anonymized data relate to each other when obtained by using different quality models, and (3) how well-suited for real-world applications are anonymized data when obtained by using different quality models. We quantified the influence of each quality model on the transformation of data by the amount of generalization and suppression used during the anonymization process. In order to address the second question, we used each model to measure the data quality of the optimal solutions that were obtained from the other models. Finally, we selected a typical real-world application scenario, i.e. statistical classification, to evaluate the third research question. For this purpose, we used logistic regression to build privacy-preserving models representing the discharge dataset in order to predict a specific class attribute, i.e. the charge of hospital stays.

Our experimental evaluation showed that (in general) different models are best suited for different application scenarios. However, we also found that one general-purpose quality model, Non-Uniform Entropy, is well-suited for general-purpose usage scenarios. This is reasonable, as anonymized data contained instance-level as well as schema-level information when using this model. Moreover, we found that statistical power decreased by only 10 %, and the prediction model exhibited good accuracy.

Individual Contributions of Thesis Author: The thesis author has significantly contributed to the development and conceptual design of the research project. Moreover, the author has contributed to the gathering, collection, acquisition or provision of data, software or sources. Further, the author has significantly contributed to the analysis and evaluation or interpretation of data, sources and conclusions drawn from them. Finally, the author has contributed to the drafting of the manuscript.

2.2 Data Quality Models for Machine Learning

Machine learning models can be built from clinical, paraclinical and biomolecular data to detect unknown relationships between biomedical parameters. In this scenario, individual-level data of patients and probands will not be made public. However, the results from the machine learning process, i.e., the models, may be shared across sites. Therefore, it has to be ensured that the so-called prediction models cannot be used to extract sensitive information. A common solution to this problem is to anonymize the original data and use the results to build so-called privacy-preserving prediction models.

A variety of different data anonymization tools have been developed in recent years. At the same time, progress has been made with methods for measuring data quality in the context of machine learning. However, most tools lack support for privacy-preserving machine learning methods. In **Ref.A.2**, we aimed to bridge this gap by integrating various machine learning techniques into ARX. In previous work, we extended ARX with a method for optimizing the quality of anonymized data for use as training data for building prediction models. In addition, we implemented supervised learning to build logistic regression models from anonymized data.

The previous implementation of the quality model had major limitations: Only one class variable was permitted, the class variable had to be considered by the privacy model, and no transformations could be applied to target variables. We overcame these limitations by re-implementing a considerable portion of the internal code of the software. Further, we implemented a generic interface to support random forest and naïve Bayes in addition to logistic regression prediction models. Then, we integrated the approach with existing privacy models, such as Differential Privacy, and we created compatibility with further data transformation techniques, such as data aggregation. Finally, we improved the ability to assess prediction performance by adding various metrics and visualizations. To illustrate our solution's high degree of versatility, we presented three case studies. We showed that our tool is able to create privacy-preserving prediction models when different types of risks have to be mitigated in order to protect the individual's privacy.

Individual Contributions of Thesis Author: The thesis author has significantly contributed to the development and conceptual design of the research project. Moreover, the author has contributed to the gathering, collection, acquisition or provision of data, software or sources. Further, the author has significantly contributed to the analysis and evaluation or interpretation of data, sources and conclusions drawn from them. Finally, the author has contributed to the drafting of the manuscript.

2.3 Efficient Anonymization While Considering Different Data Quality Models

Data anonymization is a complex task as the effectiveness highly depends on the context. For instance, the data's dimensionality, volume and statistical properties must be considered. Many different algorithms, models and methods have been proposed in the literature. However, most algorithms only support a specific combination of models. Moreover, implementations are lacking, and the number of easy-to-use tools is small.

In **Ref.A.3**, we describe how we have extended ARX with a novel approach that enables the tool to support an (almost) arbitrary combination of quality and privacy models as well as transformation methods. The basis of ARX is a globally-optimal search algorithm which applies record suppression and full-domain generalization in order to transform the input data. Generalization is based on user-defined hierarchies. This design enables high flexibility in terms of data anonymization techniques, i.e. different privacy and quality models can be plugged into the system. However, it is quite inflexible when it comes to data transformation methods, i.e. support for methods besides global generalization is inadequate, which leads to data quality loss.

To overcome these limitations, we have designed and implemented an approach in which the basic algorithm of ARX is applied to different subsets of the dataset in an iterative manner. The main advantage is that with this approach, different generalization schemes may be applied to different subsets of the data, effectively enabling local generalization in addition to global generalization. This leads to a reduction of generalization overall and, thus, better data quality. Moreover, this approach facilitates the combination of almost arbitrary privacy and quality models, as well as transformation methods.

To emphasize the potential of our new approach, we conducted an extensive experimental comparison with other data anonymization tools using six different real-world datasets. Here, we focused on multi-dimensional and local generalization. Our results show that ARX often outperforms the competitors in terms of data quality and scalability.

Individual Contributions of Thesis Author: The thesis author has significantly contributed to the development and conceptual design of the research project. Moreover, the author has contributed to the gathering, collection, acquisition or provision of data, software or sources. Further, the author has significantly contributed to the analysis and evaluation or interpretation of data, sources and conclusions drawn from them. Finally, the author has contributed to the drafting of the manuscript.

CHAPTER 3

Discussion

This thesis presents solutions for different challenges related to data quality in the context of data anonymization. This chapter summarizes and concludes the work.

3.1 Principal Results

The starting point for the research in this thesis was based on the assumption that data quality depends on the usage scenario. We investigated whether certain quality models are suited for specific usage scenarios and whether certain quality models are fit for general purposes (**Ref.A.1**). Our results indicated that different quality models are or are not suitable for different application scenarios. For instance, *Discernibility* might be suitable for anonymizing small datasets (e.g. data about rare diseases). On the other hand, *AECS* might **not** be the best choice for predictive modelling. However, our results also indicate that *Non-Uniform Entropy* is best suited for general-purpose scenarios. Moreover, we selected one specific purpose, i.e. machine learning, to investigate further (**Ref.A.2**), as previous work in this area had shown that anonymization has a detrimental impact on data quality when used for machine learning tasks. By implementing the ability to create privacy-preserving prediction models into ARX, we were able to show that it is possible to achieve very good prediction performance and still provide a high level of privacy at the same time. Finally, we realized that software support for different methods is needed to cover a broad spectrum of usage scenarios, particularly for various data quality and transformation models. Therefore, in **Ref.A.3**, we presented a novel approach which enables ARX to support a multitude of quality, privacy and transformation models in combination. By means of an extensive experimental evaluation, we were able to show that ARX often outperforms competitors. The success of our solution is illustrated by the fact that ARX has been used in many different guidelines, software collections and research projects.

3.2 Prior Work

Previous works investigating data quality in the context of data anonymization have obtained conflicting results. Some have found that anonymization has only a small impact on data quality, e.g. [KL06], while others found a non-trivial impact, e.g. [PE07]. Findings from further studies indicated that data quality may depend on both the anonymization and the analysis methods [LP04, CK06]. Since then, one particular quality model, i.e. Non-Uniform Entropy, has frequently been recommended for anonymizing biomedical data, e.g. in [EEA13]. In **Ref.A.1**, we were able to confirm this finding, particularly for the general-purpose setting.

Name	**Discl. Type**	**Opt.**	**Transf.**	**Metric**
MinGen	ID	✓	FDG,RS	Precision
Binary Search	ID	✓	FDG,RS	Precision
Incognito (ℓ-diversity / t-closeness)	(AD) ID	✓	FDG,RS	*
OLA	ID	✓	FDG,RS	*
K-Optimize	ID	✓	STG,RS	*
μ-argus	ID	✗	STG,CS	Precision
Datafly	ID	✗	FDG,RS	-
Improved Greedy Heuristics	ID	✗	FDG,RS	-
Genetic Algorithm	ID	✗	STG,RS	CM
Bottom-Up Generalization	ID	✗	STG	-
Top-Down Specialization	ID	✗	STG,VS	-
Mondrian Multidimensional	ID	✗	MDG	DM
InfoGain Mondrian	AD	✗	MDG	N.-U. Entropy
Top-Down Disclosure	AD	✗	VS	-
SPALM	MD	✓	FDG	DM
MPALM	MD	✗	MDG	Heuristics
Flash	*	✓	*	*
Lightning	*	✗	*	*

Table 3.1: Overview of data anonymization algorithms. The wildcard character "*" indicates a flexible design or that multiple values are possible. For example, the Flash algorithm may be used to counteract identity disclosure but also attribute disclosure. *(FDG: Full-Domain Generalization; STG: Subtree-Generalization; MDG: Multidimensional Generalization; RS: Record Suppression; CS: Cell Suppression; VS: Value Suppression*

For the specific use case of machine learning, early research suggested that anonymization could compromise the usefulness of data [BS08]. Subsequently, methods were developed to optimize anonymized data for training prediction models, disproving this notion. These methods initially centered on basic anonymization techniques like k-anonymity and simple prediction models in distributed settings [AP08, IKB09]. As a result, assessing the effectiveness of anonymization methods for predictive modelling became a common practice in academia [FWFY10, MKH17]. More recently, a wider range of prediction and privacy models have been explored. Some introduced general-purpose anonymization algorithms to enhance prediction performance [LLBW11, LTZS14], while others focused on privacy-preserving algorithms tailored to specific prediction models [LC11, FWJ12]. In **Ref.A.2**, we built upon this research and enhanced ARX with privacy-preserving machine learning techniques.

A great number of data anonymization algorithms have been proposed in the literature. Table 3.1 shows an overview of some of the most well-known algorithms, including the disclosure threat they aim to prevent, as well as the transformation scheme and quality models they are designed to use. The table includes both optimal and heuristic algorithms. Notably, there are two outliers, i.e. *Flash* and *Lightning*. **Flash** was originally proposed by Kohlmayer et al., together with a generic implementation framework. This enabled the usage of globally optimal algorithms that utilize full-domain generalization to obtain k-anonymous datasets [KPE$^+$12]. The framework was adopted in ARX, which, later on, was extended to support further privacy and quality models, as well as transformations and combinations thereof [PKLK14]. Furthermore, Prasser et al. integrated the **Lightning** algorithm, a heuristic counterpart to Flash [PBE$^+$16]. As we showed in **Ref.A.1-3**, embedding those two algorithms in the generic implementation framework of ARX enabled the software to support an almost arbitrary combination of anonymization techniques.

However, most approaches are tailored towards one single threat scenario represented by a specific privacy model, and they calculate data quality according to one specific quality model. Early data anonymization algorithms mainly address the identity disclosure threat. To this end, they focus on the k-anonymity privacy model. Sweeney proposed an algorithm called **MinGen**, which utilizes full-domain generalization with record suppression and exhaustively searches the whole lattice to find the globally optimal solution according to the Precision metric [Swe02b]. Samarati proposed a **Binary Search** algorithm that exploits a lattice's monotonicity property. In this context, monotonicity means that if a node at level k is not k-anonymous, each node at levels $\geq k$ is k-anonymous. Given the lattice's height h, the algorithm starts its search at level $h/2$. If a k-anonymous node is found, the search proceeds at the lower level $h/4$. Otherwise, it proceeds at the higher level $3h/4$. The search terminates when a level with at least one k-anonymous node is found while no k-anonymous node is at a lower level. In the case of multiple k-anonymous nodes at the same level, the globally optimal solution is identified according to the Precision metric [Sam01]. LeFevre et al. presented the **Incognito** algorithm, which takes advantage of the fact that if a transformed subset of the quasi-identifiers is not k-anonymous, neither is the whole dataset [LDR05]. Thus, the algorithm creates generalization lattices of each quasi-identifier subset and traverses them in a bottom-up, breadth-first manner. In this process, if a smaller subset of quasi-identifiers is found to not be k-anonymous, each larger subset can be predictively tagged as not k-anonymous as well and thus pruned from the lattice in further iterations of the search. As opposed to MinGen and the Binary Search algorithm, Incognito has been designed as metric agnostic, which means it can easily be adapted to include any quality model in order to find the globally optimal solution. An algorithm called **Optimal Lattice Anonymization (OLA)**, which utilizes a divide-and-conquer approach, was proposed by El Emam et al. [EEDI+09]. In order to speed up the search, the algorithm constructs sublattices and uses predictive tagging to exclude nodes from the search. The algorithm starts with the whole lattice and constructs sublattices for each node at level $h/2$ (the middle of the lattice). One node is randomly chosen as a starting point. The node is checked for anonymity; if it is not anonymous, all predecessor nodes are not anonymous either and can be pruned (excluded from the search). In that case, the search continues with a sublattice at level $(h/4) + 1$. On the other hand, if the node was anonymous, the search is continued at level $(h/4) - 1$, where the process repeats. The algorithm terminates when all sublattices have been visited. The choice of quality metric for determining the globally optimal solution is relatively flexible; in their original work, the authors have presented experimental results when using the Precision, Discernibility and Non-Uniform Entropy metrics. Like MinGen, the Binary Search algorithm, Incognito and OLA utilize full-domain generalization with record suppression. Bayardo and Agrawal proposed **K-Optimize**, a globally optimal algorithm which utilizes record suppression but subtree generalization instead of full-domain generalization [BA05]. Here, the search space is modelled by a set enumeration tree where each set is a power set over special alphabets generated from an ordered attribute's domain. The algorithm starts at the most general node and utilizes pruning on the successors of this node if it is found to be anonymous. The authors present the algorithm utilizing two different metrics, i.e., the Discernibility and Classification metrics. However, like the Incognito algorithm, K-Optimize has been designed as metric agnostic, which means any quality model may be adopted.

While these five algorithms produce a globally optimal solution, it has been proven that this problem is NP-hard [MW04]. To overcome this issue, various heuristic approaches have been proposed. These approaches guarantee a so-called *minimal solution*, i.e., a solution that is not optimal but "good enough" (e.g. [Swe98, BRK+13]). The following section covers a selection of these heuristic algorithms.

Hundepool et al. presented **μ-argus**, where subtree generalization and cell suppression are greedily applied to combinations of domain values that occur less frequently than the specified value of k. The minimal solution is chosen based on the Precision metric. The algorithm considers all possible combinations of attributes of size two and three. Thus, the resulting dataset may not be protected from linkage when more than three attributes are used [HW96]. **Datafly**, proposed by Sweeney, was the first algorithm to handle real-world datasets. The algorithm traverses the lattice in a bottom-up manner, prioritizing transformations associated with the highest number of distinct values, and applies full-domain generalization and record suppression. The algorithm terminates when a k-anonymous node is found [Swe98]. While a metric guides Datafly's search, it does not properly utilize a quality model or cost metric. The **Improved Greedy Heuristic**, proposed by Babu et al., is an extension of Datafly where the decision about which node to visit next is made based on the smallest possible minimal equivalence class size. When multiple transformations are found, the algorithm falls back on Datafly's strategy [BRK$^+$13]. Based on a lattice built from generalization hierarchies for a set of quasi-identifying attributes, Iyengar's **genetic algorithm** proposes to encode each lattice node as a "chromosome". The algorithm then utilizes subtree generalization and record suppression in order to find the "fittest" solution, where fitness is a metric for the amount of data distortion measured by the Classification metric [Iye02]. Wang et al. proposed a **Bottom-Up Generalization** algorithm, which starts with the node representing the original data and then traverses the lattice in a bottom-up manner to find a minimal solution for a classification task while utilizing subtree generalization [WYC04]. The search is guided by the *ILPG* metric [FWY05, FWY07], a so-called *Trade-Off Metric*. As the name implies, these types of metrics aim to account for the trade-off between quality (or information) loss and privacy gain, which lies at the core of every anonymization operation. Unlike the bottom-up generalization algorithm, the **Top-Down Specialization (TDS)** algorithm starts with the node representing a maximally generalized dataset. Then, it selects specializations in a top-down manner until no further specialization can be selected without violating the privacy constraint [FWY05,FWY07]. In addition to subtree generalization, TDS employs value suppression. Like Bottom-Up Generalization, the search is accompanied by a Trade-Off Metric, i.e. *IGPL* [FWY05, FWY07]. As opposed to the previously described algorithms, **Mondrian** utilizes multidimensional generalization in combination with local recoding. Introduced by LeFevre et al., the algorithm searches in a top-down manner while splitting the attributes' dimensions into partitions based on the medium value of a chosen attribute. The algorithm terminates when no more splits can be found without violating the privacy constraint, i.e. the partition contains at least k records [LDR06a]. Along with the Mondrian Multidimensional algorithm, the authors present a new quality model, i.e. the Discernibility Metric, which they incorporate into the algorithm.

While all previously described algorithms aim to prevent identity disclosure, fewer algorithms are available that target attribute disclosure. The algorithms **ℓ-diversity Incognito** [MKGV07] and **t-closeness Incognito** [LLV07] have been proposed as an extension to the original Incognito algorithm to prevent attribute disclosure according to the ℓ-diversity and t-closeness privacy model, respectively. Similarly, **InfoGain Mondrian** is an extension of the original Mondrian algorithm utilizing the ℓ-diversity privacy model [LDR06b]. Following the design of their counterparts, the Incognito extensions can produce globally optimal results, while InfoGain Mondrian is a heuristic algorithm which aims to calculate a minimal solution. Wang et al. introduced the **Top-Down Disclosure (TDD)** algorithm, which aims to find a minimally suppressed solution according to the ℓ-diversity privacy model. To this end, the algorithm starts at the completely suppressed dataset and iteratively "discloses" domain values. Here, the search is guided by the IGPL metric and terminates once no further solution exists, which does not violate the privacy constraint [WFP07]. The most well-known membership disclosure algorithms, **SPALM** and **MALM**, have been proposed by Nergiz et al. [NAC07]. Both algorithms aim to prevent membership disclosure, or table linkage, by utilizing the δ-presence privacy model. SPALM exploits the monotonicity property of the δ-presence privacy model, which states that when a dataset is anonymous according to δ-presence, so is its generalized version if it uses full-domain generalization. The algorithm starts with the maximally generalized dataset and creates specializations guided by the Discernibility Metric. The previously described monotonicity property is used to prune the search space. In contrast, MALM is a heuristic algorithm which finds a minimally optimal solution by applying multidimensional generalization.

Both the computer science as well as the statistics community have been paying special attention to data anonymization tools. Tools from the computer science field typically include a mechanism that allows the user to specify a level of privacy risk a priori, which will then be enforced on the data during anonymization automatically. However, often, they focus on specific privacy, quality and transformation models. Examples are the *UTD Anonymization toolbox* [UT 12] and the *Cornell Anonymization Toolkit* [Cor14]. In contrast, solutions from the statistics community do not support the same level of automatism but have to be operated in a more manual manner. Typically, they facilitate an interactive process where transformation methods are specified a priori, but privacy risks are quantified and measured a posteriori; which can be repeated until an acceptable risk level is achieved. Most well-known examples for these types of tools are *sdcMicro* [Tem08] and μ-Argus [HW96]. In contrast, as mentioned above, ARX is able to handle a multitude of different quality, privacy and transformation methods, as well as almost arbitrary combinations thereof. Furthermore, ARX supports the non-interactive approach through automation, but it also facilitates interactive anonymization by means of its graphical user interface in addition to the application programming interface.

3.3 Future Work

Our results from **Ref.A.1** indicated that some general-purpose quality models are best suited for different usage scenarios. One quality model in particular may be best suited for general-purpose use, i.e. Non-Uniform Entropy. However, we also found that high data quality, as measured by these models, does not necessarily correlate with the actual usefulness of data. Therefore, we concluded in **Ref.A.1** that in the future, it should be worthwhile to investigate special-purpose quality models, which are models that have been designed with a specific usage scenario in mind.

Ref.A.2. is a continuation of this work as the focus was on one particular special-purpose quality model. It was specifically designed to capture data quality when the data is used as a training set for creating prediction models to solve a classification task. In this work, we were able to rebut findings from previous research, which stated that due to insufficient data quality, anonymized data was not suitable for prediction tasks. However, one major limitation of this work is that we could only support three types of prediction models: logistic regression, naïve Bayes and random forest. Moreover, while finding a suitable data quality model is challenging, selecting the correct prediction model for a specific task is equally difficult. To overcome these limitations in the future, ARX could be extended with the support of other interesting prediction models, such as C4.5 decision trees or support vectors, as well as the support for benchmark studies. These studies are often performed to experimentally compare different prediction models in order to learn their suitability for specific tasks. Furthermore, besides statistical classification, it could be interesting and beneficial to support other prediction tasks that are particularly interesting in the medical domain, such as regression and time-to-event analysis.

As described previously, a significant requirement of any data anonymization software is the support of multiple usage scenarios. To achieve this, a variety of data transformation, privacy and quality models, as well as algorithms, must be supported. The work in **Ref.A.3** provides a step towards a comprehensive data anonymization tool where different methods and models can be used in combination. However, the tool exhibits some limitations that should be overcome in future work. While ARX is more scalable than other tools in the domain as it can process medium-sized datasets, i.e. datasets containing up to 50 quasi-identifiers and up to a few million rows, this is often not sufficient in the era of big data processing where datasets are multitudes larger than that. There are different ways to tackle this issue. One possibility would be to integrate more scalable data anonymization algorithms. Furthermore, an interesting approach could be distributed data anonymization, where data as well as methods are distributed across multiple machines in order to ease the load on a single machine.

CHAPTER 4

Bibliography

[ABMS15] Alessandro Armando, Michele Bezzi, Nadia Metoui, and Antonino Sabetta. Risk-based privacy-aware information disclosure. *Int J Secur Softw Eng*, 6(2):70–89, 2015.

[AP08] Charu C Aggarwal and S Yu Philip. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-Preserving Data Mining*, pages 11–52. Springer, 2008.

[ARX] ARX - Power Data Anonymization. Accessed: June 21, 2019.

[ASS+11] Jessica S Ancker, Sarah Shih, Mytri P Singh, Andrew Snyder, Alison Edwards, Rainu Kaushal, Hitec Investigators, et al. Root causes underlying challenges to secondary use of data. In *AMIA Annual Symposium Proceedings*, volume 2011, page 57. American Medical Informatics Association, 2011.

[BA05] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *International Conference on Data Engineering*, pages 217–228. IEEE, 2005.

[BJ12] Daniel 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, 2012. Available from SSRN: http://ssrn.com/abstract=2076397. Accessed 5 Jan 2018.

[BK17] Asta Bäck and Janne Keränen. Anonymisointipalvelut. Tarve ja toteutusvaihtoehdot. Liikenne- ja viestintäministeriö, 2017. https://julkaisut.valtioneuvosto.fi/handle/10024/79579.

[BRK$^+$13] Korra Sathya Babu, Nithin Reddy, Nitesh Kumar, Mark Elliot, and Sanjay Kumar Jena. Achieving k-anonymity using improved greedy heuristics for very large relational databases. *Trans. Data Priv.*, 6(1):1–17, 2013.

[BS08] Justin Brickell and Vitaly Shmatikov. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *14th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 70–78. ACM, 2008.

[CCZYM17] Constantinos Costa, Georgios Chatzimilioudis, Demetrios Zeinalipour-Yazti, and Mohamed F Mokbel. Efficient exploration of telco big data with compression and decaying. *33rd International Conference on Data Engineering*, pages 1332–1343, 2017.

[CK06] Lawrence H Cox and Jay J Kim. Effects of rounding on the quality and confidentiality of statistical data. In *Privacy in Statistical Databases: CENEX-SDC Project International Conference, PSD 2006, Rome, Italy, December 13-15, 2006. Proceedings*, pages 48–56. Springer, 2006.

[Com14] European Commission. *Research and Innovation Performance in the EU.* Publications Office of the European Union, Luxembourg, 2014.

[Cor14] Cornell Database Group. Cornell anonymization toolkit. http://sourceforge.net/p/anony-toolkit/, 2014.

[Cox80] Lawrence H Cox. Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, 75(370):377–385, 1980.

[DESG11] George T. Duncan, Mark Elliot, and Juan-José Salazar-González. *Statistical confidentiality: Principles and practice.* Springer, New York, 2011.

[DFT01] Josep Domingo-Ferrer and Vicenc Torra. Disclosure control methods and information loss for microdata. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 91–110, 2001.

[DRW16] Tanvi Desai, Felix Ritchie, and Richard Welpton. Five safes: Designing data access for research. *University of the West of England*, 2016.

[Dut18] Dutch Ministry of Justice and Security. On statistical disclosure control technologies. https://www.wodc.nl/binaries/Cahier%202018-20_2889_Full%20text_tcm28-362210.pdf, 2018.

[Dwo11] Cynthia Dwork. Differential privacy. In *Encyclopedia of Cryptography and Security*, pages 338–340. Springer, 2011.

[DWW99] AG De Waal and LCRJ Willenborg. Information loss through global recoding and local suppression. *Netherlands Official Statistics*, 14:17–20, 1999.

[EEA13] Khaled El Emam and Luk Arbuckle. *Anonymizing health data: Case studies and methods to get you started.* O'Reilly Media, Inc., 1 edition, December 2013.

[EEDI$^+$09] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682, 2009.

[EJAM11] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and B A Malin. A systematic review of re-identification attacks on health data. *PloS One*, 6(12):e28071, 2011.

[Ema13] Khaled El Emam. *Guide to the De-Identification of Personal Health Information.* CRC Press, 2013.

[EMOT16] Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. *The anonymisation decision-making framework.* UKAN, 2016.

[Eur14] European Medicines Agency (EMA). EMA/240810/2013 – European Medicines Agency Policy on Publication of Clinical Data for Medicinal Products for Human Use. 2014.

[Eur15] European Union Agency for Network and Information Security. Privacy and data protection by design. https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design, 2015.

[Eur16] Regulation (EU) 2016/679 of the European Parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Off J European Union*, L119/59, May 2016.

[Eur18] European Medicines Agency. 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. https://www.ema.europa.eu/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency- policy-publication-clinical-data_en-3.pdf, 2018.

[FAWJ18] Frank Fox, Vishal R Aggarwal, Helen Whelton, and Owen Johnson. A data quality framework for process mining of electronic health record data. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 12–21. IEEE, 2018.

[Fin18] Finnish Social Science Data Archive. Data management guidelines: Anonymisation and personal data. https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html, 2018.

[FWCY10] Benjamin C M Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):14, 2010.

[FWFY10] Benjamin C M Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S Yu. *Introduction to privacy-preserving data publishing: Concepts and techniques.* CRC Press, 1 edition, 2010.

[FWJ12] Pui Kuen Fong and Jens H Weber-Jahnke. Privacy preserving decision tree learning using unrealized data sets. *Transactions on Knowledge and Data Engineering*, 24(2):353–364, 2012.

[FWY05] Benjamin CM Fung, Ke Wang, and Philip S Yu. Top-down specialization for information and privacy preservation. In *21st international conference on data engineering (ICDE'05)*, pages 205–216. IEEE, 2005.

[FWY07] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu. Anonymizing classification data for privacy preservation. *Transactions on Knowledge and Data Engineering (IEEE)*, 19(5):711–725, 2007.

[GT09] Aristides Gionis and Tamir Tassa. k-anonymization with minimal loss of information. *Transactions on Knowledge and Data Engineering*, 21(2):206–219, 2009.

[HF11] Leroy Hood and Stephen H Friend. Predictive, personalized, preventive, participatory (p4) cancer medicine. *Nature reviews Clinical oncology*, 8(3):184, 2011.

[HIP] US Health Insurance Portability and Accountability Act of 1996. Public Law, 1996, p. 1349.

[HW96] Anco Hundepool and LCRJ Willenborg. μ-and τ-argus: Software for statistical disclosure control. In *Third International Seminar on Statistical Confidentiality*, 1996.

[IKB09] Ali Inan, Murat Kantarcioglu, and Elisa Bertino. Using anonymized data for classification. In *25th International Conference on Data Engineering*, pages 429–440. IEEE, 2009.

[Iye02] Vijay S Iyengar. Transforming data to satisfy privacy constraints. In *International Conference on Knowledge Discovery and Data Mining*, pages 279–288. ACM, 2002.

[JJB12] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[JKH$^+$17] Chunxiao Jiang, Linling Kuang, Zhu Han, Yong Ren, and Lajos Hanzo. Information credibility modeling in cooperative networks: Equilibrium and mechanism design. *IEEE Journal on Selected Areas in Communications*, 35(2):432–448, 2017.

[KHC$^+$16] Jinkyu Kim, Heonseok Ha, Byung-Gon Chun, Sungroh Yoon, and Sang K Cha. Collaborative analytics for data silos. *32nd International Conference on Data Engineering*, pages 743–754, 2016.

[KHF19] Stephan Kessler, Jens Hoff, and Johann-Christoph Freytag. Sap hana goes private: from privacy research to privacy aware enterprise analytics. *Proceedings of the VLDB Endowment*, 12(12):1998–2009, 2019.

[KL06] Arthur Kennickell and Julia Lane. Measuring the impact of data protection techniques on data utility: Evidence from the survey of consumer finances. In *International conference on privacy in statistical databases*, pages 291–303. Springer, 2006.

[Kor17] Korea Internet & Security Agency. KISA promotes training on identification of personal information.

https://www.kisa.or.kr/notice/press_View.jsp?mode=view&p_No=8&b_No=8&d_No=1570, 2017.

[KPE+12] Florian Kohlmayer, Fabian Prasser, Claudia Eckert, Alfons Kemper, and Klaus A Kuhn. Flash: efficient, stable and optimal k-anonymity. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 708–717. IEEE, 2012.

[LC11] Keng-Pei Lin and Ming-Syan Chen. On the design and analysis of the privacy-preserving SVM classifier. *IEEE Transactions on Knowledge and Data Engineering*, 23(11):1704–1717, 2011.

[LDR05] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *International Conference on Management of Data*, pages 49–60. ACM, 2005.

[LDR06a] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *International Conference on Data Engineering*, page 25. IEEE, 2006.

[LDR06b] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Workload-aware anonymization. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 277–286, 2006.

[Leo12] David Leoni. Non-interactive differential privacy: a survey. In *Proceedings of the First International Workshop on Open Data*, pages 40–52. ACM, 2012.

[LLBW11] Jiuyong Li, Jixue Liu, Muzammil Baig, and Raymond Chi-Wing Wong. Information based data anonymization for classification utility. *Data & Knowledge Engineering*, 70(12):1030–1045, 2011.

[LLV07] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[LLZM10] Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy. Slicing: A new approach for privacy preserving data publishing. *IEEE transactions on knowledge and data engineering*, 24(3):561–574, 2010.

[LMU19] LMU Munich. Conduct your study. https://www.osc.uni-muenchen.de/toolbox/resources_for_researchers/conduct_your_ study, 2019.

[LP04] Sandra Lechner and Winfried Pohlmeier. To blank or not to blank? a comparison of the effects of disclosure limitation methods on nonlinear regression estimates. In *Privacy in Statistical Databases: CASC Project Final Conference, PSD 2004, Barcelona, Spain, June 9-11, 2004. Proceedings*, pages 187–200. Springer, 2004.

[LTZS14] Mark Last, Tamir Tassa, Alexandra Zhmudyak, and Erez Shmueli. Improving accuracy of classification models induced from anonymized datasets. *Information Sciences*, 256:138–161, 2014.

[LZJ$^+$16] Xiang-Yang Li, Chunhong Zhang, Taeho Jung, Jianwei Qian, and Linlin Chen. Graph-based privacy-preserving data publication. *35th International Conference on Computer Communications*, pages 1–9, 2016.

[McK11] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity, 2011. TODO Accessed: June 21, 2019.

[Min15] Ministère des Solidarités et de la Santé. Données de santé: Anonymat et risque de ré-identification. http://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/publications/les-dossiers-de-la-drees/dossiers-solidarite-et-sante/article/donnees-de-sante-anonymat-et-risque-de-re-identification, 2015.

[Min16] Ministry of Science and ICT. A research on de-identification technique for personal identifiable information. https://spri.kr/download/18386, 2016.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *Transactions on Knowledge Discovery from Data*, 1(1):24–35, 2007.

[MKH17] Bernd Malle, Peter Kieseberg, and Andreas Holzinger. Do not disturb? classifier behavior on perturbed datasets. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 155–173. Springer, 2017.

[MW04] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-*

SIGART symposium on Principles of database systems, pages 223–228, 2004.

[NAC07] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 665–676, 2007.

[NC06] M Ercan Nergiz and Chris Clifton. Thoughts on k-anonymization. In *22nd International Conference on Data Engineering*, page 96. IEEE, 2006.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Symposium on Security and Privacy*, pages 111–125. IEEE, 2008.

[Off17] Office of the Australian Information Commissioner. The de-identification decision-making framework. https://www.oaic.gov.au/agencies-and-organisations/guides/de-identification-decision-making-framework, 2017.

[PBE+16] Fabian Prasser, Raffael Bild, Johanna Eicher, Helmut Spengler, Florian Kohlmayer, and Klaus A Kuhn. Lightning: Utility-driven anonymization of high-dimensional data. *Transactions on Data Privacy*, 9:161–185, 2016.

[PE07] Kingsley Purdam and Mark Elliot. A case study of the impact of statistical disclosure control on data quality in the individual uk samples of anonymised records. *Environment and Planning A*, 39(5):1101–1118, 2007.

[Per18] Personal Data Protection Commission of Singapore. Guide to basic data anonymisation techniques. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf, 2018.

[PGW+17] Fabian Prasser, James Gaupp, Zhiyu Wan, Weiyi Xia, Yevgeniy Vorobeychik, Murat Kantarcioglu, Klaus Kuhn, and Brad Malin. An open source tool for game theoretic health data de-identification. In *AMIA Annual Symposium Proceedings*. AMIA, 2017. Accepted for AMIA 2017 Annual Symposium (AMIA 2017).

[PKK16a] Fabian Prasser, Florian Kohlmayer, and Klaus A Kuhn. Efficient and effective pruning strategies for health data de-identification. *BMC Medical Informatics and Decision Making*, 16(1):49, 2016.

[PKK16b] Fabian Prasser, Florian Kohlmayer, and Klaus A Kuhn. The importance of context: Risk-based de-identification of biomedical data. *Methods of Information in Medicine*, 55(4):347–355, 2016.

[PKLK14] Fabian Prasser, Florian Kohlmayer, Ronald Lautenschläger, and Klaus A Kuhn. Arx - a comprehensive tool for anonymizing biomedical data. *Proceedings of the AMIA Annual Symposium*, pages 984–993, 2014.

[PMG$^+$18] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings VLDB Endowment*, 11(10):1071–1083, 2018.

[Pol18] Polish Ministry of Digitalization. Open data - Security standard. https://www.gov.pl/documents/31305/436699/OTWARTE_DANE_%E2%80%93_STANDARDY_BEZPIECZE%C5%83STWA.odt, 2018.

[Res19] Research Data Library Team. https://www.epfl.ch/campus/library/wp-content/uploads/2019/09/RDM_Walkthrough_Guide_20190930.pdf, 2019. École Polytechnique Fédérale de Lausanne (EPFL) Biblioth'eque.

[RL19] Artem Ryasik and Jan Lindquist. Data anonymization in knime. a redfield privacy extension walkthrough. https://www.knime.com/blog/data-anonymization-in-knime-a-redfield-privacy-extension-walkthrough, 2019.

[Sam01] Pierangela Samarati. Protecting respondents' identities in microdata release. *Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[SKH16] Sebastian Stammler, Stefan Katzenbeisser, and Kay Hamacher. Correcting finite sampling issues in entropy l-diversity. *International Conference on Privacy in Statistical Databases*, pages 135–146, 2016.

[Sta18] Standardization Administration of China. GB/T 35273-2017 Information Technology – Personal Information Security Specification, 2018.

[Swe98] Latanya Sweeney. Datafly: A system for providing anonymity in medical data. In *Database Security XI*, pages 356–381. Springer, 1998.

[Swe01] Latanya Sweeney. *Computational disclosure control - A primer on data privacy protection.* PhD thesis, Massachusetts Institute of Technology, 2001.

[Swe02a] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

[Swe02b] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[Tem08] Matthias Templ. Statistical disclosure control for microdata using the R-package sdcMicro. *Transactions on Data Privacy*, 1(2):67–85, 2008.

[Tem17] Matthias Templ. *Statistical Disclosure Control for Microdata: Methods and Applications in R.* Springer, 2017.

[TMF16] TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung. ANONTrain: Praktische Anwendung von Anonymisierungswerkzeugen. http://www.tmf-ev.de/Desktopmodules/Bring2Mind/ DMX/Download.aspx?EntryId=28213&PortalId=0, 2016.

[Tor17] Vicenç Torra. *Data privacy: Foundations, new developments and the big data challenge.* Springer, 2017.

[Uni18] University of Guelph. Clean and prepare your data. https://guides.lib.uoguelph.ca/CleanAndPrepareData/5, 2018.

[Uni19] University of Kassel. Management of research data. https://www.uni-kassel.de/themen/forschungsdatenmanagement/ service-hilfe/faq.html, 2019.

[US 02] 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):53182–53273, 2002.

[UT 12] UT Dallas Data Security and Privacy Lab. UTD Anonymization Toolbox. http://www.cs.utdallas.edu/dspl/cgi-bin/ toolbox/index.php, 2012.

[WDA$^+$16] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[WE18] Isabel Wagner and David Eckhoff. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):1–38, 2018.

[WFP07] Ke Wang, Benjamin CM Fung, and S Yu Philip. Handicapping attacker's confidence: an alternative to k-anonymization. *Knowledge and Information Systems*, 11(3):345–368, 2007.

[WLFW06] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 754–759, 2006.

[Wor] World Health Organization. International statistical classification of diseases and related health problems. Accessed: June 21, 2019.

[WVX+15] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, et al. A game theoretic framework for analyzing re-identification risk. *PloS One*, 10(3):e0120592, 2015.

[WW13] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013.

[WYC04] Ke Wang, Philip S Yu, and Sourav Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 249–256. IEEE, 2004.

[XHD+15] Weiyi Xia, Raymond Heatherly, Xiaofeng Ding, Jiuyong Li, and Bradley A Malin. R-U policy frontiers for health data de-identification. *Journal of the American Medical Informatics Association*, 22(5):1029–1041, October 2015.

[XJC+15] Lei Xu, Chunxiao Jiang, Yan Chen, Yong Ren, and KJ Ray Liu. Privacy or utility in data collection? a contract theoretic approach. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1256–1269, 2015.

[XWP+06] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 785–790, 2006.

[ZB15] Sherali Zeadally and Mohamad Badra. *Privacy in a Digital, Networked World: Technologies, Implications and Solutions.* Springer, 2015.

APPENDIX A

Contributions

Contents

A.1 Full-Core Publications

A.1.1 An Experimental Comparison of Quality Models for Health Data De-Identification

Eicher, Johanna, Klaus A. Kuhn, and Fabian Prasser. "An experimental comparison of quality models for health data de-identification." *MEDINFO 2017: Precision Healthcare through Informatics*. IOS Press, 2017. 704-708.

MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-704

An Experimental Comparison of Quality Models for Health Data De-Identification

Johanna Eicher, Klaus A. Kuhn, Fabian Prasser

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

Abstract

When individual-level health data are shared in biomedical research, the privacy of patients must be protected. This is typically achieved by data de-identification methods, which transform data in such a way that formal privacy requirements are met. In the process, it is important to minimize the loss of information to maintain data quality. Although several models have been proposed for measuring this aspect, it remains unclear which model is best suited for which application. We have therefore performed an extensive experimental comparison. We first implemented several common quality models into the ARX de-identification tool for biomedical data. We then used each model to de-identify a patient discharge dataset covering almost 4 million cases and outputs were analyzed to measure the impact of different quality models on real-world applications. Our results show that different models are best suited for specific applications, but that one model (Non-Uniform Entropy) is particularly well suited for general-purpose use.

Keywords:

Privacy, Personally identifiable information, Data anonymization

Introduction

The collaborative collection and sharing of sensitive individual-level data has become an important aspect of modern biomedical research. The secondary use of health data for research purposes is a typical example [1]. To protect privacy in such scenarios, a broad spectrum of safeguards must be implemented, including data use agreements and fine-grained access control [2]. On the data-level, anonymization is a central safeguard. There are various ways and definitions. One important aspect is data *de-identification*, which focuses on protecting data from re-identification. For this purpose, datasets are transformed in such a way that it becomes very difficult to link their records to identified individuals without investing a disproportionate amount of time and effort [3].

There are *rule-based* and *computational* approaches to the de-identification of health data. The Safe Harbor method of the US Health Insurance Portability and Accountability Act (HIPAA) [4] is a typical example for the former type. It specifies 18 transformation rules which describe the removal or alteration of attribute values that are associated with a high risk of re-identification (e.g. names and dates). In other jurisdictions, e.g. in Germany [3], regulations are less interpretable and computational methods to data de-identification are thus more important. Here, data is transformed (semi-) automatically to ensure that privacy risks are minimized.

The transformation of data inevitably leads to loss of information. Therefore, a balance has to be sought between an increase in privacy protection on one side and a decrease in data quality on the other. *Privacy models* and *quality models* are used to quantify the two aspects. The contradiction between the two conflicting optimization goals is typically resolved by specifying a risk threshold for the privacy model. This reduces the de-identification process to a simpler optimization problem, in which the objective is to make sure that risk thresholds are met while data quality is maximized [2].

Objective

Measuring data quality is a non-trivial issue as the nature of usefulness of data often depends on the use case [5]. As it is typically unknown in advance how the data will be analyzed, models are needed, which quantify data quality for general-purpose use. Fung et al. proposed to measure the *similarity* (which can be defined in multiple ways) between the original and the de-identified data [6]. Domingo-Ferrer et al. noted that a quality model should capture the amount of information loss for a reasonable range of data uses [7]. They introduced two characteristics, *analytically valid* and *analytically interesting*, which need to be present for a dataset to have *little* loss of information. In this context, analytical validity requires the preservation of certain statistical characteristics, while data is said to be analytically interesting if some useful attributes for further analyses remain intact [7].

A wide variety of *general-purpose quality models*, which aim to distinguish valid or interesting data from invalid or uninteresting data, have been proposed and used in scientific papers. Typically, these models define a decrease in data quality, as well as an increase in information loss, which can be quantified [5]. The notion of using information loss as an indicator for data quality is also prevalent in official statistics, namely the so-called *score*, which measures the trade-off between quality (information loss) and privacy (disclosure risk of the released data) [8]. Even though various papers have compared data de-identification algorithms, a systematic evaluation of quality models has not been conducted yet. Consequently, a guideline for selecting appropriate models for specific scenarios is missing. Potential application scenarios for de-identified data include the privacy-preserving sharing of data from research registries or health databases. De-identified data extracts may also be used to provide partners with an overview of data, which is potentially available for sharing in a fine-grained form. Finally, de-identified data can also directly be used for advanced analytics and observational research, e.g. for building predictive models.

As a first step towards the development of a guideline, we have implemented and evaluated several general-purpose quality models with the intention of answering the following questions:

1. How do common models for measuring data quality influence the way in which datasets are transformed?
2. If different models are used, how are the obtained results related to each other?

3. How well is de-identified data, obtained by using different quality models, suited for real-world applications?

Methods

Background

In data de-identification, the general attack vector assumed is *linkage* of a sensitive dataset with an identified dataset (or similar background knowledge about individuals). *Identity disclosure (or re-identification)* means that an individual is successfully linked to a specific data record [9]. This is a very important type of privacy breach, as it has legal consequences for data owners according to many laws and regulations worldwide. As a first step towards data de-identification, *directly identifying* information (such as names) must be removed [10]. The remaining attributes, which may be used for linkage, are termed *quasi-identifiers* (or indirect identifiers, or keys). Such attributes are not directly identifying, but they may be used in combination for linkage. It is further assumed that they cannot simply be removed from a dataset, as they may be required for analyses and that corresponding information is likely to be available to an attacker.

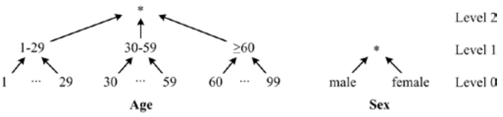
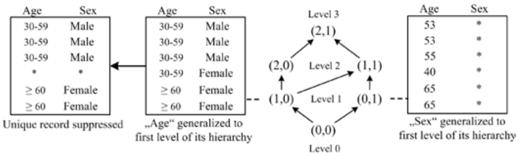


Figure 1 – Generalization hierarchies for "age" and "sex"

When data is de-identified, values of quasi-identifiers are transformed to ensure that the data fulfills privacy requirements. This can be performed with user-defined *generalization hierarchies* [11]. Examples are shown in Figure 1. Here, values of the attribute "age" are transformed into intervals, with decreasing precision on increasing *levels* of generalization. Values of the attribute "sex" can only be suppressed. Generalization hierarchies are well suited for categorical attributes, but they can also be constructed for continuous attributes through categorization.



Age	Sex
30-59	Male
30-59	Male
30-59	Male
*	*
≥ 60	Female
≥ 60	Female

Unique record suppressed

Age	Sex
30-59	Male
30-59	Male
30-59	Male
30-59	Female
≥ 60	Female
≥ 60	Female

„Age" generalized to first level of its hierarchy

Age	Sex
53	*
53	*
55	*
40	*
65	*
65	*

„Sex" generalized to first level of its hierarchy

Figure 2 – Example showing different transformations represented as a generalization lattice

In order to transform the data, *globally-optimal full-domain anonymization* algorithms have been recommended. Such algorithms construct a search space in a structure called *generalization lattice*. An example for the lattice constructed from the hierarchies from Figure 1 is shown in Figure 2. The graph displays each node that represents a single transformation, which defines generalization levels for all quasi-identifiers. An arrow denotes that a transformation is a direct *generalization* of a more *specialized* transformation, which means that it increments exactly one generalization level as defined by its predecessor. The original dataset (0, 0) is at the bottom, whereas the transformation with maximal generalization (2, 1) is at the top. The search space is then traversed to find a transformation, which results in output data that fulfills all privacy requirements, and at the same time provides optimal data quality.

Protection against re-identification is often implemented with the *k-anonymity* privacy model [9]. A dataset is k-anonymous if, regarding the quasi-identifiers, each record cannot be distinguished from at least $k - 1$ other records. This property can be used to define *equivalence classes* of indistinguishable records [12]. The basic idea is that an attacker will only be able to associate any individual with at least k records, which reduces the probability of correct linkage to not more than $\frac{1}{k}$. The left-most output dataset in Figure 2 fulfills 2-anonymity, which means the maximal re-identification risk of any record is 50%.

In order to create de-identified datasets of high quality, attribute generalization can be combined with the suppression of data records. This means that records from equivalence classes, which violate the privacy model (i.e. *outliers*), are automatically replaced with semantic-free placeholders. Because of record suppression, less generalization is required to ensure that the remaining records fulfill the privacy model, which increases the quality of de-identified datasets [13]. In the left-most dataset from Figure 2, one record has been suppressed in the output of applying the transformation (1, 0).

The aforementioned quality models are used to rank the different privacy-preserving output datasets and to select a transformation that maximizes data quality. The *Loss* model, for example, measures data granularity by analyzing the extent to which the domain of an attribute is covered by the transformed values [14]. For the datasets shown in Figure 2, the reduction in data granularity is 30%, 16% and 50%, from left to right.

Quality Models

We have implemented the following quality models (QMs) into ARX, which is an open-source de-identification tool for biomedical data [13]:

- *Average Equivalence Class Size (AECS)* is a row-oriented model, which measures the average size of equivalence classes of indistinguishable records [15].
- *Discernibility* is a row-oriented model, which measures the size of equivalence classes combined with a penalty for suppressed records [16].
- *Precision* is a cell-oriented model, which quantifies data quality by reporting the amount of distortion of attribute values. Distortion is measured as the generalization level of an attribute value relative to the height of the attribute's generalization hierarchy [17].
- *Loss* is a cell-oriented model, which measures the granularity of data by determining the fraction of an attribute's domain that is covered by the transformed values [14].
- *Ambiguity* is a row-oriented model, which measures the degree of uncertainty in the resulting data [18].
- *Kullback-Leibler (K.-L.) Divergence* is a row-oriented model, which measures differences in the distributions of equivalence class sizes [19].
- *Non-Uniform (N.-U.) Entropy* is a column-oriented model, which measures differences in the distributions of attribute values induced by data transformations [20]. It is based on the concept of mutual information, which quantifies the amount of information that can be obtained about one variable by observing the other.

Dataset

Table 1– Description of the patient discharge dataset

Attribute	Type	Description
Hospital ID	Spatial	A 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	Sensitive	Total charges for the stay

We used each quality model to de-identify a publicly available patient discharge dataset [21]. The dataset contains 3,985,166 records and 10 attributes (Table 1). In our experiments, we have defined all spatial, demographic and temporal attributes as quasi-identifiers. As we will describe later, we used the remaining sensitive attribute (charge) for determining the usefulness of output data.

Privacy Protection

We de-identified the dataset with attribute generalization and record suppression to produce output datasets, which fulfill the *k-anonymity* privacy model. We chose *k =5*, which is a typical parameter in the biomedical domain that specifies a re-identification risk of not more than 20% for each record [2].

Experimental Design

We addressed the first research question, i.e. how quality models influence the way in which datasets are transformed, analyzing how much generalization and record suppression had to be used in the de-identification process to achieve optimal data quality. The former is expressed as a *generalization degree* for each attribute, which is defined as the relative generalization level to which it was transformed. The latter is expressed as the number of *removed records*.

To answer the second question, i.e. how the results obtained with different models are related to each other, we used each model to assess the quality of the optimal solutions obtained with all other models. To make the different quantifications of quality comparable to each other, we normalized them: a value of 0% represents the original data and a value of 100% represents a dataset where all information has been removed.

Finally, to answer the third question, i.e. how well the de-identified data is suited for real-world applications, we analyzed the impact of the different methods of data transformation on typical use cases. Moreover, we employed statistical classification, which is a common application scenario for individual-level data [22]. The aim was to predict the values of a selected *class attribute* from a set of *feature attributes*. This is implemented with supervised learning, where a model is created from a training set. We used the discharge dataset to build logistic regression models [23], which were able to predict the height of the bill for hospital stays, i.e. whether the *charge* for a stay was below $10,000, between $10,000 and $50,000, or greater than $50,000.

To be able to quantify the analytical validity of the de-identified data, we created classifiers, which could be evaluated using the original input data; although they have only been trained with de-identified output data. For this purpose, we implemented the approach presented in [22] into ARX. For evaluating different predictors, we used 10-fold cross-validation. We normalized all resulting prediction accuracies into the range [0, 1], where 0% represents the accuracy of the trivial *ZeroR* method, which simply always returns the most frequent class value from the original dataset [23], and 100% represents the accuracy of a logistic regression model trained with the original, unmodified input dataset.

Results

How do common models for measuring data quality influence the way in which datasets are transformed?

Table 2 – Generalization degrees and removed records (RR).

QM	Generalization degrees	RR
AECS	5·100%, 4·0%	25%
Disc.	6·100%, 1·57%, 2·0%	0%
Precision	1·100%, 1·60%, 1·33%, 6·0%	7%
Loss	1·67%, 1·60%, 1·57%, 1·33%, 5·0%	5%
Ambiguity	4·100%, 1·67%, 1·57%, 1·50%, 1·17%, 1·0%	0%
K.-L. Div.	2·100%, 1·17%, 6·0%	21%
N.-U. Ent.	4·100%, 1·71%, 1·50%, 3·0%	10%

Table 2 shows the generalization degrees, and the number of removed records for the outputs obtained by de-identifying the discharge dataset with each quality model.

It can be seen that the fraction of removed records varied between 0% and 25%. With each quality model, at least one attribute was preserved as-is, while just the result obtained with the Loss model did not contain at least one completely generalized attribute. The dataset was transformed with high degrees of generalization when AECS, Discernibility, Ambiguity or N.-U were used. Entropy was used to quantify the loss of information. Just little generalization was used when quality was measured with Precision, Loss and K.-L. Divergence. No records were removed when Discernibility or Ambiguity were used, while a large proportion of the records was removed when quality was measured with AECS and K.-L. Divergence. With the models Precision, Loss and N.-U. Entropy, the dataset was transformed with a balanced combination of both attribute generalization and record suppression.

Table 3 – Relative information loss in percent

	QM used for de-identification						
QM used for evaluation	AECS	Discernibility	Precision	Loss	Ambiguity	K.-L. Divergence	N.-U. Entropy
AECS	**0.0004**	0.0096	0.0019	0.0024	0.0930	0.0011	0.0011
Discernibility	24.5553	**0.0390**	6.9322	5.4202	0.0424	20.5476	9.6655

Precision	66.4688	72.8116	**26.9124**	28.0155	62.9714	39.6713	62.0003
Loss	66.4688	68.4609	18.4723	**9.3869**	49.3907	38.2668	56.4513
Ambiguity	24.5547	0.0280	6.9172	5.3887	**0.0070**	20.5426	9.6612
K.-L. Divergence	83.8570	96.7961	45.6947	49.2434	72.6321	**27.8739**	55.1117
N.-U. Entropy	62.4520	65.9717	48.9954	51.8416	81.0103	44.4787	**39.4218**

How are the datasets obtained with different models related to each other?

Table 3 shows how the different models assessed the quality of output data obtained using the other models. Each column represents the result of de-identifying the data with a single model as indicated. In each row, a model was used to assess the quality of the output obtained with the other models. Consequently, the highlighted values on the diagonal represent the optimum for each model.

It can be seen that, in terms of AECS, all results had comparable data quality. However, information loss was considered very low in general. When using Discernibility and Ambiguity, results obtained with the other models were determined to be much worse and quality values differed by orders of magnitude. When using Precision and Loss, quality values were within a reasonable range, considering the transformations, which were applied to the data (Table 2). However, both models, as well as K.-L. Divergence, measured big differences between the different solutions. When using N.-U. Entropy, the quality of the results from the different models was placed in a reasonable range, and different solutions were considered to be of rather comparable quality.

How well is de-identified data obtained with different quality models suited for real-world applications?

Before building the prediction models, we performed a feature selection process. The results showed that neither the age of a patient nor the length of a stay was predictive for the prices charged by the hospitals. Therefore, we built classifiers, which predicted the charge from the spatial features *hospital-ID* and *county of residence*, the demographic parameters *sex*, *ethnicity* and *race,* as well as temporal information in form of the *admission quarter*.

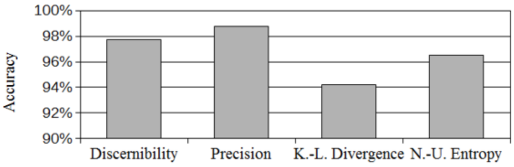


Figure 3– Relative accuracies of logistic regression models

The results obtained by training models with the output of using AECS, Loss and Ambiguity performed not very well with prediction accuracies below 30%. Figure 3 shows the results obtained using the remaining quality models. It can be seen that when using Discernibility, Precision, K.-L. Divergence and N.-U. Entropy, the de-identification process had just negligible effects on the performance of the prediction models; we measured relative accuracies between 94% and 98%. This means that the models performed almost as well as models trained with unmodified input data.

Discussion

Our experiments indicate that different models are suited best for different application scenarios. When using the AECS model, datasets were de-identified with a high degree of generalization and a high degree of record suppression. Moreover, predictive models created from the output obtained with AECS exhibited sub-optimal performance. This shows that the model is not suited well for real-world applications in biomedicine. When using the models Discernibility or Ambiguity, datasets were de-identified with attribute generalization only. This means that the models are suited well for de-identifying small datasets, e.g. from rare disease networks or data collections from sparsely populated regions, where statistical power may otherwise be reduced disproportionally. Using the models Precision, Loss or N.-U. Entropy resulted in a balanced application of both attribute generalization and record suppression. This means that the output data is well suited for providing potential data sharing partners with an overview of available data, as instance-level and schema-level information is preserved. Finally, the models Discernibility, Precision, K.-L. Divergence and N.-U. Entropy are suited well for de-identifying data that is to be used for predictive modeling. The latter two models are based on stringent information theoretic foundations, and it is thus not surprising that output obtained with them is suited well for machine learning purposes. In contrast, we did not expect to obtain such good results when using Discernibility and Precision, as both are rather simple in nature.

Our results have also shown that the utility or usefulness of data does not necessarily correlate with the degree of quality measured by general-purpose models. In future work, we also plan to investigate *special-purpose quality models*, which are models that have been designed with specific usage scenarios in mind.

The application scenario investigated in this article, statistical classification, is a well-known example of a specific application scenario. While our results have shown that data de-identified with general-purpose quality models can be suited well for this context, specialized quality models also have been proposed. They minimize the loss of information for features, which are most discriminating for a specified class attribute [6]. This has been shown to optimize output data for classification purposes [16].

Another application scenario is the de-identification of diagnosis codes for use in association studies between phenotypic and genotypic data [24]. The transactional characteristics of such data require that irrelevant inter-attribute relationships are removed, which can be achieved with specialized de-identification algorithms that also require specific data quality models. In future work, we plan to investigate such models, e.g. utility constraints [24] as well.

Conclusion

Non-Uniform Entropy is a quality model, which has frequently been recommended for de-identifying health data, e.g. by Emam et al. [25]. Based on the results of our experiments, we can confirm that the model provides the best results for general-purpose usage. With this model, de-identified data contained instance-level and schema-level information. Moreover, statistical power was reduced by only 10%. Finally, by using de-identified data with optimal quality according to this model as a training set, we were able to build a statistical classifier with good prediction accuracy.

References

[1] J. Christoph, L. Griebel and I. Leb, Secure secondary use of clinical data with cloud-based NLP services, *Methods of Information in Medicine* **54** (2015), 276–282.

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

[3] Federal Data Protection Act, version promulgated on 14 January 2003 (Federal Law Gazette I p. 66), as most recently amended by Article 1 of the Act of 14 August 2009 (Federal Law Gazette I p. 2814).

[4] U.S. health insurance portability and accountability act of 1996, Public Law, 1996, p. 1349

[5] B.C.M. Fung, K. Wang, A.W.-C. Fu and P.S. Yu, *Introduction to privacy-preserving data publishing: Concepts and techniques,* CRC Press, 1st ed., 2010, ISBN 9781420091489.

[6] B.C.M. Fung, K. Wang, R. Chen and P.S. Yu, Privacy-preserving data publishing: A survey of recent developments, *ACM Computing Surveys* **42** (2010), 14.

[7] J. Domingo-Ferrer, Sánchez and S. Hajian, Database privacy, S. Zeadally and M. Badra, editors, *Privacy in a digital, networked world: Technologies, implications and solutions*, Springer, 2015.

[8] J. Nin, J. Herranz and V. Torra, Towards a more realistic disclosure risk assessment, *Privacy in Statistical Databases*, Springer, 2008, vol. 5262 of Lecture Notes in Computer Science, 152–165.

[9] L. Sweeney, *Computational disclosure control - A primer on data privacy protection*, Ph.D. thesis, Massachusetts Institute of Technology, 2001.

[10] F. Prasser, R. Bild, J. Eicher, H. Spengler, F. Kohlmayer and K.A. Kuhn, Lightning: Utility-driven anonymization of high-dimensional data, *Transactions on Data Privacy* **9** (2016), 161–185.

[11] P. Samarati, Protecting respondents' identities in microdata release*,* *Transactions on Knowledge and Data Engineering* **13** (2001), 1010–1027.

[12] P. Samarati and L. Sweeney, Generalizing data to provide anonymity when disclosing information, in *Symposium on Principles of Database Systems* (ACM) (1998).

[13] F. Prasser, F. Kohlmayer, R. Lautenschläger and K.A. Kuhn, ARX – A comprehensive tool for anonymizing biomedical data, *AMIA Annual Symposium Proceedings, AMIA*, 2014, 984–993

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

[15] K. LeFevre, D.J. DeWitt and R. Ramakrishnan, Mondrian multidimensional k-anonymity, International Conference on Data Engineering, IEEE, 2006, 25.

[16] R.J. Bayardo and R. Agrawal, Data privacy through optimal k-anonymization*, International Conference on Data Engineering*, IEEE, 2005, 217–228.

[17] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10** (2002), 571–588.

[18] M.E. Nergiz and C. Clifton, Thoughts on k-anonymization, *International Conference on Data Engineering*, IEEE, 2006, 96.

[19] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, l-diversity: Privacy beyond k-anonymity, *Transactions on Knowledge Discovery from Data* **1** (2007), 24–35.

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

[21] D. Sanchez, S. Martinez and J. Domingo-Ferrer, Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata" (Supplementary Materials: http://arxiv.org/abs/1511.05957, Archive: http://crises-deim.urv.cat/opendata/SPD_Science.zip), *Science* **6279** (2016), 1274–1274.

[22] A. Inan, M. Kantarcioglu and E. Bertino, Using anonymized data for classification, *Proceedings - International Conference on Data Engineering* (2009), 429–440.

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

[24] G. Loukides, A. Gkoulalas-Divanis and B. Malin, Anonymization of electronic medical records for validating genome-wide association studies., *Proceedings of the National Academy of Sciences of the United States of America* **107** (2010), 7898–7903.

[25] K. El Emam and L. Arbuckle, *Anonymizing health data: Case studies and methods to get you started*, O'Reilly Media, Inc., 1st ed., 2013, ISBN 978-1-449-36307-9.

Address for correspondence

Fabian Prasser, Chair of Medical Informatics, Institute of Medical Statistics and Epidemiology, University Hospital rechts der Isar, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany; E-mail: fabian.prasser@tum.de.

A.1.2 A Comprehensive Tool for Creating and Evaluating Privacy-Preserving Biomedical Prediction Models

Eicher, Johanna, Raffael Bild, Helmut Spengler, Klaus A. Kuhn, and Fabian Prasser. "A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models." *BMC medical informatics and decision making* 20 (2020): 1-14.

Eicher *et al. BMC Medical Informatics and Decision Making* (2020) 20:29
https://doi.org/10.1186/s12911-020-1041-3

BMC Medical Informatics and Decision Making

SOFTWARE Open Access

A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models

Johanna Eicher[1*], Raffael Bild[1], Helmut Spengler[1], Klaus A. Kuhn[1] and Fabian Prasser[2,3]

Abstract

Background: Modern data driven medical research promises to provide new insights into the development and course of disease and to enable novel methods of clinical decision support. To realize this, machine learning models can be trained to make predictions from clinical, paraclinical and biomolecular data. In this process, privacy protection and regulatory requirements need careful consideration, as the resulting models may leak sensitive personal information. To counter this threat, a wide range of methods for integrating machine learning with formal methods of privacy protection have been proposed. However, there is a significant lack of practical tools to create and evaluate such privacy-preserving models. In this software article, we report on our ongoing efforts to bridge this gap.

Results: We have extended the well-known ARX anonymization tool for biomedical data with machine learning techniques to support the creation of privacy-preserving prediction models. Our methods are particularly well suited for applications in biomedicine, as they preserve the truthfulness of data (e.g. no noise is added) and they are intuitive and relatively easy to explain to non-experts. Moreover, our implementation is highly versatile, as it supports binomial and multinomial target variables, different types of prediction models and a wide range of privacy protection techniques. All methods have been integrated into a sound framework that supports the creation, evaluation and refinement of models through intuitive graphical user interfaces. To demonstrate the broad applicability of our solution, we present three case studies in which we created and evaluated different types of privacy-preserving prediction models for breast cancer diagnosis, diagnosis of acute inflammation of the urinary system and prediction of the contraceptive method used by women. In this process, we also used a wide range of different privacy models (k-anonymity, differential privacy and a game-theoretic approach) as well as different data transformation techniques.

Conclusions: With the tool presented in this article, accurate prediction models can be created that preserve the privacy of individuals represented in the training set in a variety of threat scenarios. Our implementation is available as open source software.

Keywords: Biomedical data, Prediction models, Machine learning, Classification, Privacy protection, Data anonymization

Background

The digitalization of healthcare promises to enable personalized and predictive medicine [1]. Based on digital data that characterize patients and probands at comprehensive depth and breadth [2], machine learning models can be created that are able to detect unknown relationships between biomedical parameters and enable decision support systems by using the knowledge about such relationships to infer or predict parameters (henceforth called *target variables*), e.g. diagnoses or outcomes [3]. However, in such data-driven environments, it is becoming increasingly challenging to protect the data used for creating such models from privacy breaches [4]. Data privacy involves ethical, legal and societal aspects [5] and different layers of protection mechanisms must therefore be implemented [6, 7].

*Correspondence: johanna.eicher@tum.de
[1] School of Medicine, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany
Full list of author information is available at the end of the article

On the technical level, current efforts in the area of machine learning for health data put a significant focus on distributed learning which overcomes the need to share data across institutional boundaries to create the large datasets needed for training purposes [8, 9]. Cryptographic secure multiparty computation approaches are an important technique in this context [10]. Although this solves some of the privacy issues, it is important to realize that privacy protection must be addressed on multiple levels, including the output data level where it must be ensured that the resulting prediction models cannot be used to extract personal information [11]. Prediction models, which learn from anonymized data are a common solution to this problem. The core concept behind data anonymization is to transform data in such a manner that privacy risks are reduced while the reduction of risks is balanced against a reduction of data utility [12, 13]. Several high-profile re-identification attacks have shown that simply removing all directly identifying attributes (e.g. names and addresses) is not sufficient for this purpose [14, 15]. Laws and regulations, e.g. the Privacy Rule of the U.S. Health Insurance Portability and Accountability Act (HIPAA) [16] or the European General Data Protection Regulation [17], define different approaches to address this issue.

In recent years, several easy-to-use tools have been developed that make methods of data anonymization available to a broad range of users. At the same time, various methods for addressing output data privacy in machine learning have been proposed by the research community, but robust implementations that can be applied in practice are lacking. In this article, we report on our ongoing efforts to bring both worlds together by integrating machine learning techniques into a well-known data anonymization tool. In prior work, we have laid the groundwork for the results presented in this article by (1) implementing a method into the tool that ensures that anonymized output data is suitable as training data for creating prediction models, and (2) integrating logistic regression models into the tool in such a way that they can be used to assess the performance of models created from anonymized data [18]. In this software article, we present a wide range of enhancements that significantly broaden the applicability of the approach. In detail, we

1. added a method to make anonymized output data suitable for the training of multiple models that can predict different target variables,
2. implemented additional types of prediction models to enable assessing the performance of different types of privacy-preserving machine learning techniques,
3. integrated the approach with further anonymization methods, including differential privacy, which is a state-of-the-art approach offering strong privacy protection,
4. implemented a wide range of additional metrics and visualizations for assessing the impact of privacy protection on prediction performance,
5. added support for further data transformation techniques, such as data aggregation.

The resulting tool is highly versatile, as it supports binomial and multinomial target variables, different types of prediction models and a wide range of methods of privacy protection. Moreover, all techniques have been integrated into a sound framework that supports the creation, evaluation and refinement of models through intuitive graphical user interfaces. We demonstrate the broad applicability of our approach by creating different types of privacy-preserving models for breast cancer diagnosis, diagnosis of acute inflammation of the urinary system and prediction of the contraceptive method used by women using different anonymization and prediction techniques. The results show that accurate prediction models can be created that preserve privacy in a variety of threat scenarios. Our implementation is available as open source software.

Implementation

The software described in this article has been developed by extending ARX, an open source anonymization tool which has specifically been designed for applications to biomedical data [19]. In this section, we will focus on the two most important functionalities implemented, which are (1) methods for the automated creation of privacy-preserving prediction models and (2) methods for evaluating and fine-tuning the resulting models. In the individual sections, we will describe how we addressed particularly complex challenges.

Methods for creating privacy-preserving prediction models

In predictive modeling, the goal is to predict the value of a predefined *target variable* from a given set of values of *feature variables* as accurately as possible. Typical application scenarios in medicine include knowledge discovery and decision support.

Our tool implements the common *supervised learning* approach, where a model is created from a *training set*. It focusses on *classification* tasks where target variables are categorical and values of the target variable are called *classes* [20]. To create privacy-preserving prediction models, our tool implements supervised learning from anonymized data. To maximize the performance of the resulting models it utilizes the optimization procedures provided by ARX to produce anonymized output data that is suited for this purpose.

At its core, ARX utilizes user-defined generalization *hierarchies* to transform data. A simple example is shown in Fig. 1. As can be seen, generalization hierarchies store the original attributes' values in the leaf nodes while inner nodes contain generalized representations of the values from the leaf nodes of the according subtree. When a hierarchy is used to transform the values of an attribute, all values are replaced by the corresponding inner nodes on a given *level* of the hierarchy. In the example, values of the attribute "age" are transformed into age groups by replacing them with the corresponding generalized values on level 2 of the hierarchy, while values of the attribute "sex" are left as-is (which corresponds to "transforming" them to level 0 of the hierarchy). In an abstract sense, the anonymization process implemented by ARX basically produces all possible output datasets by applying all possible combinations of generalizations to the input dataset. For each possible output, two parameters are measured: (1) privacy protection, and (2) data utility. After this process, ARX returns the transformed dataset that satisfies pre-defined privacy protection levels and which is most useful. In practice, ARX implements a wide range of pruning strategies and optimization techniques to avoid needing to analyze all possible output datasets (see, e.g. [19, 21]). Moreover, ARX supports further transformation techniques which are implemented by extending the basic anonymization process outlined in this paragraph. Furthermore, privacy protection as well as data utility can be measured using different models. We will briefly introduce the most important methods used in this article in the remainder of this section.

Privacy models

In ARX, privacy models are used to specify and quantify levels of protection. The methods for creating privacy-preserving prediction models presented in this article are compatible with all privacy models currently implemented by ARX (an overview is provided on the project website [22]). In this paper, we will use the following models to showcase our solution: (1) *k-anonymity*, which protects records from re-identification by requiring that each transformed record is indistinguishable from at least $k - 1$ other records regarding attributes that could be used in linkage attacks [15], (2) *differential privacy* which guarantees that the output of the anonymization procedure is basically independent of the contribution of individual records to the dataset, which protects output data from a wide range of risks [23, 24], and (3) a *game-theoretic model* which employs an economic perspective on data re-identification attacks and assumes that adversaries will only attempt re-identification in case there is a tangible economic benefit [25, 26].

Utility models

ARX supports a wide range of models for quantifying (and hence optimizing) the utility of output data. To optimize output towards suitability as a training set for prediction models, we have implemented the method by Iyengar [27]. The basic idea is to distinguish between the removal of *structure* and the removal of *noise* by measuring the heterogeneity of values of class attributes in groups of records that are indistinguishable regarding the specified feature variables. For instance, if the age of individuals and the occurrence of a certain disease exhibits a strong correlation, the relationship between these two attributes is most likely best captured by adequate age groups instead of more granular data. In prior work, we have already described a basic implementation of the approach [18]. However, the implementation had several important limitations, which resulted from the compressed internal data representation used by ARX [19]: (1) it only supported one class variable, (2) it required that class variables were addressed by a privacy model, and (3) it required that no transformations were applied to target variables. To

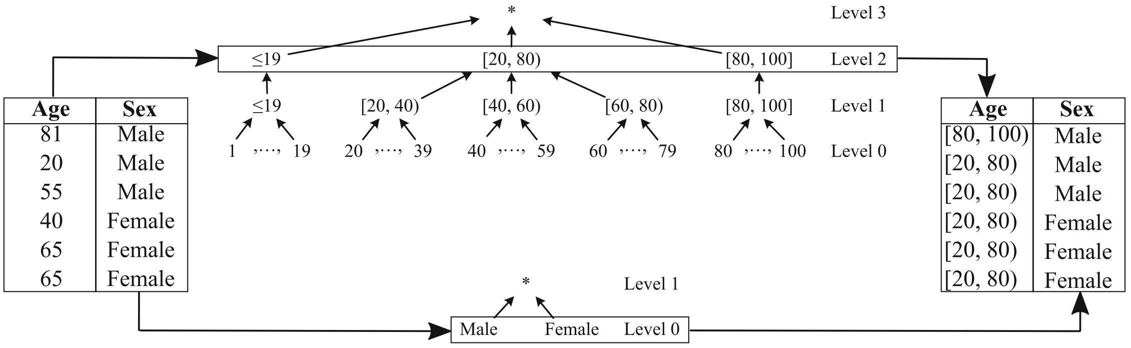


Fig. 1 Example of attribute transformation based on generalization hierarchies. Values of the attributes "age" and "sex" are transformed using level 2 and level 0, respectively, of their associated hierarchies

overcome these limitations we had to rewrite major parts of the internals of the software and the resulting utility model is now the most complex model supported. Finally, we also had to develop and implement a specialized *score function* with proven mathematical properties to support differential privacy [24].

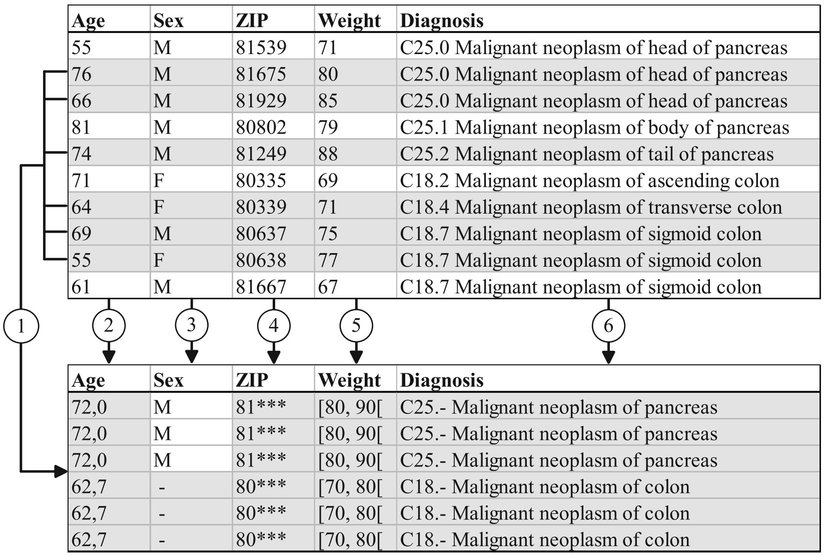
Transformation models

Based on the generic mechanism described above, ARX provides support for a wide range of transformation techniques. Different methods for transforming data can also be used in combination. Typically, this is done to preserve as much output data utility as possible and to preserve important schematic properties of data, such as the data types of variables. Figure 2 shows an example of the different methods supported: (1) *Random sampling* is a common method to reduce the certainty of attackers about the correctness of re-identifications. It is also a major building block of differential privacy in ARX [24]. (2) *Aggregation* is a method where sets of numeric attribute values are transformed into a common aggregated value. (3) *Suppression* means that values are simply removed from a dataset, which may be applied on the cell-, record- or attribute-level. (4) *Masking* is a method where individual characters are removed. (5) *Categorization* means that continuous variables are mapped to categories. (6) *Generalization* is a method where attribute values are replaced by less specific values based on user-defined generalization hierarchies or classifications, such as the International Classification of Diseases [28].

In the output dataset shown in Fig. 2, the risk of a record being re-identified correctly is not higher than 33.3% (3-anonymity). In addition, the anonymization procedure fulfills (ϵ, δ)-differential privacy with $\epsilon \approx 0.92$ and $\delta \approx 0.22$, under the assumption that all changes other than sampling have been implemented using a data-independent transformation method [24]. While support for the transformations utilized in the example is provided out-of-the-box by ARX, implementing evaluation methods for prediction models trained on this data needs careful attention, as we will describe in the next section.

Classification models

To enable users to assess the performance of different types of prediction techniques, we implemented a generic interface to prediction models and integrated three methods as is shown in Fig. 3: (1) *Logistic regression*, where the relationship between the feature variables and the target variable is expressed as a linear model which is transformed using a logarithmic function [20]. Since support for this model was already established in previous work, we only had to make minor adjustments to integrate it with the new interface. (2) *Naïve Bayes* [29], which makes strong (hence naïve) assumptions about the independence of the distributions of the feature variables based on Bayes' theorem. The only dependency is assumed to exist between the target variable and each of the feature variables. Predictions are made by simply calculating the posterior probabilities of each of the classes using the prior probability of the feature vector. (3) *Random forest* [30],



Age	Sex	ZIP	Weight	Diagnosis
55	M	81539	71	C25.0 Malignant neoplasm of head of pancreas
76	M	81675	80	C25.0 Malignant neoplasm of head of pancreas
66	M	81929	85	C25.0 Malignant neoplasm of head of pancreas
81	M	80802	79	C25.1 Malignant neoplasm of body of pancreas
74	M	81249	88	C25.2 Malignant neoplasm of tail of pancreas
71	F	80335	69	C18.2 Malignant neoplasm of ascending colon
64	F	80339	71	C18.4 Malignant neoplasm of transverse colon
69	M	80637	75	C18.7 Malignant neoplasm of sigmoid colon
55	F	80638	77	C18.7 Malignant neoplasm of sigmoid colon
61	M	81667	67	C18.7 Malignant neoplasm of sigmoid colon

Age	Sex	ZIP	Weight	Diagnosis
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon

Fig. 2 Example of different transformation schemes used in data anonymization. 1: Sampling, 2: Aggregation, 3: Suppression, 4: Masking, 5: Categorization, 6: Generalization

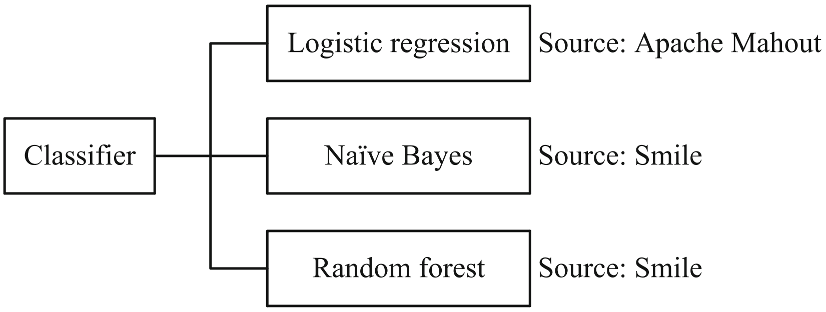


Fig. 3 Classification models supported by the current implementation. A generic wrapper has been developed to encapsulate the implementation specifics of different machine learning libraries

which belongs to the class of *ensemble learning methods*. This means that the predictions of multiple models are combined into a single prediction. The individual models are decision trees generated from independently sampled training data by selecting a random subset of the features at each split in the learning process.

We tested a wide range of implementations that are compatible with ARX's license and decided that we need to rely on different frameworks to integrate scalable implementations of different techniques. For this reason, we had to create a common interface already mentioned above to abstract away the details of specific implementations. We integrated logistic regression from Apache Mahout [31] and both naïve Bayes and random forest from Smile [32].

Assessing prediction performance

Preprocessing training data

The creation of prediction models typically involves the process of reviewing models and iteratively refining parameters to achieve optimal performance. This requires metrics for performance assessment. A commonly used method is to calculate performance measures using *k-fold* cross-validation [33]. In this process, the records of a dataset are first divided randomly into k partitions of equal size, which are then iteratively analyzed by using each of the k partitions as evaluation and all other partitions as training data. This process yields k results which are combined to derive an overall estimate of the model's performance.

When classification models are built from anonymized data, it needs to be evaluated how anonymization has affected their performance. This cannot be implemented "naively" by comparing the results of performing *k-fold* cross-validation on the anonymized data and of performing *k-fold* cross-validation on input data. Instead, a classifier must be built from transformed output data in such a way that the model is able to make predictions based on features which have not been transformed. As a result, the model can be evaluated using unmodified input data to obtain relative performance estimates [34]. 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 anonymized training data has been transformed before passing it to the classifier to make predictions [35]. Figure 4 visually contrasts both approaches. It can be seen that in the naive approach two classifiers are built from two different datasets (input and output), evaluated against these datasets and then their accuracy is compared to derive a relative performance. In our tool, the second classifier is built from output data but evaluated on (preprocessed) input data to obtain comparable results for both models.

Our tool creates privacy-preserving models by training them on anonymized data. This results in the challenge that the prediction models created can only be applied to data that has been transformed in the same way as the anonymized training dataset. Thus, we had to ensure that the resulting prediction models are able to interpret features from output data as well as input data correctly. This is challenging when the domain of attribute values is not preserved during anonymization, as in these cases, the input contains values which are not present in the output and thus the classifier would have to be evaluated with values which it has not seen during training. As a solution, we implemented a preprocessing step that accounts for the different types of transformations supported (see beginning of this section).

Whether the preprocessing step needs to be applied to a specific variable depends on the type of the variable and the transformation method utilized. Table 1 shows an overview. "N/A" indicates that the transformation method cannot be used for variables of the according type. For instance, aggregation is typically only applied to numeric attributes. It can be seen that for all types of suppression (cell, attribute, record), random sampling as well as aggregation, evaluation data does not have to be preprocessed. The reason is that the domain is being preserved

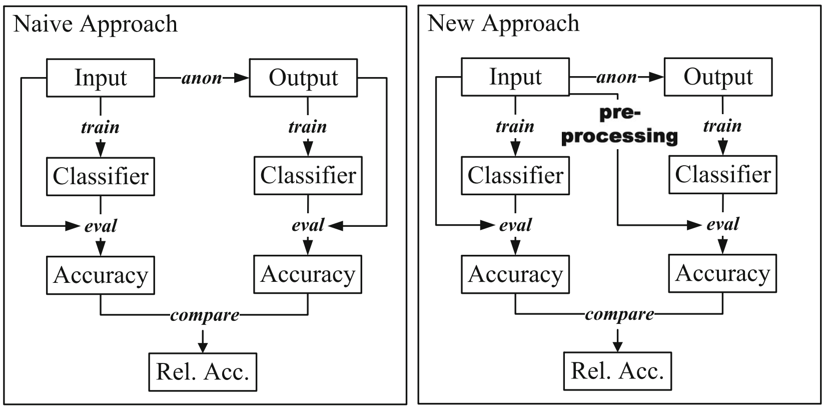


Fig. 4 Different approaches for measuring the relative performance of a privacy-preserving classifier. Our tool implements a method that makes sure that the performance of prediction models can be expressed relative to the performance of models trained on unmodified data

during transformation. With all remaining transformation schemes, data needs to be preprocessed before handing it to the classifier for evaluation. As can be seen, preprocessing only needs to be performed for attribute values that have been generalized or categorized. In both cases, this can be implemented by applying the same generalization hierarchies or categorization functions to input data that have also been used to anonymize the training dataset. During the evaluation process this is performed automatically as all relevant information on how input data has been transformed is known to the software. For the purpose of utilizing the output data generated by ARX to build a privacy-preserving prediction model outside of the software, according export functionalities (e.g. for hierarchies) are provided.

Performance assessment

All implemented classification models are able to handle multinomial classification tasks, where the target variables need not be dichotomous. The main reason behind this design decision is that we wanted our methods to integrate seamlessly with the remaining functionalities of ARX, without imposing any major restrictions. However, assessing the performance of multinomial classifiers is non-trivial and subject of ongoing research [20]. Our previous implementation therefore only supported very rudimentary performance measurements [18]. One method to overcome this limitation is the *one-vs-all* approach, in which the performance of a *n-nomial* classifier is assessed by interpreting it as a collection of *n* binomial classifiers, each of which is able to distinguish one selected class from all others.

Table 1 Overview of transformation schemes and their preprocessing requirements

Transformation scheme	Preprocessing required	
	Numeric attributes	Categorical attributes
Cell suppression	No	No
Attribute suppression	No	No
Record suppression	No	No
Generalization	Yes	Yes
Categorization	Yes	N/A
Aggregation	No	N/A
Random sampling	No	No

We decided to implement this method as it is simple and enables utilizing typical parameters for prediction performance. Our implementation currently supports the following measures: (1) *sensitivity*, also called *recall* or *true positive rate*. (2) *Specificity*, also called *true negative rate*. (3) The *Receiver Operating Characteristic (ROC)* curve, which plots the true positive rate (i.e. the sensitivity) for a single class against the false positive rate (1-specificity) [36]. The ROC curve shows the trade-off between sensitivity and specificity for every possible cut-off for a prediction, i.e. any increase in sensitivity will be accompanied by a decrease in specificity. (4) The *Area Under the ROC Curve* (ROC AUC), which summarizes the ROC performance of a classifier and which is equivalent to the probability that the classifier will assign a higher score to a randomly chosen positive event than to a randomly chosen negative event [36]. (5) The *Brier score*, which measures the mean squared distance between predicted and actual outcomes [37].

In addition to the models described previously, we always evaluate the performance of the *Zero Rule (0-R) algorithm*, which ignores the feature variables and simply always returns the most frequent class value. The performance of this simplistic "prediction model" is frequently used as a realistic baseline for assessing the performance of more sophisticated machine learning algorithms. In our tool, the performance of privacy-preserving models is reported in absolute terms as well as relative to baseline (0-R) and the selected classifier, both trained on unmodified input data.

As an additional measure specific to our application scenario, we implemented the *skill score*, which quantifies the relative accuracy of a classification model over some reference accuracy [38]. In our case, the relative accuracy is the accuracy of the classification model built from anonymized data over the accuracy of the model built from original data. Typically, the accuracy is represented by a metric such as the Brier score, leading to the following definition:

$$Brier\ skill\ score = 1 - \frac{Brier_{anonymized}}{Brier_{original}}$$

A skill score of zero means that the Brier scores for models built on output and input data are equal. If the score is in the range $]\,0, 1]$ then the model built on output data performed better and if it is in the range $[\,-\infty, 0[$, the model trained on the original data performed better.

Results

Interfaces for end users and applications

ARX's views and interfaces for data anonymization and privacy risk analysis have been described in previous publications [19, 39] and are also explained in depth on the project website [22]. Here, we will focus on the views and interfaces provided for analyzing the performance of prediction models. All methods described in the previous sections have been implemented into the Graphical User Interface (GUI) and they are also available via the software's comprehensive Application Programming Interface (API).

Figure 5 shows a screenshot of the graphical interface in which methods for configuring prediction models as well as for assessing their performance have been implemented. Areas 1 and 2 can be used to graphically assess the performance of privacy-preserving models. Both views are available side-by-side for input data and output data to allow for visual comparisons. They show basic performance parameters and ROC curves for models built with original and anonymized data, respectively. Areas 3 and 4 can be used to select target variables as well as feature variables and to configure model types and their parameters.

Case studies

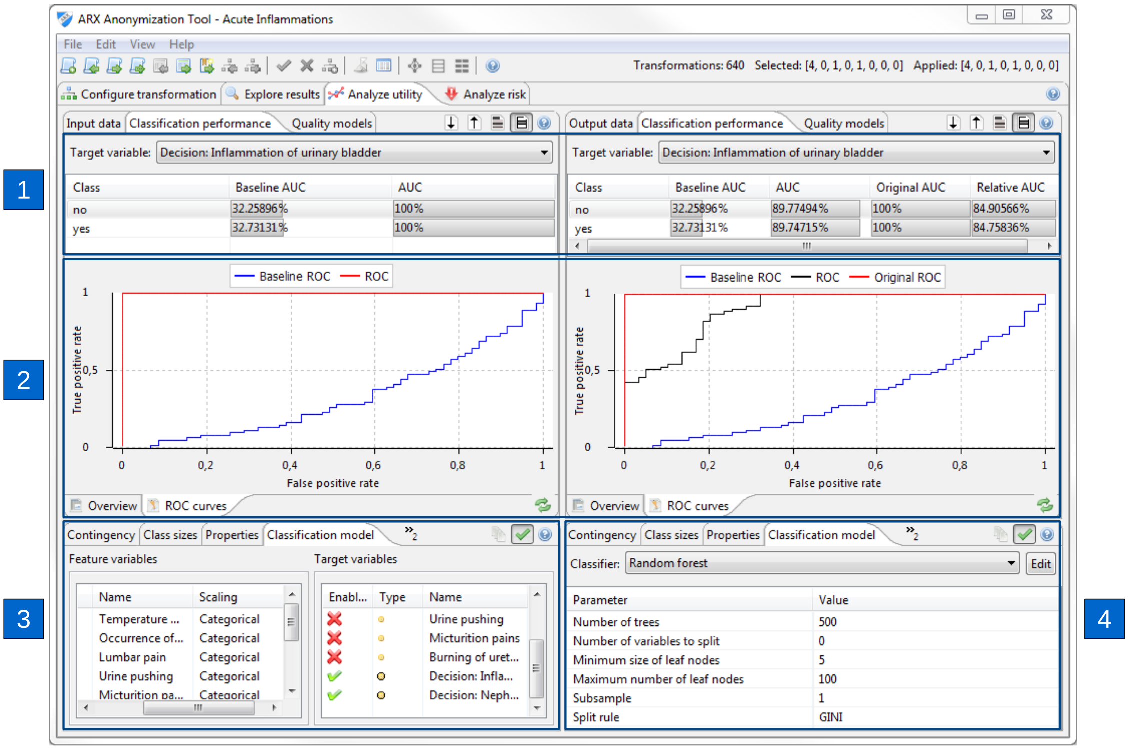
In this section, we will present three case studies to illustrate our solution and to show its practical applicability. For this purpose, we have selected three datasets to build different types of models for different biomedical prediction tasks. We have deliberately selected datasets that are challenging to anonymize as they contain a small number of records (between 120 and 1473). We will use the visualizations provided by ARX to discuss the utility and privacy protection provided by the resulting models. In all cases, we measured execution times for data anonymization as well as model building and evaluation of not more than a few seconds on commodity hardware.

Case study 1: acute inflammation of the urinary system

In the first case study, we used a dataset containing 120 records that were originally collected for testing expert systems. The task is to diagnose two diseases of the urinary system: acute inflammation of the bladder and acute nephritises. The dataset contained nine numeric and binary attributes, two of which represented the target classes. More details can be found in the original publication [40] and the publicly available version of the dataset [41]. As a privacy model we used k-anonymity, which protects the records in the training set from re-identification. We used common parameterizations of $5 \leq k \leq 25$ and random forests as prediction models. Data was transformed using aggregation, generalization and record suppression.

Figure 6 shows the results obtained for one of the two target variables (inflammation of urinary bladder). For comparison, the blue line shows the performance achieved when always returning the most frequent class attribute (0-R). In the first two plots, the ROC of models trained on unmodified training data and anonymized data is identical. We measured a relative ROC AUC (relative to the trivial classifier and to the performance of models trained on input data) of 100% for $k = 5$ and $k = 10$ and $k = 15$. For higher values of k, performance dropped to 87.72% for $k = 20$, 48.37% for $k = 25$. The Brier skill scores changed from 0 to 0.08, -0.78, -1.25 and -4.05. For $k \leq 20$, which offers a very high degree of protection [42], the resulting privacy-preserving models exhibited high prediction power.

When anonymizing data, ARX may determine that an optimal balance between privacy protection and output data utility is achieved by completely generalizing (and thereby actually removing) one or multiple attributes. This can be interpreted as automated dimensionality reduction or feature selection. Figure 7 shows that for $k =$ 15 three out of six feature variables were removed (Missings = 100%). From the results presented in the previous paragraph we can see that this had only a minor impact on prediction performance, which implies that the variables



Transformations: 640 Selected: [4, 0, 1, 0, 1, 0, 0, 0] Applied: [4, 0, 1, 0, 1, 0, 0, 0]

Target variable: Decision: Inflammation of urinary bladder

Class	Baseline AUC	AUC
no	32.25896%	100%
yes	32.73131%	100%

Target variable: Decision: Inflammation of urinary bladder

Class	Baseline AUC	AUC	Original AUC	Relative AUC
no	32.25896%	89.77494%	100%	84.90566%
yes	32.73131%	89.74715%	100%	84.75836%

Feature variables

Name	Scaling
Temperature ...	Categorical
Occurrence of...	Categorical
Lumbar pain	Categorical
Urine pushing	Categorical
Micturition pa...	Categorical

Target variables

Enabl...	Type	Name
✗	●	Urine pushing
✗	●	Micturition pains
✗	●	Burning of uret...
✓	○	Decision: Infla...
✓	○	Decision: Neph...

Classifier: Random forest

Parameter	Value
Number of trees	500
Number of variables to split	0
Minimum size of leaf nodes	5
Maximum number of leaf nodes	100
Subsample	1
Split rule	GINI

Fig. 5 Screenshot of the view implemented for assessing the performance of privacy-preserving prediction models. Area 1: Comparison of basic performance parameters, Area 2: ROC curves for models built with original and anonymized data, Area 3: Selection of feature and class variables, Area 4: Selection and configuration of model parameters

that have been removed are not predictive for the target variable. If the target variable needs to be protected from inference attacks, this information can be used as an indicator that the variables that have been removed may not needed to be transformed at all.

Finally, Fig. 8 shows re-identification risk profiles provided by ARX (cf. [39]). A risk profile summarizes the risks of all records in a dataset, by associating each possible risk level with the relative number of records which are affected. It can be seen that k-anonymity with $k = 15$ significantly reduced the risk of re-identification for all records in the dataset, highlighting the high degree of privacy protection that can be achieved with negligible effects on prediction performance.

Case study 2: breast cancer cytopathology

In the second case study, we utilized a dataset which contained 699 records collected by the University of Wisconsin Hospitals to study methods for predicting the malignancy of breast tissue from cytopathology reports. It

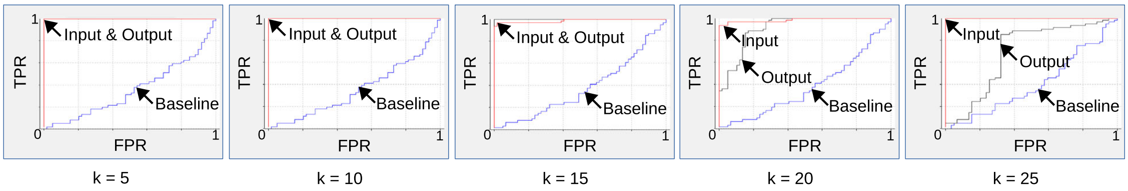
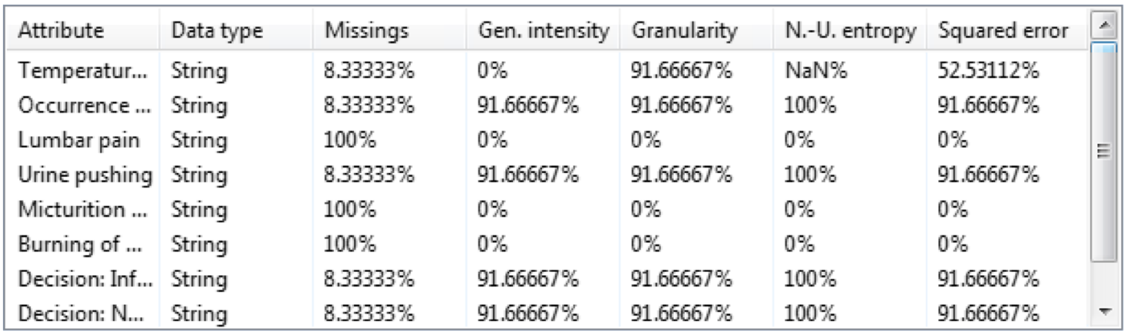


Fig. 6 ROC performance in the case study using k-anonymous data for training random forests on the acute inflammation dataset. The False Positive Rates (FPR) and True Positive Rates (TPR) are plotted against the x-axes and y-axes, respectively. It can be seen that data anonymization had a negative impact on the performance of the resulting prediction models only for $k \geq 15$



Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
Temperatur...	String	8.33333%	0%	91.66667%	NaN%	52.53112%
Occurrence ...	String	8.33333%	91.66667%	91.66667%	100%	91.66667%
Lumbar pain	String	100%	0%	0%	0%	0%
Urine pushing	String	8.33333%	91.66667%	91.66667%	100%	91.66667%
Micturition ...	String	100%	0%	0%	0%	0%
Burning of ...	String	100%	0%	0%	0%	0%
Decision: Inf...	String	8.33333%	91.66667%	91.66667%	100%	91.66667%
Decision: N...	String	8.33333%	91.66667%	91.66667%	100%	91.66667%

Fig. 7 Automated dimensionality reduction performed by ARX starting from $k = 15$ when anonymizing the acute inflammation dataset. For larger values of k, ARX performs automated dimensionality reduction during data anonymization. By comparing the results with the ROC curves in Fig. 6 it can be seen that the removal of three out of six feature variables had only a minor impact on prediction performance

contained 10 numeric and binary attributes, one of which represented the target class (malignant or benign tissue). The dataset and further details are available online [41].

For privacy protection, we utilized (ϵ, δ)-differential privacy with $\epsilon \in \{2, 1.5, 1.0, 0.5, 0.1\}$ and $\delta = 10^{-3}$. We used logistic regression as modeling technique. Implementing differential privacy requires randomization and we therefore report on the best model obtained from five anonymization processes performed for each parameterization. Data was transformed using random sampling, categorization, generalization and record suppression. The results are shown in Fig. 9.

As can be seen in the figure, prediction performance decreased with decreasing values of epsilon, which was to be expected as the degree of privacy protection increases when epsilon decreases. Moreover, the results confirm prior findings which indicated that a value of about $\epsilon = 1$ is an optimal parameterization for the differentially private anonymization algorithm implemented by ARX [24]. Furthermore, we studied the effect of randomization on the stability of the performance of the models created. The prediction model trained on unmodified input data achieved a ROC AUC of about 99.2%. For the five models created with $\epsilon = 1$ we measured a ROC AUC of between 85.8% and 92.27% (88.28% on average) which equals a relative ROC AUC of between 61.63% and 83.96% (74.80% on average) compared to baseline performance and the model trained on unmodified data. The Brier skill score varied between -1.38 and -3.45 (-2.66 on average), which is quite good considering the high degree of privacy protection provided.

Finally, Fig. 10 shows the risk profiles provided by ARX for the best model obtained using $\epsilon = 1$. As can be seen, re-identification risks were reduced to an extent even larger than in the previous case study. Moreover, we also found that ARX performed significant dimensionality

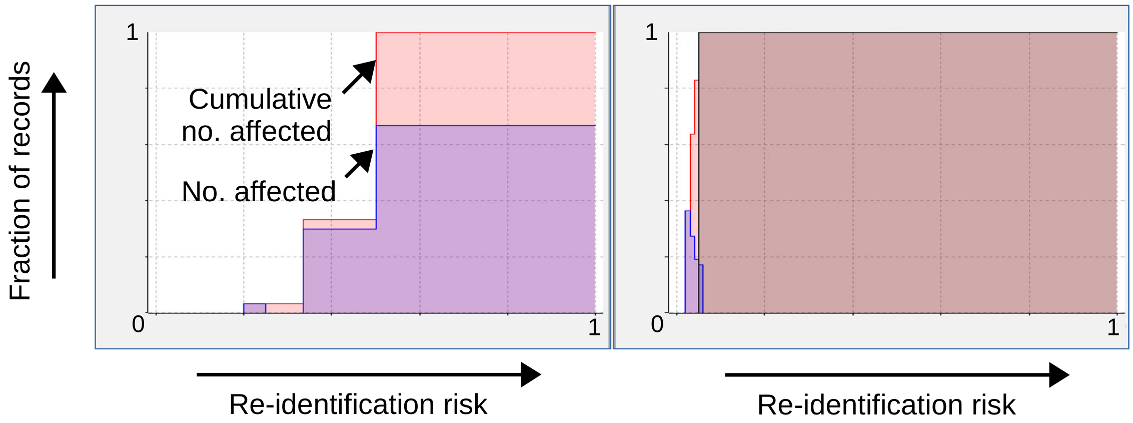


Fig. 8 Impact of data anonymization on re-identification risk profiles for the acute inflammation dataset. As can be seen, k-anonymity with $k = 15$ significantly reduced the risk of re-identification for all records in the dataset

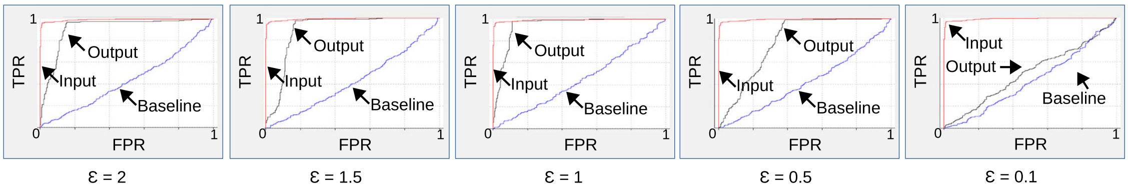


Fig. 9 ROC performance in the case study using differential privacy for training logistic regression models to predict the malignancy of breast tissue. The False Positive Rates (FPR) and True Positive Rates (TPR) are plotted against the x-axes and y-axes, respectively. It can be seen that data anonymization had a significant impact on prediction performance, but acceptable accuracy could still be observed for $\epsilon \geq 1$

reduction and that malignancy was basically predicted from a single attribute (bland chromatin).

Case study 3: use of contraceptive methods

In the third case study, we utilized a dataset consisting of 1473 records from the 1987 National Indonesia Contraceptive Prevalence Survey to predict the contraceptive method used of women based on their demographic and socio-economic characteristics. The dataset contained 10 numeric, categorical and binary attributes, one of which represented the target class (type of contraceptive method used). More details can be found in the original publication [43] and the dataset is available online [41].

For privacy protection, we employed an innovative game-theoretic method that works on the assumption that adversaries will only attack a dataset (or prediction model) if there is a tangible economic benefit. For parameterizing the method, we followed the proposal by Wan et al. [25]: the cost for the adversary of trying to re-identify an individual was set to \$4 (a number that has been derived from the costs of obtaining detailed personal information online) and the monetary benefit of including a record into the training set was assumed to be \$1200 (this number was derived from an analysis of grant funding received and data shared by the Electronic Medical Records and Genomics (eMERGE) Network [44], which is funded by the National Institute of Health (NIH)).

We considered a single free parameter G, which specified the monetary gain of the adversary in case of successful re-identification and, at the same time, the monetary loss for the data controller for each successfully re-identified record. By varying this single parameter we were able to investigate a wide variety of scenarios, in which either the data controller or the adversary was at an advantage. For prediction, we used Naïve Bayes classifiers. Data was transformed using categorization, generalization as well as cell and record suppression.

Overall, as can be seen in Fig. 11, we found that anonymizing the dataset with $G = 0, 500, 1000, 1500$ and 2000 had only a very limited impact on the performance of the resulting privacy-preserving prediction models. Models trained on unmodified input data achieved a ROC AUC of 71.82%. We were not able to observe a relationship between privacy parameters and the prediction performance of the privacy-preserving models. The reason is that the game-theoretic model contains an implicit data

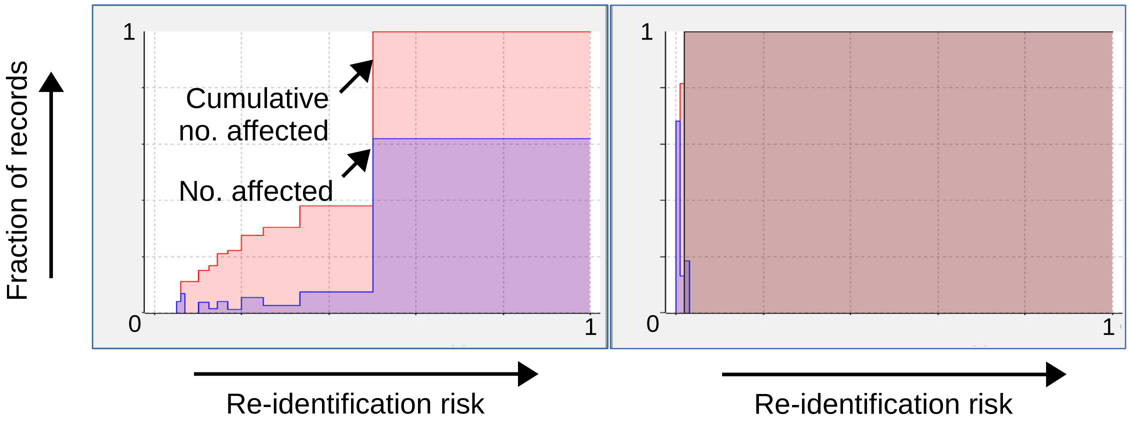


Fig. 10 Impact of data anonymization on re-identification risk profiles for the breast cancer dataset. As can be seen, the differential privacy model with $\epsilon = 1$ resulted in the strongest reductions to re-identification risks of all models used in the case studies

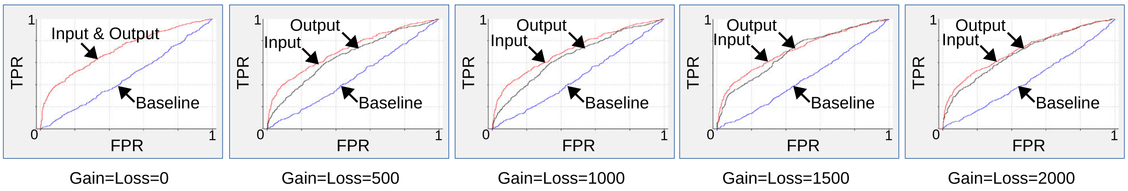


Fig. 11 Impact of data anonymization on prediction performance in the contraceptive methods case study. The False Positive Rates (FPR) and True Positive Rates (TPR) are plotted against the x-axes and y-axes, respectively. As can be seen, data anonymization using the game-theoretic model had only a very minor impact on prediction accuracy

quality model that does not directly reflect the suitability of data for training prediction models. We measured a relative ROC AUC between 77.33% and 100% (90.35% on average) and Brier skill scores between -0.04 and 0 (-0.02 on average). Analogously to the other studies, we observed a significant reduction of re-identification risks.

Discussion

Comparison with prior work

Early work has suggested that anonymization destroys the utility of data for machine learning tasks [45]. Many methods for optimizing anonymized data as a training set for prediction models have since been developed. They show that this is not actually true. Initially, these methods focused on simple anonymization techniques, such as *k-anonymity*, and simple prediction models, such as decision trees and on applications in distributed settings [35, 46]. As a result of these developments, evaluating (novel) anonymization methods by measuring the usefulness of output data for predictive modeling tasks has become a standard practice in academia [47, 48]. More recently, a broader spectrum of prediction and privacy models has been investigated. Some authors proposed general-purpose anonymization algorithms to optimize prediction performance. While most of these algorithms have been designed in such a way that the resulting anonymized data is guaranteed to provide a degree of protection based on specific privacy models only [49, 50], they allow for any type of prediction model to be used. In contrast, in other works, privacy-preserving algorithms for optimizing the performance of specific prediction models were developed [51, 52]. Many recent studies focused on sophisticated models, such as support vector machines [51, 53, 54] and (deep) neural networks [55–57]. More complex and comprehensive privacy models have also received significant attention. In particular, the differential privacy model was investigated extensively [53, 55, 56, 58–62]. It is notable, that among these more modern approaches, a variety has focused on biomedical data [56, 57, 60]. We note, however, that these developments originate from the computer science research community and if the developed algorithms are published, then typically only in the form of research prototypes.

In parallel, several practical tools have been developed that make methods of data anonymization available to end-users by providing easy-to-use graphical interfaces. Most notably, $\mu - ARGUS$ [63] and *sdcMicro* [64] are tools developed in the context of official statistics, while ARX has specifically been designed for applications to biomedical data [19]. μ-ARGUS and sdcMicro focus on the concept of *a posteriori disclosure risk control* which is prevalent in the statistics community. In this process, data is mainly transformed manually in iterative steps, while data utility, usefulness and risks are monitored continuously by performing statistical analyses and tests. ARX implements a mixture of this approach and the *a priori disclosure risk control* methodology. This means that data is anonymized semi-automatically. In each iteration, the data is sanitized in such a way that predefined thresholds on privacy risks are met while the impact on data utility is minimized. A balancing is performed by repeating this process with different settings, thereby iteratively refining output data. This approach has been recommended for anonymizing health data (see, e.g. [7, 12] and [13]) and it enables ARX to support an unprecedentedly broad spectrum of techniques for transforming data and measuring risks. All three tools provide users with methods for assessing and optimizing the usefulness of anonymized data for a wide variety of applications. ARX is, however, the only tool providing support for privacy-preserving machine learning.

Limitations and future work

Currently, our tool only supports three different types of prediction models, i.e. logistic regression, naïve Bayes and random forest, for which we could find scalable implementations that are compatible to ARX in terms of their technical basis and licensing model. However, further approaches, e.g. C4.5 decision trees and support vector machines, have also received significant attention in the literature (see e.g. [49–51, 53, 54, 58, 60, 62]). In future work, we plan to extend our implementation accordingly.

Moreover, choosing the right type of prediction model for a specific dataset and task is challenging, as there are no general recommendations [20]. Therefore, benchmark studies are often performed, in which the results of different models are experimentally compared for a specific dataset using a complex process involving the separation of data into training sets, evaluation sets and validation sets [65]. In future work, we plan to extend our implementation to support such benchmark studies for privacy-preserving models as well.

In this article we have focused on transformation techniques supported by ARX for which a preprocessing step can be implemented by applying a known transformation function to features (see "Preprocessing training data" section). The software, however, also supports transformation approaches where it is not clear how a given feature must be transformed to match the representation used for training purposes. Local generalization is an important example. In this case, the same attribute value can be transformed to different generalized representations in different records of the training set. When providing features to the model to make predictions, it is therefore unclear how the values of such attributes must be generalized. One approach to overcome this challenge is to apply all possible transformations and to then analyze which transformation results in the prediction with the highest confidence. However, this involves a high degree of complexity and we therefore plan to develop more scalable approaches in the future.

Finally, our current implementation focuses on classification tasks. In future work, we plan to provide support for further learning and prediction tasks that are of specific importance to medical research. Important examples include regression and time-to-event analysis [20].

Conclusions

In this paper, we have presented a comprehensive tool for building and evaluating privacy-preserving prediction models. Our implementation is available as open source software. We have further presented three case studies which show that, in many cases, a high degree of privacy protection can be achieved with very little impact on prediction performance. Our tool supports a wide range of transformation techniques, methods for privacy protection and prediction models. The methods supported are particularly well suited for applications to biomedical data. Notably, the truthful transformation methods implemented prevent implausible data from being created (e.g. combinations or dosages of drugs which are harmful for a patient) [66]. Moreover, methods of privacy preservation have been implemented in a way that is relatively easy to explain to ethics committees and policy makers, as they basically rely on the intuitive idea of hiding in a crowd [24]. To our knowledge, ARX is the only publicly available anonymization tool supporting a comprehensive set of methods for privacy-preserving machine learning in an integrated manner.

Availability and requirements

- **Project name**: ARX Data Anonymization Tool
- **Project home page**: https://arx.deidentifier.org/
- **Operating system(s)**: Platform independent
- **Programming language**: Java
- **Other requirements**: Java 1.8 or higher
- **License**: Apache License, Version 2.0
- **Any restrictions to use by non-academics**: No

Abbreviations

0-R: Zero rule; API: Application programming interface; AUC: Area under the curve; eMERGE: Electronic medical records and genomics; GUI: Graphical user interface; HIPAA: U.S. health insurance portability and accountability act; NIH: National institute of health; ROC: Receiver operating characteristic

Acknowledgements

The authors gratefully acknowledge the insightful discussions with Alexander Hapfelmeier about methods for evaluating the performance of multinomial classifiers.

Authors' contributions

JE and FP have developed the methods. JE, KK and FP wrote the manuscript. JE, RB, HS and FP implemented the algorithms and performed the case studies. All authors discussed the manuscript at all stages. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets used during the current study are available from the corresponding author on reasonable request

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Medicine, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany. [2]Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Straße 2, 10178 Berlin, Germany. [3]Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany.

Received: 27 June 2019 Accepted: 30 January 2020
Published online: 11 February 2020

References

1. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin oncol. 2011;8(3):184.
2. Schneeweiss S. Learning from big health care data. N Engl J Med. 2014;370(23):2161–3.
3. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24.
4. Liu V, Musen MA, Chou T. Data breaches of protected health information in the United States. JAMA. 2015;313(14):1471–3.
5. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395–405.

6. Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J Invest Med. 2010;58(1):11–18.
7. El Emam K, Malin B. Appendix B: Concepts and Methods for De-identifying Clinical Trial Data. In: Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC: The National Academies Press; 2015.
8. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255–60.
9. Shokri R, Shmatikov V. Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. New York: ACM; 2015. p. 1310–1321.
10. Dankar FK, Madathil N, Dankar SK, Boughorbel S. Privacy-Preserving Analysis of Distributed Biomedical Data: Designing Efficient and Secure Multiparty Computations Using Distributed Statistical Learning Theory. JMIR Med Inform. 2019;7(2):e12702.
11. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE; 2017. https://doi.org/10.1109/sp.2017.41.
12. El Emam K, Arbuckle L. Anonymizing health data: Case studies and methods to get you started. 1st ed. Sebastopol: O'Reilly Media, Inc.; 2013.
13. Xia W, Heatherly R, Ding X, Li J, Malin BA. R-U policy frontiers for health data de-identification. J Am Med Inform Assoc. 2015;22(5):1029–41.
14. Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Symposium on Security and Privacy. IEEE; 2008. p. 111–125.
15. Sweeney L. Computational disclosure control - A primer on data privacy protection. Cambridge: Massachusetts Institute of Technology; 2001.
16. United States. The Health Insurance Portability and Accountability Act (HIPAA). Washington: U.S. Dept. of Labor, Employee Benefits Security Administration; 2004.
17. EU General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). Off J Eur Union. 2016;1:119.
18. Prasser F, Eicher J, Bild R, Spengler H, Kuhn KA. A Tool for Optimizing De-identified Health Data for Use in Statistical Classification. In: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2017. https://doi.org/10.1109/cbms.2017.105.
19. Prasser F, Kohlmayer F. Putting statistical disclosure control into practice: The ARX data anonymization tool. In: Medical Data Privacy Handbook. Springer International Publishing; 2015. p. 111–148. https://doi.org/10.1007/978-3-319-23633-9_6.
20. Witten IH, Eibe F. Data mining: Practical machine learning tools and techniques. San Francisco: Morgan Kaufmann; 2016.
21. Prasser F, Kohlmayer F, Kuhn KA. Efficient and effective pruning strategies for health data de-identification. BMC Med Inform Decis Making. 2016;16(1):49.
22. ARX - Power Data Anonymization. http://arx.deidentifier.org/. Accessed 21 June 2019.
23. Dwork C. Differential privacy. In: Encyclopedia of Cryptography and Security. Heidelberg: Springer; 2011. p. 338–340.
24. Bild R, Kuhn KA, Prasser F. SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees. Proc Priv Enhancing Technol. 2018;2018(1):67–87.
25. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Ganta R, et al. A game theoretic framework for analyzing re-identification risk. PloS One. 2015;10(3):e0120592. Cambridge.
26. Prasser F, Gaupp J, Wan Z, Xia W, Vorobeychik Y, Kantarcioglu M, et al. An Open Source Tool for Game Theoretic Health Data De-Identification. In: AMIA Annual Symposium Proceedings. AMIA; 2017. Accepted for AMIA 2017 Annual Symposium (AMIA 2017).
27. Iyengar VS. Transforming data to satisfy privacy constraints. In: International Conference on Knowledge Discovery and Data Mining. ACM; 2002. p. 279–88.
28. World Health Organization. International statistical classification of diseases and related health problems. 2016. https://www.who.int/classifications/icd/en/. Accessed 21 June 2019.
29. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn. 1997;29(2):103–130.
30. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
31. Apache Software Foundation. Apache Mahout: Scalable machine-learning and data-mining library. 2011. http://mahout.apache.org/. Accessed 21 June 2019.
32. Smile – Statistical Machine Intelligence and Learning Engine. https://haifengl.github.io/smile/. Accessed 21 June 2019.
33. Bailey TL, Elkan C. Estimating the Accuracy of Learned Concepts. In: Proceedings of the 13th International Joint Conference on Artifical Intelligence. San Francisco: Morgan Kaufmann Publishers Inc.; 1993. p. 895–900.
34. Li T, Li N. On the tradeoff between privacy and utility in data publishing. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09. ACM Press; 2009. https://doi.org/10.1145/1557019.1557079.
35. Inan A, Kantarcioglu M, Bertino E. Using anonymized data for classification. In: 2009 IEEE 25th International Conference on Data Engineering. IEEE; 2009. https://doi.org/10.1109/icde.2009.19.
36. Fawcett T. An introduction to ROC analysis. Pattern Recog Lett. 2006;27(8):861–74.
37. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3.
38. Wilks DS. Sampling distributions of the Brier score and Brier skill score under serial dependence. Q J R Meteorol Soc. 2010;136(653):2109–18.
39. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. IEEE J Biomed Health Inform. 2017;22(2):611–22.
40. Czerniak J, Zarzycki H. Application of rough sets in the presumptive diagnosis of urinary system diseases. In: Artificial Intelligence and Security in Computing Systems. Springer; 2003. p. 41–51. https://doi.org/10.1007/978-1-4419-9226-0_5.
41. Dua D, Graff C. UCI Machine Learning Repository. 2017. http://archive.ics.uci.edu/ml. Accessed 21 June 2019.
42. European Medicines Agency. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. 20161–99. EMA/90915/2016.
43. Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci. 1990;87(23):9193–6.
44. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4(1):13.
45. Brickell J, Shmatikov V. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In: 14th International Conference on Knowledge Discovery and Data Mining (SIGKDD). Las Vegas: ACM; 2008. p. 70–78.
46. Aggarwal CC, Yu PS. A general survey of privacy-preserving data mining models and algorithms. In: Privacy-Preserving Data Mining. Springer; 2008. p. 11–52. https://doi.org/10.1007/978-0-387-70992-5_2.
47. Fung BCM, Wang K, Fu AWC, Yu PS. Introduction to privacy-preserving data publishing: Concepts and techniques. 1st ed. Boca Raton: CRC Press; 2010.
48. Malle B, Kieseberg P, Holzinger A. Do not disturb? classifier behavior on perturbed datasets. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Springer; 2017. p. 155–73. https://doi.org/10.1007/978-3-319-66808-6_11.
49. Li J, Liu J, Baig M, Wong RCW. Information based data anonymization for classification utility. Data Knowl Eng. 2011;70(12):1030–45.
50. Last M, Tassa T, Zhmudyak A, Shmueli E. Improving accuracy of classification models induced from anonymized datasets. Inf Sci. 2014;256:138–161.
51. Lin KP, Chen MS. On the design and analysis of the privacy-preserving SVM classifier. IEEE Trans Knowl Data Eng. 2011;23(11):1704–17.
52. Fong PK, Weber-Jahnke JH. Privacy preserving decision tree learning using unrealized data sets. Trans Knowl Data Eng. 2012;24(2):353–364.
53. Sazonova V, Matwin S. Combining Binary Classifiers for a Multiclass Problem with Differential Privacy. Trans Data Priv. 2014;7(1):51–70.
54. Mancuhan K, Clifton C. Statistical Learning Theory Approach for Data Classification with ℓ-diversity. In: Proceedings of the 2017 SIAM International Conference on Data Mining. SIAM; 2017. p. 651–659. https://doi.org/10.1137/1.9781611974973.73.
55. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: Proceedings of the 2016 SIGSAC Conference on Computer and Communications Security. New York: ACM; 2016. p. 308–318.
56. Esteban C, Hyland SL, Rätsch G. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. arXiv preprint arXiv:170602633. 2017.

57. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating Multi-label Discrete Electronic Health Records using Generative Adversarial Networks. arXiv preprint arXiv:170306490. 2017.
58. Friedman A, Schuster A. Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10. ACM; 2010. https://doi.org/10.1145/1835804.1835868.
59. Zhang N, Li M, Lou W. Distributed data mining with differential privacy. In: IEEE International Conference on Communications (ICC). IEEE; 2011. https://doi.org/10.1109/icc.2011.5962863.
60. Jiang X, Ji Z, Wang S, Mohammed N, Cheng S, Ohno-Machado L. Differential-private data publishing through component analysis. Trans Data Priv. 2013;6(1):19.
61. Zaman ANK, Obimbo C, Dara RA. A Novel Differential Privacy Approach that Enhances Classification Accuracy. In: Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering - C3S2E '16. ACM; 2016. https://doi.org/10.1145/2948992.2949027.
62. Zaman ANK, Obimbo C, Dara RA. An Improved Data Sanitization Algorithm for Privacy Preserving Medical Data Publishing. In: Canadian Conference on Artificial Intelligence. Basel: Springer; 2017. p. 64–70.
63. De Waal A, Hundepool A, Willenborg L. C. R. J. Argus: Software for statistical disclosure control of microdata. US Census Bureau; 1995.
64. Templ M. Statistical disclosure control for microdata using the R-package sdcMicro. Trans Data Priv. 2008;1(2):67–85.
65. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2016.
66. Dankar FK, El Emam K. Practicing differential privacy in health care: A review. Trans Data Priv. 2013;6(1):35–67.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.2 Other Publications

A.2.1 Flexible data anonymization using ARX – Current status and challenges ahead

Prasser, Fabian, **Johanna Eicher**, Helmut Spengler, Raffael Bild, and Klaus A. Kuhn. "Flexible data anonymization using ARX—Current status and challenges ahead." *Software: Practice and Experience* 50, no. 7 (2020): 1277-1304.

Received: 6 August 2019 | Revised: 27 January 2020 | Accepted: 28 January 2020

DOI: 10.1002/spe.2812

RESEARCH ARTICLE

WILEY

Flexible data anonymization using ARX—Current status and challenges ahead

Fabian Prasser[1] | **Johanna Eicher**[2] | **Helmut Spengler**[2] | **Raffael Bild**[2] | **Klaus A. Kuhn**[2]

[1]Medical Informatics Lab, Berlin Institute of Health (BIH) and Charité Universitätsmedizin Berlin, Berlin, Germany

[2]School of Medicine, Institute of Medical Informatics, Statistics and Epidemiology, Technical University of Munich, Munich, Germany

Correspondence
Fabian Prasser, Medical Informatics Lab, Berlin Institute of Health (BIH), Charité Universitätsmedizin Berlin,
Charitéplatz 1, 10117 Berlin, Germany.
Email: fabian.prasser@charite.de

Summary

The race for innovation has turned into a race for data. Rapid developments of new technologies, especially in the field of artificial intelligence, are accompanied by new ways of accessing, integrating, and analyzing sensitive personal data. Examples include financial transactions, social network activities, location traces, and medical records. As a consequence, adequate and careful privacy management has become a significant challenge. New data protection regulations, for example in the EU and China, are direct responses to these developments. Data anonymization is an important building block of data protection concepts, as it allows to reduce privacy risks by altering data. The development of anonymization tools involves significant challenges, however. For instance, the effectiveness of different anonymization techniques depends on context, and thus tools need to support a large set of methods to ensure that the usefulness of data is not overly affected by risk-reducing transformations. In spite of these requirements, existing solutions typically only support a small set of methods. In this work, we describe how we have extended an open source data anonymization tool to support almost arbitrary combinations of a wide range of techniques in a scalable manner. We then review the spectrum of methods supported and discuss their compatibility within the novel framework. The results of an extensive experimental comparison show that our approach outperforms related solutions in terms of scalability and output data quality—while supporting a much broader range of techniques. Finally, we discuss practical experiences with ARX and present remaining issues and challenges ahead.

KEYWORDS

data anonymization, de-identification, privacy, security, software tools

Fabian Prasser and Johanna Eicher contributed equally to this study.

1 | INTRODUCTION

In the era of big data processing and artificial intelligence, the race for innovation has become a race for data. The spectrum of personal data that is collected electronically covers almost all aspects of our lives. Important examples of sensitive personal information include financial transactions, data about activities in social networks, location traces collected via mobile phone networks and medical records.[1] These data bear a tremendous potential for modern technologies to enable progress in a wide range of fields, such as economics, science, and public security. Possible applications vary from product recommender systems to health care decision support, computational criminology, and terrorism informatics.[2,3] Yet, in order to unlock this potential, data often need to be published, shared with third parties or reused for other purposes than the ones for which it was originally collected. This is a challenging task, as privacy concerns and restrictions imposed by national and international data protection laws, for example, the US Health Insurance Portability and Accountability Act (HIPAA),[4] the European General Data Protection Regulation (GDPR),[5] or the Chinese national standard on the protection of personal information,[6] need to be considered.

Data privacy can be addressed on multiple levels. The *Five Safes* framework describes one approach to conceptualize relevant safeguards in data management processes.[7] First, it can be important to ensure that *projects* are safe, which for example requires organizational measures that ensure that data use is appropriate. Second, it can be important to ensure that *people* working with the data are safe and trustworthy, for example by using strong authentication and authorization measures. Third, the *data* itself can be made safe, meaning that risks of re-identification are reduced to an acceptable minimum. Fourth, safe *settings* can be set up to reduce the risk of privacy breaches during processing, for example, by means of cryptographic protocols for secure multiparty computation.[8] Finally, the disclosure risk of *output* data can also be controlled to ensure that results do not leak sensitive personal information.

Data anonymization is an important building block for achieving safe input and output data. The basic idea is to transform data in such a way that privacy risks are reduced while the reduction of risks is balanced against a reduction of data utility.[9-13] Several high-profile re-identification attacks have demonstrated that this is a complex task requiring tool support.[14,15] For instance, simply removing directly identifying attributes, such as names or social security numbers, will typically not be enough to prevent privacy breaches.[16-18] More formal approaches are required, which employ mathematical and statistical models for quantifying risks and the impact of anonymization on data usefulness. Moreover, complex algorithms must be employed to balance both aspects in a scalable manner. We note that formal data anonymization is different from basic techniques of data masking or random data generation.[19] In this work, we focus on non-interactive microdata anonymization, which means that protected records are created from the records of an input dataset[11] and we do not cover interactive query anonymization, as, for example, implemented by PINQ[20] or Airavat.[21]

1.1 | Background

The obvious first step in any data anonymization process is to remove all direct identifiers of individuals.[11] The next—and far more challenging—step is to modify the dataset in a way that reduces the risk that an attacker is able to successfully link identified or identifiable individuals to one or multiple records or other sensitive information contained in the dataset.[17,22] In this process, the risk of such privacy breaches is quantified by mathematical or statistical *privacy models* (typically involving a threshold for what level of risk is deemed acceptable) and the utility of output data is quantified by a *utility model*. Figure 1 shows an abstract overview of an anonymization algorithm: A procedure searches through the space of all possible outputs, which is defined by one or multiple data *transformation models*, to find a solution which fulfills the risk thresholds specified for the privacy model and at the same time provides optimal output according to the utility model.

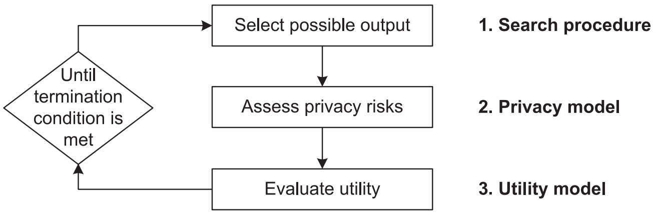
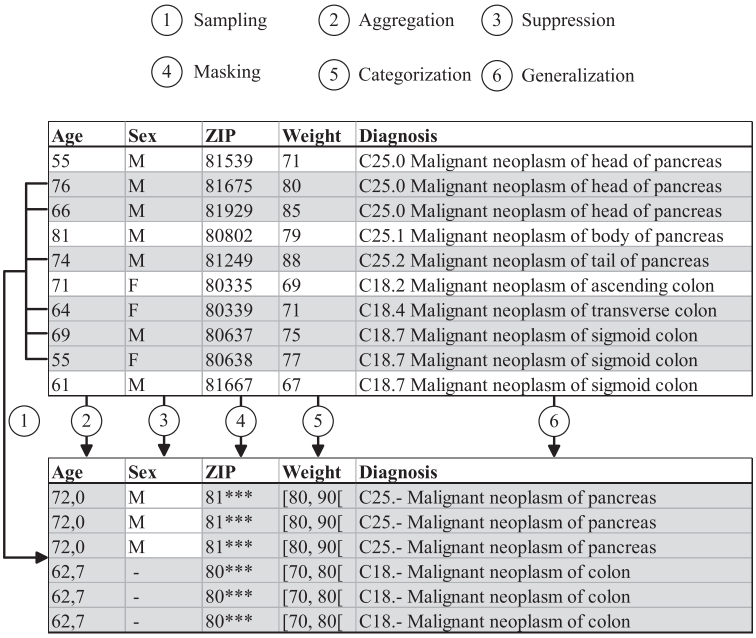


FIGURE 1 Abstract process implemented by data anonymization algorithms. A search procedure traverses the space of possible outputs while privacy models are used for assessing privacy risks and utility is evaluated using a utility model



Age	Sex	ZIP	Weight	Diagnosis
55	M	81539	71	C25.0 Malignant neoplasm of head of pancreas
76	M	81675	80	C25.0 Malignant neoplasm of head of pancreas
66	M	81929	85	C25.0 Malignant neoplasm of head of pancreas
81	M	80802	79	C25.1 Malignant neoplasm of body of pancreas
74	M	81249	88	C25.2 Malignant neoplasm of tail of pancreas
71	F	80335	69	C18.2 Malignant neoplasm of ascending colon
64	F	80339	71	C18.4 Malignant neoplasm of transverse colon
69	M	80637	75	C18.7 Malignant neoplasm of sigmoid colon
55	F	80638	77	C18.7 Malignant neoplasm of sigmoid colon
61	M	81667	67	C18.7 Malignant neoplasm of sigmoid colon

Age	Sex	ZIP	Weight	Diagnosis
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon

FIGURE 2 Example of data transformation methods. A variety of transformation techniques typically need to be combined with each other to effectively anonymize a dataset

An example using a combination of multiple transformation models is shown in Figure 2. As can be seen, a transformation might involve procedures such as taking a random sample of the records from the input dataset, aggregating numerical values and replacing them by their mean, suppressing individual values, masking parts of strings, categorizing numerical attributes, and generalizing categorical attributes. To reduce the risk of successful linkage attacks or the confidence an attacker might have in the correctness of linkage, these transformations may reduce the fidelity of data or introduce uncertainty by introducing noise.

Obviously, anonymization algorithms that support such complex transformation schemes cannot be implemented by simply searching the space of all potential output datasets for an optimal solution, as the search spaces are typically far too large. As a consequence, a wide range of heuristic strategies[23,24] and sophisticated clustering algorithms[25-29] have been developed. We emphasize, however, the importance of keeping the abstract model of data anonymization procedures implementing a specific combination of risk, utility, and transformation models in mind. For example, previous algorithms are typically only able to implement a specific combination of selected models, which severely limits their practical applicability.

As a consequence, the range of publicly available open source solutions is surprisingly small. It is well known that the effectiveness of different anonymization techniques highly depends on context, which includes the dimensionality, volume, and statistical properties of data.[12,30,31] Other important aspects that need to be considered include which types of applications or analyses the data are to be used for, whether the data will be released publicly or with additional access control and whether the data are tabular or have longitudinal or transactional characteristics. To ensure that anonymization software can be utilized for different application scenarios, different algorithms, and different methods for transforming data and quantifying reductions in usefulness must therefore be supported.[11] Moreover, many anonymization techniques involve significant computational complexity[32] which makes it challenging to implement them in a scalable manner.

1.2 | Related work

The current landscape of open source anonymization software basically consists of three types of solutions:

- First, there are tools originating from the computer science community (typically research prototypes), such as the *UTD Anonymization Toolbox*,[33] the *Cornell Anonymization Toolkit*,[34] *TIAMAT*,[35] *Anamnesia*[36] or *SECRETA*[37] and source code published as supplementary material to articles (eg, References 38 and 39). These solutions are able to automatically enforce privacy guarantees specified by users a priori. However, they usually only support a limited set of privacy models and focus on specific privacy and data transformation models.

- Second, there are tools originating from the statistics community, with *sdcMicro*[40] and μ-Argus[41] being the most prominent examples. These tools implement a more manual approach which enables them to support a wider variety of methods for measuring risks, transforming data, and analyzing the usefulness of output data. Privacy risks are typically quantified after transformations have been applied (a posteriori), which leads to an interactive anonymization process involving repeated and incremental transformations of a dataset.
- More recently, a wide range of commercial solutions has become available, often as a result to the requirements laid out in the GDPR. These closed-source tools focus on commercial markets. Typically, little is known about the underlying algorithms and they are not available for experimental evaluations and comparisons.

The ARX Data Anonymization Tool positions itself between these extremes with the aim of providing open software achieving a high degree of automation while at the same time providing supporting a wide range of techniques. In the past, various individual features and functionalities of ARX have been described in specific publications. Examples include anonymization methods based on statistical models,[42] game-theory,[43] differential privacy,[44] and an initial version of ARX's support for privacy-preserving data mining.[45] In addition, we have published two overview articles about ARX over the course of the years. The first article, which was published in 2014, focused on version 2.2.0 of ARX[46] while the second article, which was published in 2015, covered version 3.0.0 and introduced the application programming interface.[47] However, previous versions of ARX provided only limited support for complex data transformation models. We addressed this limitation in the work described in this article.

1.3 | Contributions

In the data anonymization space, it is of significant importance to distinguish between privacy models, transformation models, utility models, and anonymization algorithms. In general, a wide range of models needs to be supported to be able to address different real-world anonymization problems. However, prior algorithms typically only support a specific combination of methods. While previous versions of ARX already supported multiple privacy and utility models, only a small set of transformation techniques was available. In this work, we present a novel approach that has been implemented into the software to support (almost) arbitrary combinations of privacy and utility models with a wide range of data transformation techniques while preserving scalability.

We first present the core design principles that enable ARX to support multiple techniques for measuring privacy risks as well as output data utility while providing computational efficiency. Second, we present a novel approach for extending this design to significantly improve its genericity and flexibility regarding supported transformation methods. Next, we review the spectrum of methods supported and discuss their compatibility within the enhanced anonymization framework of the software. Then we present an extensive experimental comparison with related software. Our results show that ARX often outperforms other solutions in terms of scalability and—at the same time—output data quality, all while supporting a much broader spectrum of techniques. Finally, we discuss practical experiences with ARX, present remaining challenges and outline how we plan to address them in future work.

2 | FLEXIBLE DATA ANONYMIZATION IN ARX

2.1 | Basic design

At its core, ARX uses a highly efficient *globally-optimal search algorithm* for transforming data with *full-domain generalization* and *record suppression*. The transformation of attribute values is implemented through domain generalization hierarchies, which represent valid transformations that can be applied to individual-level values. Two examples are shown in Figure 3. Here, values of an attribute "age" are transformed into intervals with decreasing precision over increasing levels of generalization. Values of the attribute “sex“ can only be suppressed. We note that assigning generalization level zero to an attribute leaves its values unchanged. In ARX, generalization hierarchies can be specified by the user or created

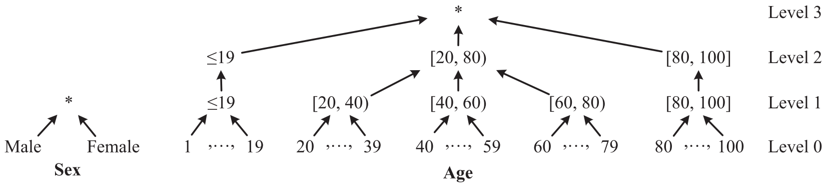
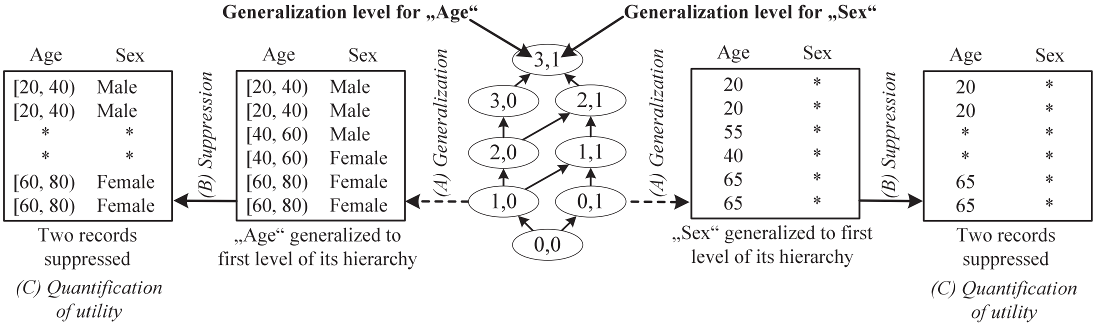


FIGURE 3 Examples of domain generalization hierarchies. The hierarchy to the left specifies possible generalizations of values for the attributes sex and the hierarchy to the right specifies generalizations for the attribute age

automatically for categorical and continuous attributes. In the latter case, this is accomplished by specifying functions for performing on-the-fly categorization of the value domain (eg, creating a grouping of heights or weights).

With full-domain generalization, all values of an attribute are transformed to the same generalization level in all records.[11] The set of all possible combinations of generalization levels for all attributes forms a *generalization lattice*, where each element is called a *generalization scheme*. The generalization lattice for the example hierarchies from Figure 3 together with an example dataset to which various generalization schemes have been applied is shown in Figure 4. Each node represents a single generalization scheme, which defines generalization levels for all attributes in the dataset. An arrow between two schemes indicates that they differ by exactly one generalization level. The transformation (0,0) represents the original dataset whereas the transformation (3,1) represents the dataset which results from maximal generalization. Referring to the overview from Figure 1, the optimal scheme from the lattice can be determined by going through all schemes one-by-one. In each step, the generalization scheme is applied (Step A), all records that do not adhere to the privacy requirements are suppressed (Step B) and the utility of the resulting output dataset is calculated (Step C). In the end, the optimal solution (ie, the output dataset with the highest utility) is returned. In the example, the privacy requirement is k-anonymity with $k = 2$, which means that each record must be indistinguishable from at least one other record (see Section 2.3 for more details on privacy models). In both output datasets created through generalization, the records three and four violate the privacy requirement and thus they have to be suppressed. After this, output data utility is measured to enable selecting the optimal solution. A simple utility model would be the number of cells that have not been suppressed (ie, that have a value different from "*"). In this case, the output dataset on the left would have a utility of eight while the output dataset on the right would have a utility of four. In practice, more sophisticated utility models are typically used, as is described in Section 2.3.

Anonymization algorithms using full-domain generalization are among the oldest approaches that have been developed in the field. Well-known examples include globally-optimal algorithms, such as Incognito[48] or *OLA*[49] and heuristic algorithms for data of higher dimensionality, such as *DataFly*.[23] ARX implements its own algorithms, *Flash* and *Lightning*, that significantly outperform prior approaches in the low-dimensional[50] as well as the high-dimensional setting,[12] respectively. Both algorithms make heavy use of ARX's compressed in-memory data representation[47] and advanced pruning-strategies.[31] Moreover, ARX employs a specialized record-suppression strategy that enables the software to suppress individual records for a specific generalization scheme, even when the privacy model used can only be evaluated


Generalization level for „Age"
Generalization level for „Sex"
3,1
3,0
2,1
2,0
1,1
1,0
0,1
0,0
(B) Suppression
(A) Generalization
(A) Generalization
(B) Suppression
Two records suppressed
„Age" generalized to first level of its hierarchy
„Sex" generalized to first level of its hierarchy
Two records suppressed
(C) Quantification of utility
(C) Quantification of utility

Age	Sex
[20, 40)	Male
[20, 40)	Male
*	*
*	*
[60, 80)	Female
[60, 80)	Female

Age	Sex
[20, 40)	Male
[20, 40)	Male
[40, 60)	Male
[40, 60)	Female
[60, 80)	Female
[60, 80)	Female

Age	Sex
20	*
20	*
55	*
40	*
65	*
65	*

Age	Sex
20	*
20	*
*	*
*	*
65	*
65	*

FIGURE 4 Example of full-domain generalization. It shows a generalization lattice and the results of applying two generalization schemes to a dataset followed by the suppression of records that do not adhere to the privacy requirements

for the overall dataset[13] (an example is *average re-identification risk*; further privacy models supported by ARX will be described in Section 2.3).

An important advantage of this class of search-based algorithms is that they are generic, which means that a wide range of privacy and utility models can be plugged into the system. The most important downside is that they are very inflexible in terms of supported transformation schemes and global generalization does not adjust well to the multidimensional distribution of data. This typically results in significant reductions to the quality of output data.

2.2 | Implementing advanced transformation methods

To overcome these limitations, we developed an approach for using the scalable basic algorithms of ARX as building blocks for implementing a wider range of more flexible transformation models. The basic idea is to iteratively apply the full-domain generalization algorithm to different subsets of an input dataset, resulting in different generalization schemes being used for the different subsets.

Horizontal partitioning strategy: What is needed for this purpose, is a partitioning strategy that reduces the overall degree of generalization applied. Such a strategy can be constructed using the basic algorithms provided by the software as follows. ARX supports the specification of a *limit on the number of suppressed records*. Moreover, records that have been suppressed may either be considered when calculating the overall utility of a transformed output dataset or they may be *ignored entirely* (ie, when calculating data utility, suppressed records are considered to be unmodified). To automatically partition and anonymize a dataset with n records, users only need to specify a *limit on the maximal number of partitions* (p) that can be created. From this limit, the minimal number of records in each partition can be derived ($n_p = \frac{n}{p}$). As is illustrated in Figure 5, ARX then sets the suppression limit accordingly and anonymizes the dataset while ignoring the impact of record suppression on data utility. This process is then iteratively repeated for the records that have been suppressed in the previous step until less than n_p suppressed records remain.

Vertical partitioning strategy (ie, grouping or clustering): To also support data aggregation, we developed a clustering strategy that is also based upon ARX's core algorithms as follows. The basic idea is to use the generalization scheme computed in each iteration not to transform the dataset, but to determine the clusters of values that need to become indistinguishable. In a subsequent *postprocessing step*, attributes of records within these clusters are then made indistinguishable by applying *aggregation functions* to the values from the input dataset of selected attributes (hence, vertical partitions) within each cluster (returning, eg, the mean or dynamic intervals). Vertical partitioning is performed automatically by ARX for attributes for which the user has configured aggregate functions. Further details on the horizontal as well as the vertical partitioning strategy, including pseudocode and examples, are provided in Appendices A and B.

As a result of the implementation of these two partitioning approaches, the software now supports combinations of four different types of transformation methods, which are listed in Table 1. With the new horizontal partitioning strategy, ARX can be configured to apply the same transformation scheme to all records in a dataset (*full-domain generalization*) or to apply different transformation schemes to different subsets of the records (*multi-dimensional generalization*).[51] The maximal number of transformations that may be used can be specified. ARX always guarantees that identical records in the input dataset will be transformed identically. With the new vertical partitioning strategy, hierarchies can also be

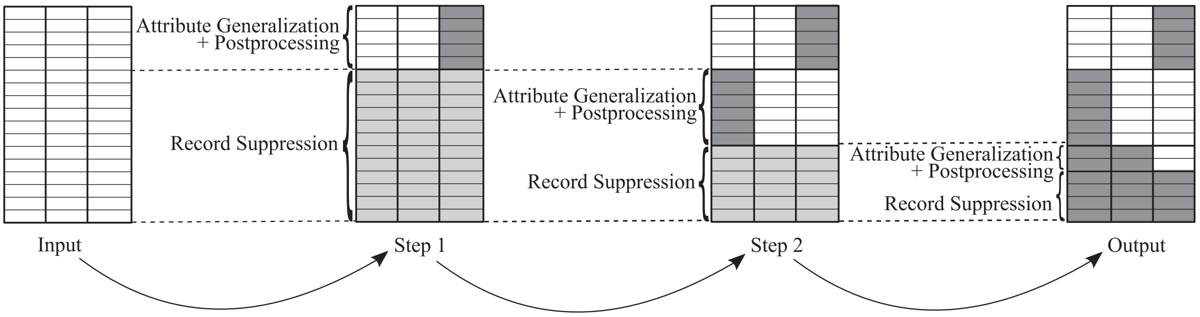

Attribute Generalization + Postprocessing
Record Suppression
Attribute Generalization + Postprocessing
Record Suppression
Attribute Generalization + Postprocessing
Record Suppression
Input
Step 1
Step 2
Output

FIGURE 5 Recursive application of the core transformation process for horizontal partitioning. ARX is able to apply full-domain generalization of attribute values followed by record suppression recursively to different subsets of a dataset

TABLE 1 Overview of transformation models supported by ARX

Transformation model		Type of attribute		Supported in prior versions
Type	**Implementation**	**Categorical**	**Numeric**	
Generalization	Multi-dimensional generalization	✓	✓	—
	Full-domain generalization	✓	✓	✓
	Top- and bottom-coding	—	✓	✓
	Categorization	—	✓	✓
Suppression	Cell-level	✓	✓	–
	Attribute-level	✓	✓	✓
	Record-level	✓	✓	✓
Sampling	Random	✓	✓	✓
	By query	✓	✓	✓
Microaggregation	Arithmetic and geometric mean	—	✓	–
	Median and mode	✓	✓	–
	Set	✓	✓	–
	Interval	—	✓	–

used to form clusters in which sets of attribute values can then be transformed into a common value by user-specified aggregation functions. Here, we have implemented support for the arithmetic and geometric mean, intervals, sets as well as median and mode for numerical attributes and sets, median and mode for categorical attributes. With the addition of the two partitioning schemes, we were able to extend ARX with six new transformation methods.

If transformation rules have been specified that only enable a suppression of values, a global transformation process will result in *attribute suppression*, while a local transformation process will result in a cell suppression scheme.[52] Independently of the specific transformation models specified, ARX may return a solution in which some of the records have been suppressed (typically only a tiny fraction). Generalization hierarchies can also be represented as functions, which can be used to perform on-the-fly categorization of continuous attributes during anonymization. Top- and bottom-coding can be implemented by using hierarchies that truncate values exceeding a user-specified range. We note that ARX contains multiple methods and wizards to automatically or semi-automatically construct hierarchies to apply these transformation methods. Finally, ARX supports drawing a sample from the input dataset. Methods that can be used for this purpose include matching a dataset against another dataset, querying the dataset using an expressive query language and random sampling. This can be used to relate a dataset to an underlying population table or to reduce privacy risks. Random sampling is further used to introduce randomness into the differential privacy mechanism supported by the software (see next section).

2.3 | Compatibility of methods

ARX supports a wide range of privacy and utility models. In this section, we discuss their compatibility with the horizontal and vertical partitioning strategies integrated into the software. The use of horizontal partitioning requires that privacy models can be enforced independently on different subsets of the data and that utility can be estimated by calculating it independently for different subsets. The use of vertical partitioning requires utility to be estimated accordingly.

ARX implements a wide range of privacy models that address different threats, such as *membership disclosure*, *attribute disclosure*, and *identity disclosure*.[11] Moreover, the privacy models address different assumptions about the intent and background knowledge of adversaries, such as the *prosecutor model*, *journalist model*, and the *marketer model*.[67] *Syntactic models* enforce restrictions on the structure of data, *statistical models* estimate risks in relationship to a larger underlying population or the success probabilities of attacks while *semantic models* have more direct relationships to mathematical notions of privacy. An overview of the models supported by ARX is shown in Table 2. Many models are supported in different variants.

An overview of the compatibility of the privacy models supported by ARX with different transformation techniques is provided in Table 3. Most incompatibilities are due to the way in which sampling is used in the software to implement privacy models. The method for taking a sample of the dataset is used to implement differential privacy, to specify population tables and to implement the horizontal partitioning algorithm. Consequently, privacy models that use sampling can currently not be combined with local transformation models. This is one of the shortcomings of the current development stage of the software that we plan to address in future work (see Section 6). Moreover, we note that in some cases it is also not obvious whether the privacy guarantees specified by a model also hold when data are partitioned. We have formally proven this for most models, but not yet for population uniqueness. For this reason, local transformation is currently deactivated for this model in the software. The current version of the differential privacy algorithm implemented in ARX is not compatible with the horizontal or vertical partitioning methods, as carefully randomized partitioning schemes would be required to ensure that privacy is not violated.[44]

In ARX, many different data utility models can be used as optimization functions. As is shown in Table 4, the software supports *general-purpose models*, which can be utilized when it is unknown in advance how output data will be used, and *special-purpose* (or *workload-aware*) models which quantify the usefulness of data for specific applications.[11] Utility models typically estimate data utility by quantifying the amount of information loss, for example, by measuring differences or similarities between the input and the output dataset. Models can roughly be classified as measuring information loss on the *attribute-level*, *cell-level*, *record-level*, or *dataset-level*. Typical examples for changes on these levels are differences in the distributions of attribute values, reductions in the granularity of data, differences in the distinguishability of records, or changes to overall scores, such as the accuracy of prediction models trained on the data. Notably, its strong support of methods for building and evaluating prediction models makes ARX also one of the most comprehensive tools available for privacy-preserving data mining.

Table 5 outlines the compatibility of the utility models with the transformation techniques supported by ARX. Incompatibilities resulting from vertical partitioning arise when using microaggregration operators. During the anonymization process, utility is only estimated for affected cells based on generalization. Incompatibilities resulting from horizontal partitioning are due to the fact that the frequencies of values in the input and output dataset are only known for the partition that is currently being processed. We emphasize that all utility models supported by ARX can still be used with all transformation methods. The quantification of utility reported by the system may be slightly off, however.

TABLE 2 Overview of privacy models supported by ARX

Privacy model	Type	Disclosure model	Attacker model	Population table
δ-Presence[53]	Syntactic/statistical	Membership	Journalist	✓
k-Anonymity[54]	Syntactic/statistical	Identity	Prosecutor	—
Average risk[42]	Syntactic/statistical	Identity	Marketer	—
k-Map[54]	Syntactic/statistical	Identity	Journalist	✓
k-Map with frequency estimators[55,56]	Statistical	Identity	Journalist	—
Population uniqueness[57-60]	Statistical	Identity	Marketer	—
ℓ-Diversity[61,62]	Syntactic/statistical	Attribute	Prosecutor	—
t-Closeness[63]	Syntactic/statistical	Attribute	Prosecutor	—
δ-Disclosure privacy[64]	Syntactic/statistical	Attribute	Prosecutor	—
β-Likeness[65]	Syntactic/statistical	Attribute	Prosecutor	—
Game-theoretic model (prosecutor)[43,66]	Semantic	Identity	Prosecutor	—
Game-theoretic model (journalist)[43,66]	Semantic	Identity	Journalist	✓
(ϵ, δ)-Differential privacy[44]	Semantic	All	All	—

TABLE 3 Compatibility matrix of privacy models and transformation models in ARX

Privacy model	Multi-dimensional generalization	Full-domain generalization	Top- and bottom-coding	Categorization	Cell-level suppression	Attribute-level suppression	Record-level suppression	Random sampling	Sampling by query	Microaggregation (all functions)
δ-Presence	—	✓	✓	✓	—	✓	✓	—	—	✓
k-Anonymity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Average risk	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
k-Map	—	✓	✓	✓	—	✓	✓	—	—	✓
k-Map with frequency estimator	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Population uniqueness	(✓)	✓	✓	✓	(✓)	✓	✓	✓	✓	✓
ℓ-Diversity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
t-Closeness	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
δ-Disclosure privacy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
β-Likeness	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Game-theoretic model (prosecutor)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Game-theoretic model (journalist)	—	✓	✓	✓	—	✓	✓	—	—	✓
(ϵ, δ)-Differential privacy	—	✓	✓	✓	—	✓	✓	—	—	—

Brackets indicate functionality that is currently deactivated.

TABLE 4 Overview of utility models supported by ARX

Utility model	Type	Scope	Optimization	Visual analysis
Missings	Generic	Cell	✓	✓
Granularity/loss[68]	Generic	Cell	✓	✓
Precision[22]	Generic	Cell	✓	✓
Nonuniform entropy[69,70]	Generic	Attribute	✓	✓
Average distinguishability[51]	Generic	Record	✓	✓
Discernibility[32,49]	Generic	Record	✓	✓
Ambiguity[29]	Generic	Record	✓	✓
Record-level entropy[66]	Generic	Record	✓	✓
Sum of squared errors	Generic	Record	—	✓
Publisher benefit[43]	Special purpose	Record	✓	✓
Classification accuracy[45,68,71]	Special purpose	Datasets	✓	✓

3 | EXPERIMENTAL DESIGN

3.1 | Tools and algorithms

In previous work, we have already shown that ARX outperforms prior algorithms in terms of scalability and/or data utility when implementing global data transformation schemes.[12,44,50] In this article, we show that this is also true for local generalization schemes enabled by the horizontal and vertical partitioning strategies described in Section 2.1. For this purpose, we compare our tool to related software. Specifically, we focus on the following transformation schemes:

- *Multi-dimensional generalization*: Solves an anonymization problem by generalization. Values are transformed by replacing them with values from the provided hierarchies. Identical records will also be transformed identically.[51]
- *Local generalization*: Solves an anonymization problem by local generalization. Generalization can be performed without hierarchies, for example, by creating sets of values or intervals and identical records can be transformed differently.[51]

In our evaluation, we focus on tools that implement highly automated anonymization processes, analogously to ARX. Moreover, the privacy models implemented by these tools interpret datasets as population data describing one individual per record. When calculating frequencies, missing values are treated as an own category that only matches other missing values. As a baseline for evaluating the performance of multi-dimensional generalization, we used the well-known Mondrian algorithm[51] as implemented by the open source UTD Anonymization Toolbox (version 2012).[33] Following a top-down partitioning approach, Mondrian starts off with the trivial partition which contains all records of the dataset and keeps partitioning until no further partitions can be formed without violating the privacy requirements specified. As a baseline regarding local generalization, we used the authors' implementation of the algorithm proposed by Sánchez et al[38] (details can be found in the supplementary material of the article[72]). This approach interprets categorical attributes as integer-valued, clusters records based on their centroids and then forms groups of indistinguishable records in each cluster by replacing values with corresponding intervals. We note that the competing algorithms have specifically been designed for the respective data transformation schemes implemented, while ARX supports all of them in an integrated manner using a single algorithm. When implementing local generalization with ARX we employ an aggregate function to generalize values within clusters in the output dataset. Finally, we note that in all experiments attributes were either generalized by replacing them with values from a generalization hierarchy or by replacing them with intervals. In the experiments with local generalization, all attributes were interpreted as numbers, as this is the approach implemented by the algorithm by Sánchez et al[38] We note that this comes without loss of generality, as the dynamic forming of intervals over numbers representing categories is equivalent to the forming of sets containing the values encoded by the numbers contained in the interval.

TABLE 5 Compatibility matrix of utility models and transformation models in ARX

Utility model	Multi-dimensional generalization	Full-domain generalization	Top- and bottom-coding	Categorization	Cell-level suppression	Attribute-level suppression	Record-level suppression	Random sampling	Sampling by query	Microaggregation (all functions)
Missings	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Granularity/loss	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Precision	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Nonuniform entropy	(✓)	✓	✓	✓	(✓)	✓	✓	✓	✓	(✓)
Average distinguishability	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Discernibility	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ambiguity	(✓)	✓	✓	✓	(✓)	✓	✓	✓	✓	(✓)
Record-level entropy	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Sum of squared errors	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Publisher benefit	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Classification accuracy	(✓)	✓	✓	✓	(✓)	✓	✓	✓	✓	(✓)

Brackets indicate that output data utility can only be approximated.

3.2 | Datasets

We used six real-world datasets, most of which have already been utilized for evaluating previous work on data anonymization: (1) *US Census*, an excerpt from the 1994 census database, which serves as the de facto standard for evaluations of anonymization algorithms, (2) *Competition*, introduced in the KDD data mining competition in 1998, (3) *Crash Statistics*, NHTSA crash statistics from their Fatality Analysis Reporting System, (4) *Time Use Survey*, data from the American Time Use Survey, (5) *Health Interviews*, results from the Integrated Health Interview Series, and (6) *Community Survey*, responses to the American Community Survey, an ongoing survey conducted by the US Census bureau on demographic, social, and economic characteristics from randomly selected people living in the United States. The sizes of the datasets on disk range between 2.52 MB (US Census) and 107.56 MB (Health Interviews). To ensure compatibility with the algorithm by Sánchez et al and to simplify the distribution of data together with the source code used in the experiments, we performed dictionary encoding on all categorical attributes.[38] The datasets have different characteristics, which are listed in Table 6:

- *Dimensionality*, that is, the number of attributes. With 30 attributes the Community Survey dataset is of *high* dimensionality. All other datasets contain either eight or nine attributes and are of *medium* dimensionality.
- *Volume*, that is, the number of records. The datasets US Census, Competition, and Community Survey contain between 30 162 and 68 725 records and are of *low* volume. With a size of 100 937 and 539 253 records, respectively, the datasets Crash Statistics and Time Use Survey are of *medium* volume while Health Interviews is a *high* volume dataset comprising 1 193 504 records.
- *Identifiability*, which is based on the number of unique patterns of attribute values contained in the data. Each such combination has the potential to identify individuals in the dataset and thus the number of patterns can be used for risk estimation.[73] We have calculated the number of these so-called *minimal sample uniques* (MSUs) using the SUDA2 algorithm provided by sdcMicro, modified to print the number of MSUs identified. In addition to the overall number of MSUs per dataset we report the average number of MSUs per cell. The more MSUs the higher is the risk of re-identification and therefore identifiability. While Community Survey and Competition are of *high* and *medium* identifiability, respectively, all other datasets are of *low* identifiability.

For reference, further properties of the datasets are presented in Appendix C. As a rule of thumb, higher dimensionality, volume, or identifiability can be expected to increase execution times and decrease output data utility. We note that some of the evaluation datasets are samples from a larger population, which have been created using complex sampling designs. These aspects could be used to derive more exact risk estimates during data anonymization. The tools considered in our evaluation, however, only implement privacy models that make worst-case assumptions and they do not implement mechanisms for considering complex data structures. Hence, we did not include special variables, such as strata variables or sampling weights, into our evaluation datasets and assumed that all datasets describe one individual per record. We emphasize that this is a frequent assumption in many domains, for example, in medical research, which is also often made when comparing automated data anonymization procedures. Moreover, this approach allows for a fair comparison between the tools covered in this section. We will discuss its limitations in Section 6.

TABLE 6 Overview of the datasets and their complexity in terms of dimensionality, volume as well as identifiability

	Dimensionality		Volume		Identifiability		
Dataset	**Attributes**	**Complexity**	**Records**	**Complexity**	**MSUs**	**MSUs/cell**	**Complexity**
US Census	9	Medium	30 162	Low	62 809	0.23	Low
Competition	8	Medium	63 441	Low	791 475	1.56	Medium
Crash Statistics	8	Medium	100 937	Medium	175 271	0.22	Low
Time Use Survey	9	Medium	539 253	Medium	321 406	0.07	Low
Health Interviews	9	Medium	1 193 504	High	2 888 220	0.27	Low
Community Survey	30	High	68 725	Low	15 708 409	7.6	High

As a rule of thumb, higher degrees of complexity can be expected to increase execution times and decrease output data utility.
Abbreviation: MSU, minimal sample unique.

3.3 | Configuration and setup

When selecting privacy models to use in the evaluation, the individual methods supported by the tools and algorithms listed above must be considered. ARX supports all models presented in Table 2. The Mondrian algorithm from the UTD Anonymization Toolbox, however, only supports k-anonymity and the algorithm by Sánchez et al supports only k-anonymity and t-closeness. We therefore decided to present results for the k-anonymity privacy model, because it is the only model supported by all competitors. We are well aware of the weaknesses of k-anonymity and emphasize that ARX also supports multiple more recent models, as described in the previous sections.

Common parameterizations for k-anonymity used in the literature are $k = 2, 3, 5, 10$, which equal thresholds for prosecutor re-identification risk of 50%, 33%, 20% and 10%. We vary this parameter and the number of attributes that must be protected from linkage (the so-called *quasi-identifiers [QI]*) to study the effect of different risk thresholds and data dimensionality on output data utility as well as scalability. We note that increasing the number of quasi-identifiers is a simple way to significantly increase the number of anonymization problems studied and that it can also provide more detailed insights into the effect of data dimensionality on the algorithms' performance. When varying k we included all quasi-identifiers and when varying the number of quasi-identifiers we used $k = 5$. We evaluated the scalability of the different solutions by measuring elapsed real *execution times*. In order to obtain stable results, we calculated averages over multiple runs of each algorithm (the number of runs for each experiment was determined based on the stability of runtime measurements). For practical reasons, we introduced a hard time limit of 3600 seconds and runs that did not terminate within that time frame were cancelled.

To evaluate output data utility, we used a simple and intuitive general-purpose model, called *Granularity*, which measures the value-level precision of output data.[68] For reference, a formal definition is presented in Appendix D. All utility measurements have been normalized into a range of [0, 1], such that 100% represents an unmodified dataset, and 0% represents the a dataset from which all information has been removed. We note that general-purpose utility models have limitations regarding their ability to capture the usefulness of output data for specific application scenarios, for example, regression modeling. However, at the extreme points of general-purpose utility estimates, such models also provide a good indicator for the usefulness of data for specific applications. For example, a general-purpose utility of close to 100% indicates that almost no changes have been made to the data, which typically also corresponds with usefulness for performing concrete analyses. Analogously, a general-purpose utility of 50%, for example, indicates that significant changes have been made to the data, which typically also significantly impacts usefulness for specific applications.

The experiments were performed on a desktop machine equipped with a quad-core 3.2 GHz Intel Core i5 CPU running a 64-bit Windows NT kernel and a 32-bit JVM (1.8.0_202_x86). All tools tested leveraged only one of the CPU cores of the benchmark system. Our implementation of the benchmark and the datasets used are available online.[74]

4 | RESULTS OF EXPERIMENTS AND DISCUSSION

4.1 | Comparison with the UTD Anonymization Toolbox

Figure 6 shows the execution times measured when performing multidimensional generalization. We note that in some settings we were not able to process the datasets Crash Statistics, Health Interviews, and Community Survey with the implementation of the Mondrian algorithm from the UTD Anonymization Toolbox, since the application terminated with an error. In the figure, this is indicated by "x". Regarding the other setups, it can be seen that higher volume or identifiability resulted in higher execution time (Time Use Survey, Health Interviews). With ARX execution times increased with increasing privacy parameters, while with the UTD Anonymization Toolbox execution times decreased with increasing privacy protection. For processing the high-dimensional dataset, ARX needed not more than 1000 seconds, while all other datasets could be processed in not more than 100 seconds. The UTD Anonymization Toolbox needed significantly more time in all cases.

Figure 7 shows the data utility measured in the experiments. It can be seen that in all cases ARX returned output data to which almost no modifications had been made. The results show that data utility slightly decreased when the degree of privacy protection increased. When using the UTD Anonymization Toolbox, however, significant changes were made to input data, resulting in utility estimates as low as 60% in some cases. It can further be seen that with ARX output data utility decreased monotonically when the number of quasi-identifiers increased. This trend could generally also be observed for the UTD Anonymization Toolbox. Some instabilities could however be observed when processing the Time

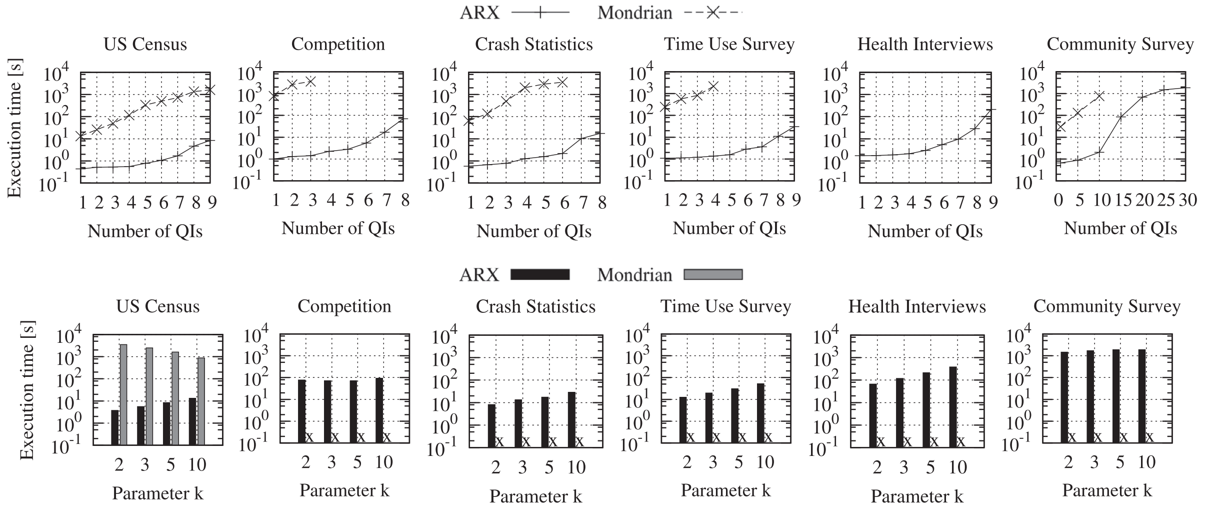


FIGURE 6 Comparison of the execution times of ARX and the UTD Anonymization Toolbox. Note: On *y*-axes, logarithmic scaling was used. In the bar charts, the symbol "x" indicates a missing data point due to the algorithm exceeding the time limit or terminating with an error

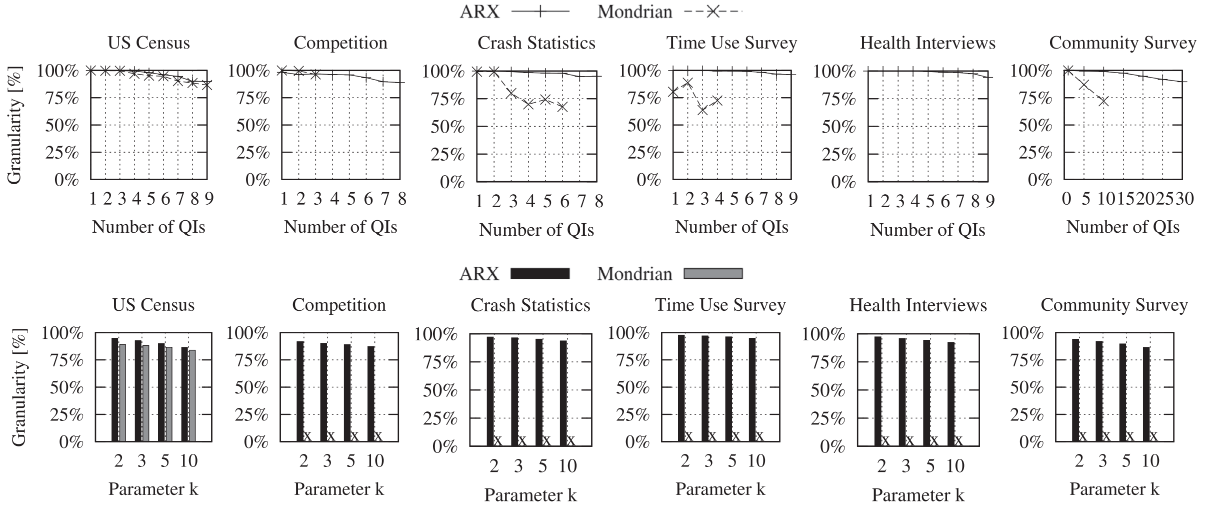


FIGURE 7 Comparison of the data utility obtained using ARX and the UTD Anonymization Toolbox. Note: On *y*-axes, logarithmic scaling was used. In the bar charts, the symbol "x" indicates a missing data point due to the algorithm exceeding the time limit or terminating with an error

Use Survey and Crash Statistics datasets. We note that the fact that ARX is able to significantly reduce the uniqueness of records in the Community Survey dataset with only about 10% reduction in data granularity implies that correlations exist between many of the attributes of the dataset. We conclude that, in our experiments with multidimensional generalization, the algorithm implemented by ARX exhibited significantly higher scalability than the Mondrian algorithm implemented by the UTD Anonymization Toolbox and at the same time provided higher degrees of output data utility.

4.2 | Comparison with the algorithm by Sánchez et al

Figure 8 shows the execution times measured when comparing ARX to the local generalization algorithm by Sánchez et al. It can be seen that ARX performed comparable to this algorithm in low-dimensional settings, while ARX performed

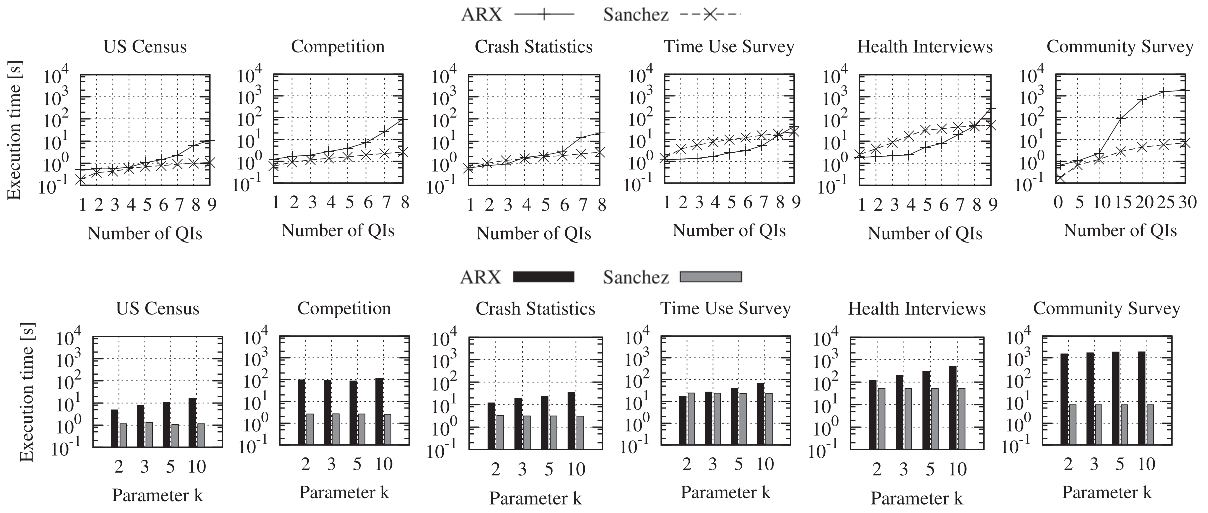


FIGURE 8 Comparison of the execution times of ARX and the algorithm by Sánchez et al. Note: On *y*-axes, logarithmic scaling was used

worse in high-dimensional settings. It can further be seen that ARX was less scalable when processing the dataset with high identifiability. The results also show that ARX outperformed the algorithm by Sánchez et al when processing the Time Use Survey dataset, which has the second highest volume of the datasets considered, but very low identifiability. This can be explained by the fact that the optimizations implemented into ARX are particularly effective when identifiability is low[31,50] and that the runtime complexity of the approach by Sánchez et al is dominated by sorting the dataset. This also implies that the performance of the algorithm by Sánchez et al mostly depends on the number of records contained in a dataset, which is also reflected by our results.

Figure 9 shows the data utility measured in the experiments. It can be seen that in all cases ARX returned output datasets to which almost no modifications had been made. When using the approach by Sánchez et al, however, significant changes were made to input data, again resulting in utility estimates as low as 60% in some cases. This is remarkable, as the transformation method implemented by ARX is less flexible, as it always guarantees that identical records in input data are transformed to identical records in the output dataset. Again, data utility decreased monotonically when risk

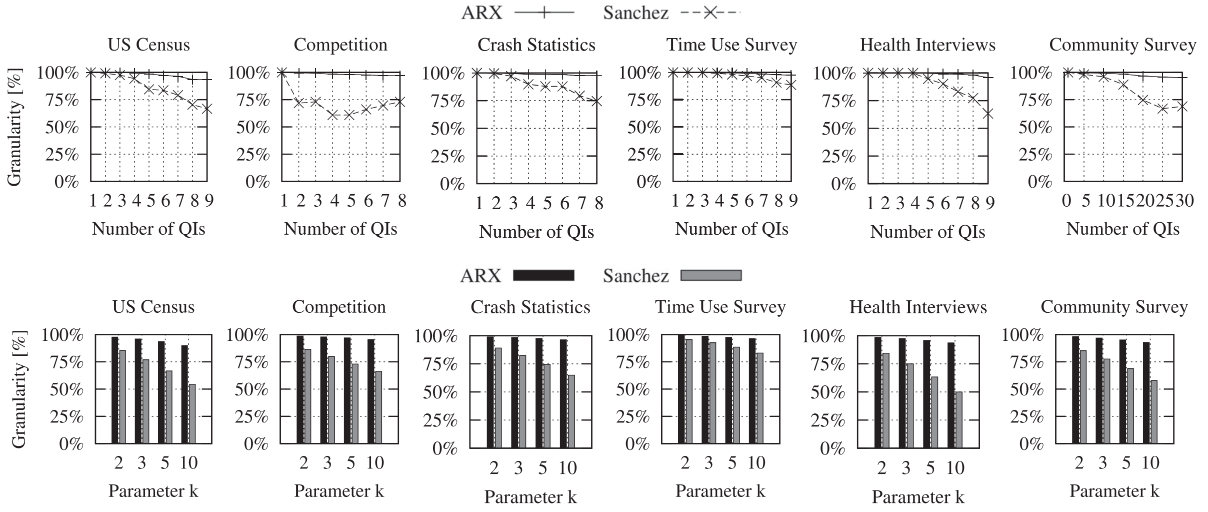


FIGURE 9 Comparison of the data utility obtained using ARX and the algorithm by Sánchez et al

thresholds increased but the effect was much stronger when using the algorithm by Sánchez et al. We conclude that, in our experiments with local generalization, the approach by Sánchez et al exhibited higher scalability but the algorithm implemented by ARX provided higher degrees of output data utility.

5 | SUMMARY AND PRACTICAL EXPERIENCES

In this article, we have presented an overview of the current development state of the ARX Data Anonymization Tool. We have described recent extensions to the software that enable users to utilize a wide variety of data transformation methods that were previously only supported by specific tools or algorithms. We have presented the results of an extensive experimental evaluation which has shown that ARX often outperforms related software. The development of methods that make ARX so flexible was not only a major methodological challenge, but it also contributed significantly to the success of the software. To illustrate this, we briefly present some examples of official policies and guidelines, research projects, and data publishing activities that have made use of the software.

On the level of guidelines, ARX has for example been mentioned by the European Medicines Agency as a solution for implementing quantitative risk assessments when implementing Policy 0070[75] on the sharing of data from clinical trials.[76] Moreover, ARX has been listed in a guideline by the European Union Agency for Network and Information Security (ENISA) on methods for implementing privacy and data protection by design principles.[77] Another guideline mentioning ARX has been released by the UK Anonymization Network (UKAN), which is an organization promoting and advising on best practices in data anonymization.[78] The document has also been adapted by the Office of the Australian Information Commissioner.[79] ARX has also been covered in a comprehensive analysis of anonymization tools released by the Directorate for Research, Studies, Evaluation and Statistics of the central administration of the French Ministry of Social Affairs and Health.[80] It has further been mentioned in a report on requirements and implementation options for anonymization services by the Finnish Ministry of Transport and Communications,[81] in a guide to data anonymization by the Personal Data Protection Commission of Singapore,[82] a security standard released by the Polish Ministry of Digitalization,[83] a report on data anonymization by the Dutch Ministry of Justice and Security[84] as well as a report by the Korean Ministry of Science and ICT.[85] These examples show the importance of open source anonymization tools for supervisory authorities.

On the level of scientific data management, various institutions have included ARX into software collections. Examples include the Finnish Social Science Data Archive,[86] EPFL,[87] the University of Guelph,[88] the University of Munich,[89] and the University of Kassel.[90] The graphical frontend of ARX is also frequently used in training courses. For example, the Korea Internet & Security Agency (KISA) and the TMF e.V., the umbrella organization for networked medical research in Germany, offer regular training programs.[91,92] ARX has further been covered in many handbooks on the topic.[40,93,94] Recently ARX has also been integrated into the big data processing framework KNIME,[95] and one of ARX's core algorithms has been selected to form the backbone of SAP HANA Data Anonymization.[96]

ARX has also been used in several research projects, mostly through its application programming interface. One important area is research on privacy-preserving big data analytics platforms. For example, Costa et al described a platform for big data management in the telecommunication sector that offers privacy-enhancing features through ARX.[97] Kim et al proposed a distributed analytics platform based on ensemble learning for healthcare data. They used data anonymized with ARX as a baseline in experimental comparisons.[98] A second line of research using ARX focuses on trust and access control. An interesting example is the article by Armando et al, which describes a risk-aware access control framework for information disclosure. The presented prototype includes a risk mitigation module which uses adaptive anonymization operations implemented on top of ARX.[99] Another example is the work by Jiang et al, in which game-theoretic methods have been used to develop a credibility model in cooperative networks and ARX has been included in the evaluation.[100] The development of new data anonymization methods is another area in which ARX is frequently utilized. An interesting example is the work by Stammler et al, who have used ARX to implement and evaluate an enhanced variant of the ℓ-diversity privacy model which uses an asymptotically unbiased estimator for the Shannon entropy.[62] Li et al have proposed and implemented a graph-based framework for privacy-preserving data publishing, which they evaluated by comparing the output of their framework with the output of ARX.[101] Moreover, Xu et al proposed a contract-based approach to handle the trade-off between privacy and utility, which has been implemented on top of ARX.[102] Finally, Park et al have developed a data synthesis mechanism based on Generative Adversarial Networks and they used ARX as a baseline technology in their evaluation.[103]

ARX has also been used to anonymize datasets for public and private dissamination. However, since official guidelines unfortunately do not usually provide specific instructions on how data needs to be anonymized, only little information is publicly available on practical applications. One example is the work by Kuzilek et al describing the Open University Learning Analytics Dataset, which is a representative subset of student data collected at the Open University. The data were anonymized using ARX in a process that has been certified by the Open Data Institute.[104] As another example, Ursin et al have used ARX to assess and manage the re-identification risk of a large dataset from the Norwegian Cervical Cancer Screening Program.[105]

6 | LIMITATIONS AND CHALLENGES AHEAD

ARX's flexibility and a relatively intuitive and easy-to-use interface are key factors that contributed to the software's success. However, we emphasize that the methods implemented by the software are complex from a mathematical and statistical perspective and, as a consequence, anonymization in real-world settings can usually only be carried out by experts. For example, risks models must be selected according to the context, and risks must then be reduced precisely to an extent that ensures that the data are reliably protected. In addition, one must be aware of the intended use of a dataset to ensure that the anonymized data remain useful. Moreover, there are several limitations that we plan to address in future work.

First, ARX does currently not support many methods provided by data anonymization tools from the statistics community, such as sdcMicro. Important examples include methods for considering the effect of complex sampling designs on re-identification risks when anonymizing data or different means of calculating the frequency of records for risk estimation. The main reason why we have not yet implemented such techniques is that they are not frequently used in the area of health data privacy, which is our primary application domain. However, we plan to extend the software in this direction in future work. Another area of future work is to compare ARX to other algorithms using transformation methods not studied in this article. Important examples include cell suppression and methods for aggregating continuous variables in such a way that they remain continuous and keep their scale of measure (eg, replacing them by the mean within clusters).

Second, while ARX is much more scalable than many other solutions in the field, it can currently only be used to anonymize medium-sized datasets with up to a few million rows and up to 50 quasi-identifying variables. Nowadays, data controllers often need to deal with gigabytes and terabytes of data, with in some cases hundreds of attributes that need to be protected. One example is large sparse datasets used for creating machine learning models.[106] Due to its high degree of automatization, ARX is well suited for implementing anonymization operators that can then be distributed amongst a large number of nodes to enable or speed up the processing of very large datasets. However, integrating appropriate strategies for distributing data and processing the results obtained from different nodes is challenging. This is particularly true for ARX, where parallelization strategies must be implemented carefully to not impact the flexibility of the software.

Another important area of future development is to improve ARX's abilities to process high-dimensional data along two axes. First, we plan to improve the scalability of finding solutions to anonymization problems with a high number of quasi-identifiers by implementing an alternative to the algorithm currently used by the software. The genetic algorithm proposed by Wan et al for anonymizing genetic data is an interesting candidate[66] but integrating it is challenging due to the different context in which it was proposed. Second, we plan to improve the utility of output data in high-dimensional settings by implementing methods to better handle complex inter-attribute relationships.[107] One possible solution to this problem is to treat the data as transactional, that is, set-valued, and to employ specific privacy models, such as k^m-anonymity,[108] which is implemented by Anamnesia[36] and SECRETA.[37]

Another related area with significant challenges ahead is to improve the compatibility of the privacy models implemented with local transformation methods. In this context, we plan to redesign our sampling subsystem to ensure that also models that rely on sampling can be used when applying local data transformation. Moreover, for some models, for example, those that use statistical models to estimate population uniqueness, it is not yet clear whether their privacy guarantees hold in the local transformation context. We plan to formally analyze this and to develop variants that can be used with local transformations if needed. These steps are also needed to guarantee privacy-preservation in the distributed settings outlined above. Finally, our differential privacy algorithm[44] needs to be extended with differentially private procedures which incorporate the horizontal and vertical partitioning methods.

We further plan to include more methods from the area of statistical disclosure control and further less formal transformation methods into ARX. An important example is the SUDA2 algorithm,[73] which can be used to implement various types of risk analysis and anonymization and which is frequently used in the statistics community. Furthermore, we plan to include data masking techniques (eg, for random data generation and shuffling) into the software to enable users to combine formal methods of data anonymization with a wide range of such basic transformation operations.

Finally, we are working on many features to make the software even more reliable and usable in practical applications. For example, we have recently integrated the data anonymization operations provided by ARX into the ETL environment

Pentaho Data Integration,[109] and we are working to integrate them into further environments, such as Talend Open Studio.[110] A significant challenge in this process is to not negatively impact the flexibility of the software.

ACKNOWLEDGEMENTS

The authors would like to thank the various contributors to ARX. They are particularly grateful to the contributions of Florian Kohlmayer, Fabian Stahnke, and Martin Waltl, who have contributed to the development and evaluation of the horizontal and vertical partitioning methods described in this article. Moreover, we would like to thank the anonymous reviewers for their comments and suggestions which have greatly improved this manuscript.

ORCID

Fabian Prasser https://orcid.org/0000-0003-3172-3095
Johanna Eicher https://orcid.org/0000-0003-4871-0282
Raffael Bild https://orcid.org/0000-0002-7398-5598

REFERENCES

1. Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques; 2014. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.
2. Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans Knowl Data Eng*. 2005;17(6):734-749.
3. Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. *MIS Q*. 2012;36(4):1165-1188.
4. US Department of Health and Human Services Office for Civil Rights. Standards for privacy of individually identifiable health information: final rule. *Fed Reg*. 2002;67(157):53181.
5. Council of the European Union, European Parliament. Regulation (EU) 2016/679 of the European parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46. *Off J Eur Union*. 2016;59(L119):1-88.
6. Standardization Administration of China. GB/T 35273-2017 information technology – personal information security specification; 2018.
7. Desai T, Ritchie F, Welpton R. *Five Safes: Designing Data Access for Research*. Bristol: University of the West of England; 2016.
8. Cramer R, Damgård IB, Nielsen JB. *Secure Multiparty Computation*. Cambridge: Cambridge University Press; 2015.
9. Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Min Knowl Disc*. 2005;11(2):195-212.
10. Xia W, Heatherly R, Ding X, Li J, Malin BA. R-U policy frontiers for health data de-identification. *J Am Med Inform Ass*. 2015;22(5):1029-1041.
11. Fung BCM, Wang K, Fu AW-C, Yu PS. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Boca Raton, FL: CRC Press; 2010.
12. Prasser F, Bild R, Eicher J, Spengler H, Kohlmayer F, Kuhn KA. Lightning: utility-driven anonymization of high-dimensional data. *Trans Data Priv*. 2016;9(2):161-185.
13. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. *IEEE J Biomed Health Inform*. 2018;22(2):611-622.
14. Leoni D. Non-interactive differential privacy: a survey. Paper presented at: Proceedings of the 1st International Workshop on Open Data; 2012:40-52.
15. El Emam K, Jonker E, Arbuckle L, Malin BA. A systematic review of re-identification attacks on health data. *PLoS One*. 2011;6(12):e28071.
16. Duncan GT, Elliot M, Salazar-González J-J. *Statistical Confidentiality: Principles and Practice*. New York, NY: Springer; 2011.
17. Narayanan Arvind, Shmatikov Vitaly. Robust de-anonymization of large sparse datasets. *Symposium on Security and Privacy*. Piscataway, NJ: IEEE; 2008;111–125.
18. Sweeney L. Computational disclosure control - a primer on data privacy protection (PhD thesis). Massachusetts Institute of Technology; 2001.
19. Hundepool A, Domingo-Ferrer J, Franconi L, et al. *Statistical Disclosure Control*. Hoboken, NJ: John Wiley & Sons; 2012.
20. McSherry FD Privacy integrated queries: an extensible platform for privacy-preserving data analysis. Paper presented at: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data; 2009:19-30.
21. Roy I, Setty STV, Kilzer A, Shmatikov V, Witchel E. Airavat: security and privacy for MapReduce. *NSDI*. 2010;10:297-312.
22. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncert Fuzz Knowl-Based Syst*. 2002;10(05):571-588.
23. Sweeney L. Datafly: a system for providing anonymity in medical data. *Database Security XI*. Boston, MA: Springer; 1998:356-381.
24. Babu KS, Reddy N, Kumar N, Elliot M, Jena SK. Achieving k-anonymity using improved greedy heuristics for very large relational databases. *Trans Data Priv*. 2013;6(1):1-17.
25. Soria-Comas J, Domingo-Ferrer J, Sánchez D, Martinez S. t-closeness through microaggregation: strict privacy with enhanced utility preservation. *IEEE Trans Knowl Data Eng*. 2015;27(11):3098-3110.

26. Byun JW, Kamra A, Bertino E, Li N. Efficient k-anonymization using clustering techniques. Paper presented at: Proceedings of the International Conference on Database Systems for Advanced Applications; 2007:188-200.
27. Gionis A, Mazza A, Tassa T. k-Anonymization revisited. Paper presented at: Proceedings of the 24th International Conference on Data Engineering; 2008:744-753.
28. Goldberger J, Tassa T. Efficient anonymizations with enhanced utility. Paper presented at: Proceedings of the International Conference on Data Mining; 2009:106-113.
29. Nergiz ME, Clifton C. Thoughts on k-anonymization. Paper presented at: Proceedings of the 22nd International Conference on Data Engineering; 2006:96.
30. Zakerzadeh H, Aggarwal CC, Barker K. Managing dimensionality in data privacy anonymization. *Knowl Inf Syst*. 2016;49(1):341-373.
31. Prasser F, Kohlmayer F, Kuhn KA. Efficient and effective pruning strategies for health data de-identification. *BMC Med Inform Decis Mak*. 2016;16(1):49.
32. Bayardo RJ, Agrawal R. Data privacy through optimal k-anonymization. Paper presented at: Proceedings of the 21st International Conference on Data Engineering; 2005:217-228.
33. UT Dallas Data Security and Privacy Lab. UTD anonymization toolbox; 2012. http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php.
34. Cornell Database Group. Cornell anonymization toolkit; 2014. https://sourceforge.net/projects/anony-toolkit/.
35. Dai C, Ghinita G, Bertino E, Byun J-W, Ninghui L. TIAMAT: a tool for interactive analysis of microdata anonymization techniques. *Proc VLDB Endow*. 2009;2(2):1618-1621.
36. OpenAIRE. Anamnesia; 2019. https://amnesia.openaire.eu/index.html.
37. Poulis Giorgos, Gkoulalas-Divanis Aris, Loukides Grigorios, Skiadopoulos Spiros, Tryfonopoulos C. SECRETA: a system for evaluating and comparing relational and transaction anonymization algorithms. Paper presented at: Proceeding of the 17th International Conference on Extending Database Technology; 2014:620-623.
38. Sánchez D, Martínez S, Domingo-Ferrer J. Comment on "Unique in the shopping mall: on the reidentifiability of credit card metadata". *Science*. 2016;351(6279):1274-1274.
39. Fung Benjamin C M. Selected publications; 2019. http://dmas.lab.mcgill.ca/fung/publicationsBySelection.htm.
40. Templ M. *Statistical Disclosure Control for Microdata: Methods and Applications in R*. Cham, Switzerland: Springer; 2017.
41. Hundepool A, Willenborg L. ARGUS: software packages for statistical disclosure control. In: Payne R, Green P, eds. *COMPSTAT*. Physica, Heidelberg; 1996;341-345.
42. Prasser F, Kohlmayer F, Kuhn KA. The importance of context: risk-based de-identification of biomedical data. *Methods Inf Med*. 2016;55(4):347-355.
43. Prasser F, Gaupp J, Wan Z, et al. An open source tool for game theoretic health data de-identification. Paper presented at: Proceedings of the AMIA Annual Symposium; 2017:1430-1439.
44. Bild R, Kuhn KA, Prasser F. SafePub: a truthful data anonymization algorithm with strong privacy guarantees. *Proc Priv Enhanc Technol*. 2018;2018(1):67-87.
45. Prasser F, Eicher J, Bild R, Spengler H, Kuhn KA. A tool for optimizing de-identified health data for use in statistical classification. Paper presented at: Proceedings of the 30th International Symposium on Computer-Based Medical Systems; 2017:169-174.
46. Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX - A comprehensive tool for anonymizing biomedical data. Paper presented at: Proceedings of the AMIA Annual Symposium; 2014:984-993.
47. Prasser F, Kohlmayer F. Putting statistical disclosure control into practice: the ARX data anonymization tool. *Medical Data Privacy Handbook*. Cham: Springer; 2015:111-148.
48. Le Fevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain k-anonymity. Paper presented at: Proceedings of the International Conference on Management of Data; 2005:49-60.
49. El Emam K, Dankar FK, Issa R, et al. A globally optimal k-anonymity method for the de-identification of health data. *J Am Med Inform Ass*. 2009;16(5):670-682.
50. Kohlmayer F, Prasser F, Eckert C, Kemper A, Kuhn KA. Flash: efficient, stable and optimal k-anonymity. Paper presented at: Proceedings of the International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing; 2012:708-717.
51. Le Fevre Kristen, De Witt David J, Ramakrishnan Raghu. Mondrian multidimensional k-anonymity. Proceedings of the 22nd International Conference on Data Engineering. 2006;:25–25.
52. Ohno-Machado L, Vinterbo S, Dreiseitl S. Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance. *J Am Med Inform Ass*. 2002;9(Suppl 6):S115-S119.
53. Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. Paper presented at: Proceedings of the International Conference on Management of Data; 2007:665-676.
54. Sweeney L. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 2002;10(05):557-570.
55. El Emam K, Dankar FK. Protecting privacy using k-anonymity. *J Am Med Inform Assoc*. 2008;15(5):627-637.
56. Pannekoek J. Statistical methods for some simple disclosure limitation rules. *Statistica Neerlandica*. 1999;53(1):55-67.
57. Chen G, Keller-McNulty S. Estimation of identification disclosure risk in microdata. *J Off Stat*. 1998;14(1):79.
58. Hoshino N. Applying Pitman's sampling formula to microdata disclosure risk assessment. *J Off Stat*. 2001;17(4):499-520.
59. Zayatz Laura Voshell. Estimation of the percent of unique population elements on a microdata file using the sample. Statistical Research Division Report Number: Census/SRD/RR-91/08; 1991.

60. Dankar F, El Emam K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. *BMC Med Inform Decis Mak*. 2012;12(1):66.
61. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: privacy beyond k-anonymity. Paper presented at: Proceedings of the 22nd International Conference on Data Engineering; 2006:24.
62. Stammler S, Katzenbeisser S, Hamacher K. Correcting finite sampling issues in entropy l-diversity. Paper presented at: Proceedings of the International Conference on Privacy in Statistical Databases; 2016:135-146.
63. Li N, Li T, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity. Paper presented at: Proceedings of the 23rd International Conference on Data Engineering; 2007:106-115.
64. Brickell J, Shmatikov V. The cost of privacy: destruction of data-mining utility in anonymized data publishing. Paper presented at: Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining; 2008:70-78.
65. Cao J, Karras P. Publishing microdata with a robust privacy guarantee. *Proc VLDB Endow*. 2012;5(11):1388-1399.
66. Zhiyu W, Yevgeniy V, Weiyi X, et al. A game theoretic framework for analyzing re-identification risk. *PLoS One*. 2015;10(3):e0120592.
67. El Emam K, Arbuckle L. *Anonymizing Health Data: Case Studies and Methods to Get You Started*. 1st ed. Sebastopol, CA: O'Reilly Media, Inc.; 2013.
68. Iyengar VS. Transforming data to satisfy privacy constraints. Paper presented at: Proceedings of the International Conference on Knowledge Discovery and Data Mining; 2002:279-288.
69. Gionis A, Tassa T. k-anonymization with minimal loss of information. Paper presented at: Proceedings of the European Symposium on Algorithms; 2007:439-450.
70. Prasser F, Bild R, Kuhn KA. A generic method for assessing the quality of de-identified health data. Paper presented at: Proceedings of the Medical Informatics Europe (MIE2016 @ HEC2016); 2016:312-316.
71. LeFevre K, DeWitt DJ, Ramakrishnan R. Workload-aware anonymization techniques for large-scale datasets. *ACM Trans Database Syst*. 2008;33(3):1-47.
72. Sánchez D, Martínez S, Domingo-Ferrer J. Supplementary materials for "How to avoid reidentification with proper anonymization" – comment on "Unique in the shopping mall: on the reidentifiability of credit card metadata". arXiv:1511.05957v22015.
73. Manning AM, Haglin DJ, Keane JA. A recursive search algorithm for statistical disclosure assessment. *Data Min Knowl Disc*. 2008;16(2):165-196.
74. A benchmark of different transformation models supported by ARX; 2019. https://github.com/arx-deidentifier/transformation-benchmark.
75. European Medicines Agency. EMA/240810/2013 - European Medicines Agency policy on publication of clinical data for medicinal products for human use; 2014. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.
76. European Medicines Agency. 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; 2018. https://www.ema.europa.eu/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-3.pdf.
77. European Union Agency for Network and Information Security. Privacy and data protection by design; 2015. https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design.
78. Elliot M, Mackey E, O'Hara K, Tudor C. *The anonymisation decision-making framework*. Manchester: UKAN; 2016.
79. Office of the Australian Information Commissioner. The de-identification decision-making framework; 2017. https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-decision-making-framework/.
80. Ministère des Solidarités et de la Santé. Données de santé: Anonymat et risque de ré-identification; 2015. https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/publications/les-dossiers-de-la-drees/dossiers-solidarite-et-sante/article/donnees-de-sante-anonymat-et-risque-de-re-identification.
81. Bäck Asta, Keränen Janne. Anonymisointipalvelut. Tarve ja toteutusvaihtoehdot Liikenne- ja viestintäministeriö; 2017. https://julkaisut.valtioneuvosto.fi/handle/10024/79579.
82. Personal Data Protection Commission of Singapore. Guide to basic data anonymisation techniques; 2018. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-_v1-(250118).pdf.
83. Polish Ministry of Digitalization. Open data - Security standard; 2018. https://dane.gov.pl/media/ckeditor/2018/11/06/security-standard_2018.odt.
84. Dutch Ministry of Justice and Security. On statistical disclosure control technologies; 2018. https://www.wodc.nl/binaries/Cahier2018-20_2889_Fulltext_tcm28-362210.pdf.
85. Ministry of Science and ICT. A research on de-identification technique for personal identifiable information; 2016. https://www.fsd.tuni.fi/aineistonhallinta/en/anonymisation-and-identifiers.html.
86. Finnish Social Science Data Archive. Data management guidelines: anonymisation and personal data; 2018. https://www.fsd.tuni.fi/aineistonhallinta/en/anonymisation-and-identifiers.html.
87. Research Data Library Team. RDM Walkthrough Guide. École polytechnique fédérale de Lausanne (EPFL) Bibliothèque. URL: https://www.epfl.ch/campus/library/wp-content/uploads/2019/09/RDM_Walkthrough_Guide_20190930.pdf.
88. University of Guelph. Clean and prepare your data; 2018. https://guides.lib.uoguelph.ca/CleanAndPrepareData/5.
89. LMU Munich. Conduct your study; 2019. https://www.osc.uni-muenchen.de/toolbox/resources_for_researchers/conduct_your_study/index.html.
90. University of Kassel. Management of research data; 2019. https://www.uni-kassel.de/themen/forschungsdatenmanagement/service-hilfe/faq.html.

91. Korea Internet & Security Agency. KISA promotes training on identification of personal information. https://www.kisa.or.kr/notice/press_View.jsp?mode=view&p_No=8&b_No=8&d_No=1570.
92. TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung. ANONTrain: Praktische Anwendung von Anonymisierungswerkzeugen. http://www.tmf-ev.de/Desktopmodules/Bring2Mind/DMX/Download.aspx?EntryId=28213&PortalId=0.
93. Domingo-Ferrer J, Sánchez D, Hajian S. Database privacy. *Privacy in a Digital, Networked World*. Basel, Switzerland: Springer; 2015.
94. Torra V. *Data privacy: Foundations, new developments and the big data challenge*. Basel, Switzerland: Springer; 2017.
95. Data Anonymization in KNIME. A redfield privacy extension walkthrough; 2019. https://www.knime.com/blog/data-anonymization-in-knime-a-redfield-privacy-extension-walkthrough.
96. Stephan K, Jens H, Johann-Christoph F. SAP HANA goes private: from privacy research to privacy aware enterprise analytics. *Proc VLDB Endow*. 2019;12(12):1998-2009.
97. Costa C, Chatzimilioudis G, Zeinalipour-Yazti D, Mokbel MF. Efficient exploration of telco big data with compression and decaying. Paper presented at: Proceedings of the 33rd International Conference on Data Engineering; 2017:1332-1343.
98. Kim J, Ha H, Chun B-G, Yoon S, Cha SK. Collaborative analytics for data silos. Paper presented at: Proceedings of the 32nd International Conference on Data Engineering; 2016:743-754.
99. Armando A, Bezzi M, Metoui N, Sabetta A. Risk-based privacy-aware information disclosure. *Int J Secur Softw Eng*. 2015;6(2):70-89.
100. Jiang C, Kuang L, Han Z, Ren Y, Hanzo L. Information credibility modeling in cooperative networks: equilibrium and mechanism design. *IEEE J Select Areas Commun*. 2017;35(2):432-448.
101. Li X-Y, Zhang C, Jung T, Qian J, Chen L. Graph-based privacy-preserving data publication. Paper presented at 35th International Conference on Computer Communications; 2016:1-9.
102. Xu L, Jiang C, Chen Y, Ren Y, Liu KJR. Privacy or utility in data collection? a contract theoretic approach. *IEEE J Select Topics Signal Process*. 2015;9(7):1256-1269.
103. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. *Proc VLDB Endow*. 2018;11(10):1071-1083.
104. Kuzilek J, Hlosta M, Zdrahal Z. Open university learning analytics dataset. *Scientific Data*. 2017;4:170171.
105. Ursin G, Sen S, Mottu J-M, Nygård M. Protecting privacy in large datasets: first we assess the risk; then we fuzzy the data. *Cancer Epidem Prevent Biomar*. 2017;26(8):1219-1224.
106. Domingos PM. A few useful things to know about machine learning. *Commun ACM*. 2012;55(10):78-87.
107. Aggarwal CC. On k-anonymity and the curse of dimensionality. Paper presented at: Proceedings of the 31st International Conference on Very Large Data Bases; 2005:901-909.
108. Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data. *Proc VLDB Endow*. 2008;1(1):115-125.
109. Prasser F, Spengler H, Bild R, Eicher J, Kuhn KA. Privacy-enhancing ETL-processes for biomedical data. *Int J Med Inform*. 2019;126:72-81.
110. Bowen J. *Getting Started with Talend Open Studio for Data Integration*. Birmingham: Packt Publishing Ltd; 2012.

How to cite this article: Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—Current status and challenges ahead. *Softw Pract Exper*. 2020;50:1277–1304. https://doi.org/10.1002/spe.2812

APPENDIX A PSEUDOCODE OF THE ALGORITHM

The core of the flexible transformation process described in this article is a routine which performs full-domain attribute generalization followed by record suppression and value aggregation (the latter is also called vertical partitioning). This process is sketched in Figure A1. The suppression limit `s` is used to specify that not more than `s` records may be suppressed. The process of optimal full-domain generalization followed by record suppression and aggregation is encapsulated in the call to the method `generalizeAndAggregate`. This method is not described in further detail, as the underlying algorithms Flash and Lightning have been covered in previous publications.[12,50] The only difference to the original algorithms is that the effect of record suppression is ignored when calculating the utility of the output produced by the available generalization schemes.

Figure A2 illustrates how the the method `transformRecords` is being applied to subsets of the records from the input dataset to implement the horizontal partitioning strategy. The pseudocode is formulated iteratively rather than recursively for ease of understanding.

In line 7, full-domain generalization is performed on the dataset `d`, resulting in the dataset `t`. `t` may contain records, which either have been transformed (ie, values generalized or aggregated) or which have been suppressed. The original versions of suppressed records are then extracted from `t` via the method `extractSuppressionCandidates` in line 8.

```
Dataset transformRecords(Dataset d, PrivacyParameter p, Integer s)
{
  Arx arx = new Arx();
  arx.addPrivacyModel(new PrivacyModel(p));
  arx.setSuppressionLimit(s);
  return arx.generalizeAndAggregate(d);
}
```

FIGURE A1 Pseudocode illustrating the method `transformRecords` [Colour figure can be viewed at wileyonlinelibrary.com]

```
 1  Dataset transform(Dataset d, PrivacyParameter p, Integer partitions)
 2  {
 3    Dataset result = {};
 4    for (Integer remaining = partitions; remaining > 0; remaining--)
 5    {
 6      Integer s = |d| - |d| / remaining;
 7      Dataset t = transformRecords(d, p, s);
 8      d = extractSuppressionCandidates(t);
 9      if (|d| == |t|) { // If all records have been suppressed
10        result = union(result, t);
11        return result;
12      }
13      Dataset a = extractTransformedRecords(t);
14      result = result ⊠ a;
15    }
16  }
```

FIGURE A2 Pseudocode illustrating the anonymization method [Colour figure can be viewed at wileyonlinelibrary.com]

These records are then processed in the next iteration if the termination condition (line 9) is not met. In line 13, the method `extractTransformedRecords` returns all records which have been subject to attribute generalization or aggregation. These are then added to the intermediate result (line 14). The parameter `partitions` (also called *p* in Section refsec:advanced) determines the maximal number of iterations. In each iteration, the suppression limit used when calling `transformRecords` is calculated appropriately in line 6 to guarantee that the condition in line 9 is satisfied within at most `partitions` iterations. The choice of `partitions` balances execution times against data quality.

APPENDIX B EXAMPLE ILLUSTRATING THE APPROACH

In this section, we provide an example illustrating an application of our algorithm. In this process, we use the example dataset from Figure 4, which we have extended with two additional attributes *height* and *income*. Domain generalization

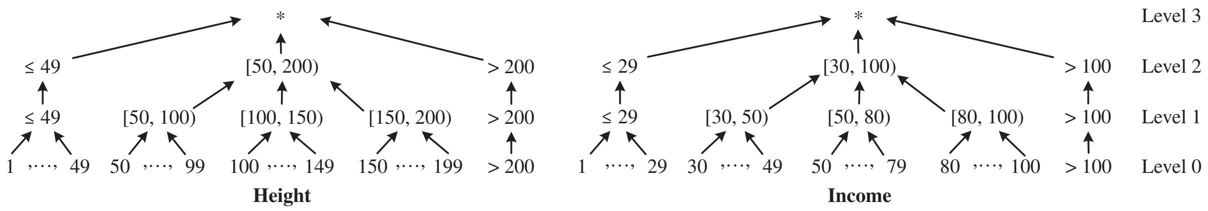
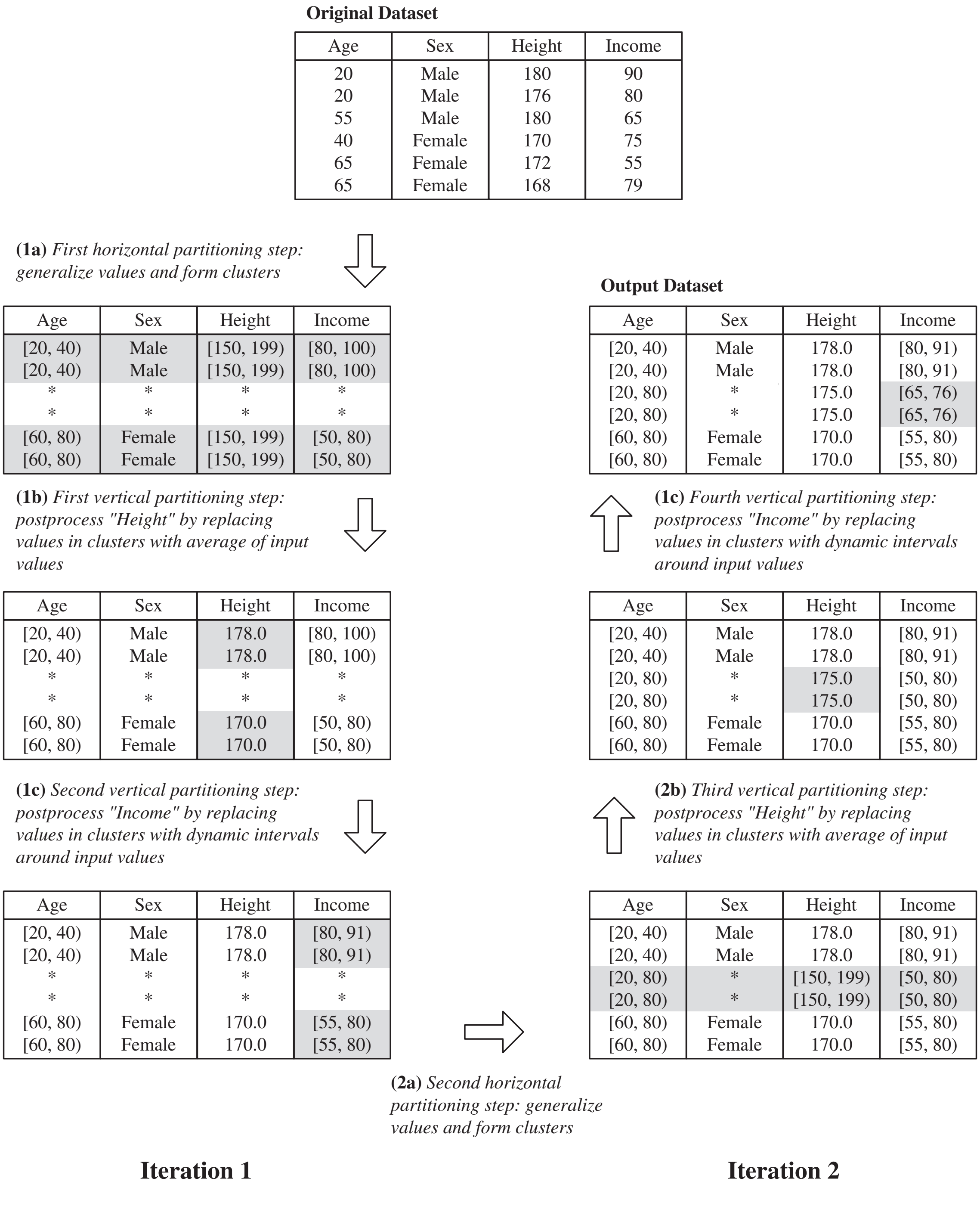


FIGURE B1 Domain generalization hierarchies for the additional attributes. The hierarchy to the left specifies possible generalizations of values of the attribute height and the hierarchy to the right specifies possible generalizations of values of the attribute income

hierarchies for the attributes *age* and *sex* have been provided in Figure 3. Hierarchies for the additional attributes are presented in Figure B1 below.

Figure B2 shows the original dataset as well as all steps executed to generate a 2-anonymous output dataset by applying local generalization to the attributes *age* and *sex* as well as aggregating *height* by replacing values with the arithmetic mean and aggregating *income* by generating dynamic intervals around values in each cluster. The algorithm terminates after two iterations, where each iteration consists of three steps. Cells transformed in each step are highlighted in grey.



Original Dataset

Age	Sex	Height	Income
20	Male	180	90
20	Male	176	80
55	Male	180	65
40	Female	170	75
65	Female	172	55
65	Female	168	79

(1a) *First horizontal partitioning step: generalize values and form clusters*

Age	Sex	Height	Income
[20, 40)	Male	[150, 199)	[80, 100)
[20, 40)	Male	[150, 199)	[80, 100)
*	*	*	*
*	*	*	*
[60, 80)	Female	[150, 199)	[50, 80)
[60, 80)	Female	[150, 199)	[50, 80)

(1b) *First vertical partitioning step: postprocess "Height" by replacing values in clusters with average of input values*

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 100)
[20, 40)	Male	178.0	[80, 100)
*	*	*	*
*	*	*	*
[60, 80)	Female	170.0	[50, 80)
[60, 80)	Female	170.0	[50, 80)

(1c) *Second vertical partitioning step: postprocess "Income" by replacing values in clusters with dynamic intervals around input values*

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 91)
[20, 40)	Male	178.0	[80, 91)
*	*	*	*
*	*	*	*
[60, 80)	Female	170.0	[55, 80)
[60, 80)	Female	170.0	[55, 80)

(2a) *Second horizontal partitioning step: generalize values and form clusters*

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 91)
[20, 40)	Male	178.0	[80, 91)
[20, 80)	*	[150, 199)	[50, 80)
[20, 80)	*	[150, 199)	[50, 80)
[60, 80)	Female	170.0	[55, 80)
[60, 80)	Female	170.0	[55, 80)

(2b) *Third vertical partitioning step: postprocess "Height" by replacing values in clusters with average of input values*

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 91)
[20, 40)	Male	178.0	[80, 91)
[20, 80)	*	175.0	[50, 80)
[20, 80)	*	175.0	[50, 80)
[60, 80)	Female	170.0	[55, 80)
[60, 80)	Female	170.0	[55, 80)

(1c) *Fourth vertical partitioning step: postprocess "Income" by replacing values in clusters with dynamic intervals around input values*

Output Dataset

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 91)
[20, 40)	Male	178.0	[80, 91)
[20, 80)	*	175.0	[65, 76)
[20, 80)	*	175.0	[65, 76)
[60, 80)	Female	170.0	[55, 80)
[60, 80)	Female	170.0	[55, 80)

Iteration 1

Iteration 2

FIGURE B2 Example illustrating the partitioning strategies. Cells transformed in each step are highlighted in grey

- In step (1a), the first horizontal partitioning step, the dataset is generalized and clusters are formed. To this end, a generalization scheme is applied to the original dataset resulting in three clusters each containing two records. The second cluster contains two suppressed records. These will be transformed in the next iteration. The attributes *age* and *sex* are transformed using the generalization hierarchies.
- In step (1b), the first vertical partitioning step, the attribute height is aggregated by replacing the values in each cluster with the average of the associated values from the input dataset.
- Finally, in the last step of the first iteration, (1c), which constitutes the second vertical partitioning step, the attribute *income* is aggregated by replacing the values in each cluster with dynamic intervals around the associated input values.

In the second iteration, the same process is repeated in steps (2a), (2b), and (2c) for the two records suppressed in the first iteration, resulting in the final output dataset.

APPENDIX C SPECIFICATION OF THE DATASETS USED IN THE EXPERIMENTS

In this appendix, we present more details about the datasets used in the experiments. We note that we used the datasets to compare the performance (in terms of scalability and output data utility) of different anonymization algorithms to each other and not to perform case studies using a specific anonymization algorithm. The properties of the datasets which are most important for this comparison (ie, volume, dimensionality, uniqueness of data) are listed in Table 6. For reference, we list further details about the attributes of the datasets in this section. We note that in practice the selection of quasi-identifiers needs to be performed in a context-specific manner considering additional safeguards such as access restrictions (see Section 6). Analogously to many other studies using the same or similar datasets, we therefore simply selected a set of privacy-relevant attributes for each dataset to perform the comparison. As can be seen in the following paragraphs, the selected attributes included demographics (eg, age, marital status, sex), social parameters (eg, education, insurance coverage), financial data (eg, income), and health parameters (eg, weight, health problems). Finally, we note that all datasets are also available in our online repository.[74]

Table C1 presents a list of the attributes of the "US Census" dataset, which comprises eight categorical attributes and one numeric attribute. The heights of the generalization hierarchies used for anonymization varied between two and five. The dataset contains an excerpt from the 1994 US census database from which records containing "null" values have been removed. We note that this dataset is a de facto standard dataset for comparing anonymization algorithms and that we have removed records containing "null" values only to replicate the setup most commonly used, not because the algorithms studied are not able to handle missing data (see Section 3.1). Further information is available online: http://archive.ics.uci.edu/ml/datasets/adult.

Table C2 presents a list of the attributes of the "Competition" dataset. As can be seen, the dataset comprises two categorical and six numeric attributes. The heights of the generalization hierarchies used for anonymization varied between two and six. The dataset originates from the 1998 KDD data mining competition. Further information is available online: http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html.

Table C3 presents a list of the attributes of the "Crash statistics" dataset, which comprises seven categorical attributes and one numeric attribute. The heights of the generalization hierarchies used for anonymization varied between two and six. We note that the attributes "ideathmon" and "ideathday" are categorical, because the dataset contains special categories for missing values ("not applicable" and "unknown"). The dataset originates from the Fatality Analysis Reporting System (FARS) of the US National Highway Traffic Safety Administration (NHTSA) and can be accessed here: ftp://ftp.nhtsa.dot.gov/FARS/.

Table C4 presents a list of the attributes of the "Time Use Survey" dataset, which comprises eight categorical attributes and one numeric attribute. The heights of the generalization hierarchies used varied between two and six. The dataset originates from the American Time Use Survey. Further information is available online: http://atusdata.org/index.shtml.

Table C5 presents a list of the attributes of the "Health Interviews" dataset. As can be seen, the dataset comprises five categorical and four numeric attributes. The heights of the generalization hierarchies used for anonymization varied between two and six. The dataset originates from the US Integrated Health Interview Series. Further information is available online: https://nhis.ipums.org/nhis/.

Table C6 presents a list of the attributes of the "Community Survey" dataset, which comprises 27 categorical and three numeric attributes. The heights of the generalization hierarchies used for anonymization varied between two and five.

TABLE C1 Specification of the "US Census" dataset

Attribute	Data type	Distinct values	Hierarchy height
sex	Categorical	2	2
age	Numeric	72	5
race	Categorical	5	2
marital-status	Categorical	7	3
education	Categorical	16	4
native-country	Categorical	41	3
workclass	Categorical	7	3
occupation	Categorical	14	3
salary-class	Categorical	2	2

The table presents a list of the attributes contained in the dataset, which consists of 30 162 records.

TABLE C2 Specification of the "Competition" dataset

Attribute	Data type	Distinct values	Hierarchy height
ZIP	Numeric	13 294	6
AGE	Numeric	94	5
GENDER	Categorical	6	2
INCOME	Numeric	7	3
STATE	Categorical	53	2
RAMNTALL	Numeric	814	5
NGIFTALL	Numeric	81	5
MINRAMNT	Numeric	58	5

The table presents a list of the attributes contained in the dataset, which consists of 63 441 records.

TABLE C3 Specification of the "Crash Statistics" dataset

Attribute	Data type	Distinct values	Hierarchy height
iage	Numeric	99	6
irace	Categorical	20	3
ideathmon	Categorical	14	4
ideathday	Categorical	33	4
isex	Categorical	3	2
ihispanic	Categorical	10	3
istatenum	Categorical	51	4
iinjury	Categorical	8	3

The table presents a list of the attributes contained in the dataset, which consists of 100 937 records.

TABLE C4 Specification of the "Time Use Survey" dataset

Attribute	Data type	Distinct values	Hierarchy height
Region	Categorical	4	3
Age	Numeric	83	6
Sex	Categorical	3	2
Race	Categorical	23	3
Marital status	Categorical	7	3
Citizenship status	Categorical	6	3
Birthplace	Categorical	155	3
Highest level of school completed	Categorical	18	4
Labor force status	Categorical	6	3

The table presents a list of the attributes contained in the dataset, which consists of 539 253 records.

TABLE C5 Specification of the "Health Interviews" dataset

Attribute	Data type	Distinct values	Hierarchy height
YEAR	Numeric	13	6
QUARTER	Numeric	4	3
REGION	Categorical	4	3
PERNUM	Numeric	25	4
AGE	Numeric	86	5
MARSTAT	Categorical	10	3
SEX	Categorical	2	2
RACEA	Categorical	16	2
EDUC	Categorical	26	2

The table presents a list of the attributes contained in the dataset, which consists of 1 193 504 records.

TABLE C6 Specification of the "Community Survey" dataset

Attribute	Data type	Distinct values	Hierarchy height
Insurance purchased	Categorical	2	2
Workclass	Categorical	10	3
Divorced	Categorical	3	2
Income	Numeric	464	5
Sex	Categorical	2	2
Mobility	Categorical	4	2
Military service	Categorical	5	2
Self-care	Categorical	3	2
Grade level	Categorical	17	3
Married	Categorical	3	2
Education	Categorical	25	4
Widowed	Categorical	3	2
Cognitive	Categorical	3	2
Insurance Medicaid	Categorical	2	2
Ambulatory	Categorical	3	2
Living with grandchildren	Categorical	3	2
Age	Numeric	93	4
Insurance employer	Categorical	2	2
Citizenship	Categorical	5	3
Indian Health Service	Categorical	2	2
Independent living	Categorical	3	2
Weight	Numeric	561	5
Insurance Medicare	Categorical	2	2
Hearing	Categorical	2	2
Marital status	Categorical	5	3
Vision	Categorical	2	2
Insurance Veteran's Association	Categorical	2	2
Relationship	Categorical	18	3
Insurance Tricare	Categorical	2	2
Childbirth	Categorical	3	2

The table presents a list of the attributes contained in the dataset, which consists of 68 725 records.

The dataset contains the data collected in the state of Massachusetts during the year 2013 as responses to the American Community Survey (ACS), an ongoing survey conducted by the US Census Bureau on demographic, social and economic characteristics from randomly selected people living in the US. Further information is available online: https://www.census.gov/programs-surveys/acs/.

APPENDIX D DEFINITION OF THE UTILITY MODEL USED IN THE EXPERIMENTS

In this appendix, we present a formal definition of the utility model used in the experiments. We denote the number of records in the dataset with n and the number of attributes in the dataset with m. The "granularity" model is a general-purpose utility measure based on the "loss" model proposed by Iyengar.[68] It is defined as:

$$1 - \frac{1}{m} \sum_{1 \leq x \leq m} \text{loss}(x), \tag{D1}$$

where $\text{loss}(x) \in [0, 1]$ returns the information loss for attribute x.

The information loss for an attribute x, denoted by $\text{loss}(x) \in [0, 1]$, is defined as the average information loss over all values of this attribute in the dataset:

$$\text{loss}(x) = \frac{1}{n} \sum_{1 \leq y \leq n} \text{loss}(x, y). \tag{D2}$$

The information loss per value, denoted by $\text{loss}(x, y) \in [0, 1]$, is calculated depending on the type of the attribute x and the transformation applied to the attribute:

1. For categorical and numeric attributes transformed using an associated generalization hierarchy, information loss per cell is defined as:

$$\text{loss}(x, y) = \frac{\text{leafs}(x, \text{value}(x, y)) - 1}{\text{leafs}(x, \text{root}(x)) - 1}, \tag{D3}$$

   where $\text{value}(x, y)$ returns the value of attribute x in record y, $\text{root}(x)$ returns the value of the root node of the generalization hierarchy for attribute x and $\text{leafs}(x, v)$ returns the number of leaf nodes rooted at the value v in the hierarchy of attribute x.
2. For numeric attributes which have been transformed into intervals (either by using a generalization hierarchy in which inner nodes represent intervals or by dynamic aggregation into intervals), information loss per cell is defined as:

$$\text{loss}(x, y) = \frac{|\text{upper}(\text{value}(x, y)) - \text{lower}(\text{value}(x, y))|}{|\max(x) - \min(x)|}, \tag{D4}$$

   where $\text{value}(x, y)$ returns the value of attribute x in record y, $\text{lower}(v)$ returns the lower bound of the interval described by value v, $\text{upper}(v)$ returns the upper bound of the interval described by value v, $\min(x)$ returns the smallest value of attribute x in the input dataset and $\max(x)$ returns the largest value of attribute x in the input dataset.
3. For values which have been suppressed, information loss is defined as:

$$\text{loss}(x, y) = 1, \text{if value}(x, y) \text{ is suppressed}, \tag{D5}$$

   which equals the information loss measured for attribute values which have been completely generalized or transformed into an intervals covering the complete domain of a numeric attribute.
4. For all other values, information loss is defined as:

$$\text{loss}(x, y) = 0, \text{in all other cases}, \tag{D6}$$

which implies that the model is not able to capture changes to data utility caused by other types of transformation, for example, by aggregating numeric values by replacing them with their mean.

We note that the model has been implemented in a manner that takes care of a wide range of edge cases. For example, it is made sure that no division by zero occurs should the domain of a variable consist of only one value and it is considered whether upper or lower bounds of intervals are inclusive or exclusive should the domain of a variable consist of integer values only. In summary, the model returns values in the range $[0, 1]$, where the original dataset has a utility of 100% and a transformed dataset in which all attribute values have been removed (either by generalization, suppression or by replacing them with intervals covering the complete domain of the attribute) has a utility of 0%.

Finally, we note that the fact that this model is not able to capture changes to data utility caused by aggregation operators other than the forming of dynamic intervals (eg, operators which replace values with their mean) is not relevant for the experiments presented in this article. The reason is that we only used generalization, suppression and replacement by dynamic intervals as other transformation operators are not supported by the tools to which we compared our software. ARX does, however, support further utility models such as the sum of squared errors, which can be used to analyze the impact of further types of aggregation (see Section 2.3).