A Quantitative Analysis of the Performance and Scalability of De-identification Tools for Medical Data

Zhiming Liu, Nafees Qamar, and Jie Qian

United Nations University
International Institute for Software Technology
Macau SAR China
{lzm,nqamar,qj}@iist.unu.edu

Abstract. Recent developments in data de-identification technologies offer sophisticated solutions to protect medical data when, especially the data is to be provided for secondary purposes such as clinical or biomedical research. So as to determine to what degree an approach– along with its tool– is usable and effective, this paper takes into consideration a number of *de-identification* tools that aim at reducing the *re-identification* risk for the published medical data, yet preserving its statistical meanings. We therefore evaluate the residual risk of re-identification by conducting an experimental evaluation of the most stable research-based tools, as applied to our Electronic Health Records (EHRs) database, to assess which tool exhibits better performance with different quasi-identifiers. Our evaluation criteria are quantitative as opposed to other descriptive and qualitative assessments. We notice that on comparing individual disclosure risk and information loss of each published data, the μ-Argus tool performs better. Also, the generalization method is considerably better than the suppression method in terms of reducing risk and avoiding information loss. We also find that sdcMicro has the best scalability among its counterparts, as has been observed experimentally on a virtual data consisted of 33 variables and 10,000 records.

1 Introduction

Interoperable electronic health records are one of the current trends characterizing and empowering the most recent Health Information Systems (HISs). With the advent of EHR standardizations, for instance, HL7 and *open*EHR [8], sharable EHR systems are now at the edge of practice. Notably, huge amount of patients' EHRs is being stored, processed, and transmitted across several healthcare platforms and among clinical researchers for online diagnosis services and other clinical research. Alternatively, the secondary use of de-identified data could be for instance in health system planning, public health surveillance, and generation of de-identified data for system testing [6]. However, if EHRs are directly made available to the public (i.e., without applying a de-identification technique), occurrences of serious data confidentiality issues are very likely to

J. Gibbons and W. MacCaull (Eds.): FHIES 2013, LNCS 8315, pp. 274–289, 2014.

occur. In reality, hospitals have confidential agreements with patients, which strictly forbid them not to disclose any identifiable information on individuals. Further to that, laws such as HIPAA [5] explicitly state the confidentiality protection on health information, where any sharable EHR system must legally comply with.

De-identification is defined [5] as a technology to remove the identifiable information such as name, and SSN from the published dataset so that the medical data may not be re-identified, even if it is being offered for a secondary use. Specifically, it is meant to deal with data privacy challenges by protecting the data under a maximum tolerable disclosure risk while still preserving the data of an acceptable quality.

One naive approach on confidentiality protection of patient's data is to remove any identifiable information (i.e., patient's name, SSN, etc.) of an EHR. However, adversary can still re-identify a patient by inferring from external information. A research [17] indicates that 87 percent of the population of U.S. can be distinguished by sex, date of birth and zip code. Such a combination of attributes, which can uniquely identify an individual, is defined as quasi-identifiers. More specifically, we can define quasi-identifiers as the background information about one or more people in the dataset. If an adversary has knowledge of these quasi-identifiers, it makes it possible to recognizing an individual and taking the advantage of his clinical data. On the other hand, we can find out most of these quasi-identifiers have statistical meanings in clinical researches. Thus, there exists a paradox between reducing the likelihood of disclosure risk and retaining the data quality. For instance, if any information of patient's residence were excluded from the EHR, it would disable related clinical partners to catch the spread of a disease. Conversely, releasing data including total information of patient's residence, sex and date of birth would bring a higher disclosure risk.

In recent years, several typical privacy criteria (i.e., k-anonymity [19], l-diversity [13], and t-closeness [11]) and anonymization methods (e.g., generalization, suppression, etc.) have been proposed. A detailed description on the formal definition of anonymity can be found in [16]. Based on these endeavors, a number of research-based de-identification tools (i.e., CAT, μ-Argus, and sdcMicro) now exist that offer data anonymization services to avoiding with the disclosure risks of patients' original data and other legal pitfalls. Each tool has its sample demonstration and even some of these have been applied on real datasets [21]. Nonetheless, these methods and tools lack in providing a sufficient evidence of their adoptability as well as usability. Thus, an experimental evaluation is dearly needed that could provide a systemic and directly usable analyses of these tools. The study should also allow choosing the most appropriate tool for de-identifying healthcare data such that any healthcare organization could know the efficacy of a tool before opting for it. To the extent of our knowledge, our conducted study is the first work that finds answers to such questions by examining the characteristics of a de-identification tool with respect to its ability to minimizing data closure risks and avoiding the distortion of results which is as much important as the de-identification process itself.

We propose an experiment on our EHR database to evaluate the performance and effectiveness of each de-identification tool. Then we find the most suitable tool for releasing EHRs by judging its capability of minimizing data disclosure risk and the distortion of de-identified data. To ensure meaningful quantitative analyses, we successfully borrowed a dataset from a local dialysis center in Macau SAR China, consisting of 1000 electronic health records. This moderate-size dataset could provide necessary quasi-identifiers for finding the possibilities of linking back an entry to the original patient even after applying a de-identification tool. Some contemporary and partly similar works include [6] and [9] that also evaluate such tools. However, most of these efforts target the technical details of the internal functioning and anonymization processes and methods such as [6], instead of providing insights on the usability and effectiveness of tools against re-identifiability of a published medical dataset. Another study [7] also summarizes some anonymization techniques. It discusses operations, metric and optimality principles of recent anonymization algorithms, and shows weakness of these algorithms through examples of different attack models. However, it does not provide a comparison of these techniques by means of quantitative analyses, or a criterion to follow, in order to find a best anonymization solution for a certain type of data.

Organization: In Section 2, we briefly introduce the experimented tools. Section 3 analyses the EHR database and then lists the potential quasi-identifiers. Section 4 introduces the design of experiment. Section 5 presents the results of our experiments. We also discuss the limitations of this study in Section 6. Section 7 draws some important conclusions and lists future directions.

2 State-of-the-Art Tools for Data De-identification

A number of research groups [14][20][22] are actively developing their de-identification tools, aiming to enabling their users to have more confidence in publicly-published dataset. They have adopted different approaches that reflect their particular interests and expertise. However, all these tools include a similar anonymization process in which a privacy criterion can be iteratively approximated. In this paper we include the following most stable tools.

CAT (Cornell Anonymization Kit). [22] is developed by a database group at Cornell University. This tool anonymizes data using generalization, which is proposed by [1] as a method that specifically replaces values of quasi-identifiers into value ranges. This tool also provides graphical user interface, which eases users' operations like adjusting parameters of a privacy criterion or checking current disclosure risk. Users can apply anonymization process iteratively until they obtain a satisfactory result. To ensure privacy criterion, users have to delete unsafe data manually. Therefore, there is no optimal principle implemented in this tool. In terms of usability, this tool presents contingency tables and density graphs between original and anonymous data, which implicitly offers users an

Table 1. Featuring the three de-identification tools

Tools	**Input Data**	**Privacy Criterion**	**Anonymization Approach**	**Data Evaluation**
CAT	Meta and microdata	l-diversity, t-closeness	Generalization	Comparison, Risk analysis
μ-Argus	Meta and microdata	k-anonymity	Global recoding, Local Suppression, etc.	Risk analysis
sdcMicro	Database	k-anonymity	Global recoding, Local Suppression, etc.	Comparison, Risk analysis

intuitive way to learn the information loss that caused during a de-identification process.

***μ*-Argus.** [14] is part of the CASC project `http://neon.vb.cbs.nl/casc/`, which is partly sponsored by the European Union. μ-Argus is an acronym for Anti-Re-identification General Utility System. This tool is based on a view of safety and unsafety of microdata that is used at Statistics Netherlands, which means the rules it applies to protect data comes from practice rather than the precise form of rules. Besides handling the specific requirements of Statistics Netherlands, this tool also implements general methods for producing safe data. In particular, it supports de-identification approaches such as global recoding, local suppression, top and bottom coding, the Post RAndomisation Method (PARM), aggregation, swapping, synthetic data and record linkage, which enable a variety of selections to enhancing data security against some foreseeable re-identification risks. Users are allowed to apply their strategies through a graphical user interface and make adjustments upon an observation of the re-identifiable risk of the results. Privacy criterion is guaranteed by an automatic mechanism, in which unsafe variables in record are removed.

sdcMicro. [20] is developed by Statistics Austria based on *R* as a highly extensive system for statistical computing. Since *R* is an open platform, it offers a facility for designing and writing functions for particular research purposes. Like μ-Argus, this tool implements several anonymization methods considering different types of variables. Users are able to try out several settings of global recording method iteratively, while have a detailed look at each step of the anonymization. Since anonymization process is applied via scripts, all the steps can easily be reproduced. In addition, this tool provides functions for the measurement of disclosure risk and the data utility for numerical data.

Table 1 illustrates a preliminary summary of the similarities and differences of these tools, allowing an security specialist to have a better intuition of the techniques behind their automations.

3 Electronic Health Records of Patients

From an ongoing collaborative work with the Kiang Wu Hospital Dialysis Center Macau SAR China, we have implemented a software system for capitalizing on its electronic health records. Our acquired test database consists of 1000 EHR samples in which a total of 183 variables have been recorded.

De-identifying such a moderate-size dataset is considerably challenging since the de-identified data would always have a chance of re-identification attack, if published. Suppose that while responding to an organization's request asking for a published dataset on patients' infectious disease histories, the corresponding quasi-identifiers (already known by an intruder) can indirectly cause a disclosure of patient's information. An adversary could determine one of the quasi-identifiers referenced to a female born on 12/04/64, sent to Kiang Wu Hospital Dialysis Center last Friday, and living in Taipa (Macau) is exactly his neighbor. Then he could find out that his neighbor has an infectious disease history of HCV (i.e., an acronym for Hepatitis C). Even though the likelihood of such a scenario is relatively difficult but yet possible with or without using the automated re-identification attacks. For VVIP personalities such a leakage can bring about far more catastrophic results than for general public records.

Preliminaries: Here, we consider a subset of the combination of the following variables in the database: Gender, Date of Birth, Place of Birth, Province of Residence, and Zip Code as a set of quasi-identifiers. From now on we use the following abbreviations:

QID = quasi-identifier, ZC = zip code, DoB = date of birth,
YoB = year of birth, DoR = district of residence, PoB = place of birth

Given a quasi-identifier, a set of records which have the same values of this quasi-identifier is defined as an anonymity set; the number of distinct values of this quasi-identifier in the database indicates the number of anonymity sets; the number of patients who share a specific value of this quasi-identifier represents the anonymity set size k. Here, anonymity is the state of being not identifiable within a set of subjects, the anonymity set. We choose quartiles as a means of indicating the value distribution of the anonymity set size for each quasi-identifier.

Table 2 shows the statistical characteristics of anonymity set size k for various quasi-identifiers. The second column indicates the number of anonymity sets in our database for a given quasi-identifier. Generally, during the de-identification process, the larger the number of distinct anonymity sets, the less information distortion on the published dataset because the anonymity set tends to be smaller in that case and removing one affects only little on the overall dataset. The min and max values denote the size of smallest and size of largest anonymity set.

According to Table 2, it is clear that some quasi-identifiers lead to particularly high disclosure risks, because more than half of their anonymity sets are smaller than 2, which means a large portion of patients can be unambiguously identifiable by that quasi-identifier. For instance, for {ZC+DoB}, we can find that 'k=1' is

Table 2. Anonymity set size k for various quasi-identifiers

Quasi-identifiers	**Numbers of sets**	**Min.**	**1st Qu.**	**Median**	**3rd Qu.**	**Max**
ZC	38	9	20	25	31	51
ZC+gender	76	2	10	13	16	30
ZC+DoB	997	1	1	1	1	2
ZC+YoB	659	1	1	1	2	5
ZC+PoB	280	1	1	1	2	37
ZC+gender+YoB	804	1	1	1	1	4
ZC+gender+PoB	341	1	1	1	2	22
gender+DoB	998	1	1	1	1	2
gender+YoB	70	5	10	13	19	38
gender+DoR	14	55	62	72	77	91
gender+PoB	44	2	5	7	9	369
gender+DoR+PoB	191	1	1	2	2	67
gender+PoB+YoB	336	1	1	1	2	31
gender+DoR+YoB	398	1	1	2	3	11
gender+DoR+PoB+YoB	638	1	1	1	2	9

up to the 3rd quartile, which means at least 75 percents of the patients are unambiguously identifiable by zip code and date of birth. Also, some quasi-identifiers are weaker because their smallest anonymity set is more than 5, such as {ZC}, {gender+DoR} and {gender+YoB}. Overall, it turns out that quasi-identifier that contains date of birth, place of birth and year of birth are most identifiable.

We also found that the size of anonymity sets for which quasi-identifiers contain 'place of birth' has a significant increase between the third quartile and max value. It means that a relatively large group of patients converge to one characteristic. This is because most of the patients of Kiang Wu Hospital Dialysis Center are Macau citizens. Consequently, patients who were born elsewhere are of sparse distribution and more likely to be unambiguously identifiable by their {gender+PoB} or {ZC+PoB}. Table 2 also clearly shows that year of birth, a reduction of date of birth, increases the de-identifiability: the median anonymity set size for {gender+YoB} is 13, whereas for {gender+DoB} is only 1.

Table 3 shows the actual number of patients that belongs to those anonymity sets, for example, for {ZC+DoB}, only two patients can be found in anonymity sets that have $k \leq 5$. The larger the value in the columns 'k=1' and '$k \leq 5$', the larger the portion of the patients that is covered by anonymity sets of small sizes, and the stronger the quasi-identifier identify patients. The number indicates that {ZC+DoB} is the strongest quasi-identifier, because almost all patients have k=1. However, zip code alone is a weaker quasi-identifier, because none of patients is in the first two columns.

Similarly, {gender+DoB} is a very strong quasi-identifier mainly because date of birth poses a significant privacy risk for nearly all the patients in our database. In this experiment, we replaced date of birth to year of birth before the experiment.

The numbers for {ZC+gender+YoB} indicates that 63.7 percent of the patients can be unambiguously identified by this quasi-identifier.
For {gender+DoR+PoB+YoB}, it shows that nearly half of the patients can be unambiguously identified.

Table 3. Number of EHR data per anonymity set size, for various quasi-identifiers

Quasi-identifiers	k**=1**	k**≤5**	k**≤10**	k**≤50**
ZC	0	0	9	949
ZC+gender	0	2	179	1000
ZC+DoB	994	1000	1000	1000
ZC+YoB	418	1000	1000	1000
ZC+PoB	199	294	309	1000
ZC+gender+YoB	637	1000	1000	1000
ZC+gender+PoB	237	333	664	1000
gender+DoB	994	1000	1000	1000
gender+YoB	0	10	188	1000
gender+DoR	0	0	0	0
gender+PoB	0	57	242	294
gender+DoR+PoB	90	294	304	542
gender+PoB+YoB	240	354	575	1000
gender+DoR+YoB	134	864	989	1000
gender+DoR+PoB+YoB	435	958	1000	1000

4 The Assessment Criteria

In order to assess the performance and effectiveness of the listed de-identification tools with our EHR database, we design our experiment of the following four main aspects.

1. Selection of Quasi-identifiers. Judging from Table 3, we found {ZC+gender +YoB} (denoted as QID^1) and {gender+DoR+PoB+YoB} (denoted as QID^2) are the most representative quasi-identifiers for this database (note that we excluded the quasi-identifiers that contained date of birth).

2. Selection of Privacy Criteria. Our comparison on the tools is independent on the parameter of selected privacy criteria. Specifically, the factors we think that affect the performance of a tool are its optimization algorithms and approaches. To ease our comparison, we provided k-anonymity for this dataset. In this experiment, we set the parameter k to 2, which means the minimum value of anonymity set size that is safe for QID^1 and QID^2.

3. Dimensions of Comparison. Three dimensions of comparison are identified. The first dimension is the individual disclosure risk of the published datasets regarding the above quasi-identifiers. An accurate measure in terms of the individual risk on a quasi-identifier was defined as the following formula [10].

$$\xi = \frac{1}{n} \sum_{k=1}^{K} f_k r_k \tag{1}$$

For a quasi-identifier, f_k denotes the size of $k - th$ anonymity set of the database; r_k denotes the probability of re-identification of a $k - th$ anonymity set; K depends on which k-anonymity to be preserved (for 2-anonymity K=2); n denotes the total number of the records. A higher number indicates that the published database undergoes a higher probability of disclosing patient's privacy. Generally, individual disclosure risk is related to the threshold value. Suppose that a threshold r^* has been set on the individual risk (see formula (1)), unsafe records are those for which $r_k \leq r^*$. When threshold value is set to 0.5, it ensures the dataset to achieve 2-anonymity. Similarly, when it is 0.2, it requires the dataset to achieve 5-anonymity.

The second dimension is the information loss for the published datasets. A strict evaluation of information loss must be based on a comparison between original dataset and published dataset. A metric called *Prec* has been proposed by Sweeny [18]. For each quasi-identifier, *Prec* counts the ratio of the practical height applied to the total height of the generalization hierarchy. Consequently, the more the variables are generalized, the higher the information loss. However, *Prec* has been criticized not considering the size of the generalized cells. Also, it does not account for the information loss caused by suppression method. Another commonly used metric is DM* [3], which addresses on the weakness of *Prec.* But it has also been criticized by [12] because it does not give intuitive results when the distributions of the variables are non-uniform. Therefore, these two metrics are not suitable for this experiment.

As described in Section 3, for a quasi-identifier QID, the distribution of its anonymity set size (denoted as F (QID)) should be equivalent to the distribution of the values of variables in QID. If some of these values are modified in the anonymization process, it will have an impact on F (QID). Such an impact depends on the frequency of the modified values in total values. Therefore, it is feasible to calculate the information loss of anonymization by comparing F (QID) of original data and published data. Looking into Table 2, it is clear that the anonymity set size of {gender+PoB} (denoted as QID^3) and {gender+DoR+PoB} (denoted as QID^4) has a significant increase between the third and forth quartile than other quasi-identifiers. In other words, the individual disclosure risk has a significant decrease in the third and fourth quartile because the larger the anonymity set size, the safer the published data is. To simplify the results, we measure the information loss in terms of the slope of anonymity set size for each QID^3 and QID^4 in the third and fourth quartile.

The information loss for a quasi-identifier is:

$$\lambda = \frac{\frac{\partial R'}{\partial Num'}}{\frac{\partial R}{\partial Num}} = \frac{\frac{k(n+1)' - k(n)'}{Sum'}}{\frac{k(n+1) - k(n)}{Sum}} \tag{2}$$

Where k(n) represents the anonymity set size at the n-th quartile of the original dataset, its primed version k(n)' is the result of the published data; Sum represents the number of distinct anonymity sets in the dataset, its primed version Sum' is the number after de-identification. The above formula usually yields a positive value. A higher number suggests a higher information loss of the original dataset.

The third dimension is the scalability of these tools. A virtual dataset consisted of 33 variables and 10,000 records are used as a test case, which includes 4 numeric variables, 3 categorical ones, and the rest are plain-text. We construct this synthetic dataset by enlarging the sample dataset of μ-argus from 4,000 to 10,000 records, of which the additional 6,000 records are copies that randomly selected from the original dataset. Using this relatively large database, we evaluate the ability of these tools to deal with large data sets.

4. Principal Methods Used for De-identification. Although different methods for acquiring k-anonymity criterion have been implemented in these tools, we present here a broad classification depending on the main techniques used to de-identify quasi-identifiers. Specifically, we classify anonymization methods in two categories as follows: Generalization and Suppression, as proposed in [2]. Other methods, which randomly replace the values of quasi-identifiers (e.g., adding noise), distort the individual data in ways that often result in incorrect clinical inferences. As these methods tend to have a low acceptance among clinical researchers, we decided not to apply them to the EHR database.

Generalization: provides a feasible solution to achieving k-anonymity by transforming the values in a variable to the optimized value ranges referencing to the user-defined hierarchies. Particularly, global recording means that generalization is performed on the quasi-identifiers across all of the records, which ensures all the records have the same recoding for each variable.

Suppression: means the removal of values from data. There are three general approaches to suppression: case-wise deletion, quasi-identifier removal, and local cell suppression, where CAT applies the first approach; μ-argus and the sdcMicro applied the third approach. For the same affected number of records, casewise deletion always has a higher degree of distortion on the dataset than local cell suppression. In most case, suppression leads to less information loss than generalization because the former affects single records whereas the latter affects all the records in the dataset. However, the negative effect of missing values should be considered.

5 Experimental Results

Before starting our experiment, we indexed our EHR database into microdata and metadata. For instance, we mapped 7 identifiable variables in PatientRecord table to categorical variables, 1 to numerical variables, 9 to string variables and removed 18 variables that were either illegal to release (i.e., patient's name, SSN) or irrelevant to research purpose (i.e., time stamp, barcode). We also truncated the value of date of birth variable into year of birth.

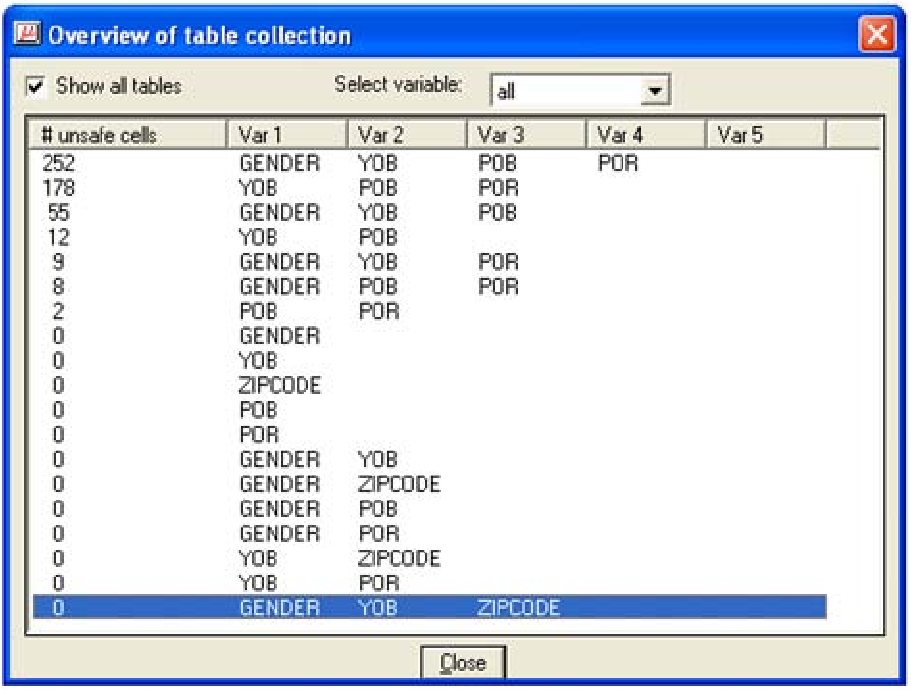
We started by anonymizing our dataset using μ-Argus. First, we specified the combination of variables to be inspected as QID^1 and QID^2 with the threshold set to 1 (maximum value of anonymity set size k, which is considered unsafe). It should be noted that the individual risk model was restricted in μ-Argus because there was an overlap between the quasi-identifiers. Then the tool counted the number of the unsafe records that are unambiguously identifiable for each combination of variables. By following its user's manual, the first anonymization method we applied was global recoding. Specifically, 22 different values in place of birth variable were equivalently generalized to 8 categories; 35 different values in year of birth variable were generalized to 12 categories; the last digit of zip code was removed. As shown in figure 1, the number of unsafe records decreased from 637 to 0 and 435 to 252, respectively for QID^1 and QID^2. It is clear that global recoding significantly decreases the risk of re-identification on QID^1. However, for QID^2, 252 out of 1000 patients remain to be unambiguously identifiable.

After dealing with categorical variables, we found that micro aggregation method was not practical, because the minimum frequency of the numeric variable is far above the minimum requirement for safe anonymity set size. Then we applied local suppression method to protect the remaining unsafe records. This led to 75 values in gender variables and 121 values in place of birth variables suppressed from the dataset.

In what follows we used sdcMicro. Due to the character encoding issue on ODBC, we collated our dataset from Traditional Chinese to UTF-8, which resulted in character loss on some of the values in place of birth and district of residence variables. Then we used freqCalc function in sdcMicro to calculate the number of unsafe records for QID^2. The result shows that 411 records could be unambiguously identified by QID^2, contrast to 435 in Table 3, which indicates an inaccuracy deviation of 5.5% on QID^2.

Similarly, we first applied the sdcMicro function globalRecode to the dataset. It turns out year of birth variable generalized to the same 12 categories, which reduced the number of unsafe records to 244 and 254, respectively for QID^1 and QID^2.

Then the function localSupp could be used to apply local suppression method. Using the threshold value of 0.5 (to achieve 2-anonymity as mentioned in Section IV), localSupp was first applied to QID^1. This led to a suppression of 244 values in zip code variable and 20 values in year of birth variable. Again, calculating the number of unsafe records for this quasi-identifier, we found that the published dataset reached 4-anonymity and the maximum value of individual risk decreased to 0.143. For QID^2, we notice that most of the unsafe records has a



☑ Show all tables

Select variable: all

# unsafe cells	Var 1	Var 2	Var 3	Var 4	Var 5
252	GENDER	YOB	POB	POR	
178	YOB	POB	POR		
55	GENDER	YOB	POB		
12	YOB	POB			
9	GENDER	YOB	POR		
8	GENDER	POB	POR		
2	POB	POR			
0	GENDER				
0	YOB				
0	ZIPCODE				
0	POB				
0	POR				
0	GENDER	YOB			
0	GENDER	ZIPCODE			
0	GENDER	POB			
0	GENDER	POR			
0	YOB	ZIPCODE			
0	YOB	POR			
0	GENDER	YOB	ZIPCODE		

Fig. 1. An overview of unsafe records for various quasi-identifiers

re-identification risk over 0.89. With the threshold value to 0.89, suppression of 254 values in place of birth variable were done. We observed only 3 records with anonymity set size k=1. Then suppression (threshold value = 0.5) was applied, 3 values in district of residence variable were suppressed.

The left side of figure 2 shows the distribution of individual risk of the original dataset for QID^2, while the right side shows the result of the published dataset. It is clear that the maximum value of individual risk decreased from 1.0 to 0.5. After three suppressions were done, for each quasi-identifier, the dataset satisfied 2-anonymity.

The third tool is CAT. As the tool restricts one quasi-identifier per anonymization process, we specified two quasi-identifiers respectively. Since CAT doesn't provide k-anonymity directly, we choose t-closeness criteria instead. We first provided t-closeness criteria on the QID^1 with a threshold value t to 0.5, which means the maximum value of individual disclosure risk is 0.5. This led to generalization method applied to year of birth and zip code variables. Specifically, every ten values in zip code variable were generalized into one category, which addressed the same effect on the published dataset as a truncation of the last digit of this variable; every two values in year of birth variable were generalized into one category.

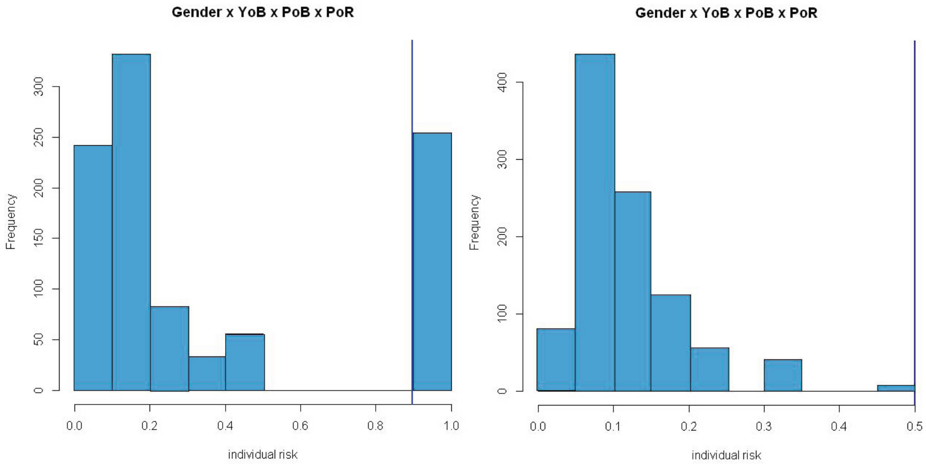


Fig. 2. The individual risk of the dataset for QID^2

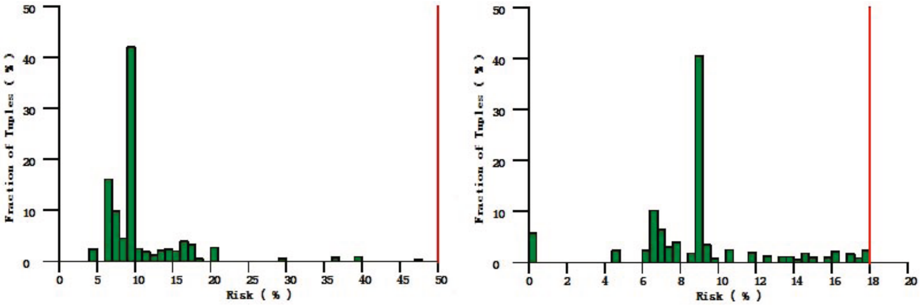


Fig. 3. The individual risk of the dataset for QID^1

As the left side of figure 3 shows, the current maximum value of individual disclosure risk is 0.5. After deleting 58 records, the maximum value decreased to 0.18. Looking into the right side of figure 3, which presents the distribution of individual disclosure risk of QID^1 on the published dataset, we found that less than 20 percent of the records have the risk above 0.1 and 2-anonymity was reached.

We then provided the t-closeness criteria with a threshold value t set to 0.978 on QID^2. This led to generalization method applied on year of birth variable, place of residence variable and place of birth variable. In particular, the values in place of residence variable were mapped into one category. The values in year of birth variable were equivalently mapped to 12 categories. The values in place of birth variable were equivalently mapped to 5 categories. After removing 57 records, the maximum value of individual risk decreased to 0.15.

Table 4. Two dimensions of comparison for various quasi-identifiers

Maximum level	**CAT**	**sdcMicro**	**μ-Argus**
ξ for QID^1	0.149	0.143	0
ξ for QID^2	0.402	0.500	0.384
λ for QID^3	3.600	2.028	1.682
λ for QID^4	86.631	3.261	1.783

Table 5. Scalability of three tools

Indicators	**CAT**	**sdcMicro**	**μ-Argus**
pass testcase	Yes	Yes	Yes
maximum vars in QID	4	33	5
maximum QID combination	1	6	5

For each quasi-identifier, these de-identification tools were able to publish the EHR dataset that satisfy 2-anonymity. We then analyzed the published datasets in terms of two aspects: 1) individual disclosure risk and 2) information loss.

Here we calculate the individual disclosure risk ξ of all published dataset using formula (1). Table 4 indicates that μ-Argus has produced safer dataset than the others, because it could protect patient's privacy under the lowest maximum individual risk for both quasi-identifiers. In particular, all records in the dataset produced by μ-Argus satisfy 2-anonymity for QID^1 as ξ equals 0. Following, we evaluated each published dataset in terms of their information loss λ (see formula (2)). Since CAT generalized all the values in the place of residence variable into one category, it led to a significant information distortion on QID^4. In contrast to CAT and sdcMicro, μ-Argus takes the lowest information loss for both quasi-identifiers to reach 2-anonymity.

Finally, we apply a virtual dataset of 10,000 records on these tools so as to evaluate their scalability. To compare the scalability of each tool, we first check whether it can load the test case. Then we examine the maximum variables in a quasi-identifier as well as the combination of quasi-identifiers. When evaluating maximum combination, we set the number of variables in a quasi-identifiers to the least maximum variables. Here, this experiment is carried out by a personal computer running the Windows 7 operating system.

As Table 5 shows, all three tools are capable of handling test case. For CAT, it can only handle one quasi-identifier at a time, which limits its scalability a lot. For μ-Argus, it has a limitation of its acceptance of variables of quasi-identifier. When exceeding its maximum, it fails with an error message which reports that the program ran out of memory (Note that this error is caused by the tool itself since memory resource is still adequate.) Since sdcMicro is able to handle 33 variables as a quasi-identifier and 6 combinations, we cannot observe its limitation using test case. Judging from Table 5, we conclude that sdcMicro has the best scalability among three tools.

6 Discussion on Results

Numerical methods are proposed to anonymize quasi-identifiers in order to avoid disclosing individual's sensitive information. However, not all these de-identifications methods such as masking were implemented in the discussed de-identification tools. Therefore, we are unable to report on the effectiveness of those methods on our EHR database. One of the constraints that our experiment has is the exclusion of a commercially available tool, i.e., The Privacy Analytics Risk Assessment Tool (PARAT) (`http://www.privacyanalytics.ca/privacy-analytics-inc-releases-version-26-of-parat/`, which is the only commercial product available so far. However, it has been reported that [4]) PARAT performs better than the algorithm implemented in CAT. Furthermore, PARAT, with its risk estimator, is able to produce more accurate de-identification results than the one incorporated in μ–Argus.

The results in this paper not only show the performance of the de-identification tools, but it also indicates the differences among tools based on the adopted algorithms to optimize the generalization steps. For instance, 254 values in place of birth variable were suppressed in sdcMicro, while all the values were generalized to 8 categories in μ-Argus. As μ-Argus generalized more variables than sdcMicro, it benefits from less records being suppressed and, thus, the statistical meanings of these variables can be preserved. This also shows a specialty of our controlled experiment that shows that the generalization method causes a lower information loss than suppression when the latter takes certain percent of the total records. Consequently, as applied to our EHR database, generalization method is more suitable than the suppression method.

For the purpose of comparison, we consider k-anonymity as the only privacy criteria that may lead to attribute disclosure problem on patient's clinical data. More considerably, no de-identification approach is being applied to clinical variables (i.e., infectious disease, blood type in PatientRecord table) leading an attacker to discover a patient's clinical information merely on finding a small variation in those clinical variables. Such an anticipated problem will also be resolved in the future development of our de-identification component for the ongoing EHR project. Likewise, identifying and implementing access control rules for external stakeholders accessing particular de-identified medical data is a complex task. Therefore, an appropriate access control and corresponding validation mechanism [15] must be placed to ensure better protection of any medical data to be offered for a secondary purpose.

7 Conclusion and Future Directions

This paper presented a rigorous assessment of the state-of-the-art de-identification tools that are available to researchers to publish datasets using anonymization techniques. The tools that have been evaluated, are CAT (Cornell Anonymization Kit), μ-Argus, and sdcMicro. We also discussed the significant features of each tool, their underlying anonymization methods, and the privacy criteria adopted.

Following, we analyzed the EHR database in terms of two categories: anonymity set size k and number of EHR data per anonymity set size for 15 quasi-identifiers. Our one of the important findings included that quasi-identifiers that contain place of birth and year of birth variables were the most identifiable. We selected two quasi-identifiers to be observed and anonymized. We also included two formulas, based on which the published dataset of each tool could be examined with respect to two dimensions: individual disclosure risk and information loss. For each tool, we outlined the anonymization process and provide 2-anonymity. Finally, we calculated the information loss and individual risk of each published dataset. As μ-Argus produced the safest records and caused the lowest information loss among these tools, it makes it more appropriate de-identification tool for anonymizing our EHR database. However, the study revealed that sdcMicro has the best scalability among three tools.

In this paper we evaluated the research-based de-identification tools dealing with structured data only. Before applying any of the de-identification tools, it is however important to know specific user requirements for de-identifying medical data. One of the future research directions includes investigating the de-identifications tools for unstructured data (e.g., clinical notes, reports, summaries, etc.), that we consider particularly relevant and usable for de-identifying legacy healthcare databases to avoid and mitigate data compromises.

Acknowledgments. This work has been supported by the project SAFEHR funded by Macao Science and Technology Development Fund. We are also grateful to the Kiang Wu Hospital Dialysis Center Macau SAR China, for their continuous support.

References

[1] di Vimercati, S.D., Foresti, S., Livraga, G., Samarati, P.: rotecting privacy in data release. In: 11th International School on Foundations of Security Analysis and Design, pp. 1–34 (2011)

[2] Emam, K.E.: Methods for the de-identification of electronic health records for genomic research. Genome Medicine 3(4), 25 (2011)

[3] Emam, K.E., Dankar, F.K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., Bottomley, J.: Research paper: A globally optimal k-anonymity method for the de-identification of health data. J. Am. Med. Inform. Assoc. (JAMIA) 16(5), 670–682 (2009)

[4] Emam, K.E., Dankar, F.K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., Bottomley, J.: Research paper: A globally optimal k-anonymity method for the de-identification of health data. JAMIA 16(5), 670–682 (2009)

[5] Fitzgerald, T.: Building management commitment through security councils. Information Systems Security 14(2), 27–36 (2005)

[6] Fraser, R., Willison, D.: Tools for de-identification of personal health information (September 2009), http://www.infoway-inforoute.ca/index.php/.../624-tools-for-de-identification-of-personal-health-information

[7] Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv. 42(4) (2010)

[8] Garde, S., Hovenga, E.J.S., Buck, J., Knaup, P.: Ubiquitous information for ubiquitous computing: Expressing clinical data sets with openehr archetypes. In: MIE, pp. 215–220 (2006)

[9] Gupta, D., Saul, M., Gilbertson, J.: Evaluation of a deidentification (de-id) software engine to share pathology reports and clinical documents for research. American Journal of Clinical Pathology, 176–186 (2004)

[10] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E.S., Seri, G., Wolf, P.-P.D.: Handbook on statistical disclosure control (December 2006)

[11] Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: 23rd International Conference on Data Engineering (ICDE 2007), pp. 106–115 (2007)

[12] Li, T., Li, N.: Optimal k-anonymity with flexible generalization schemes through bottom-up searching. In: IEEE International Conference on Data Mining Workshops (ICDMW 2006), pp. 518–523 (2006)

[13] Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: 23rd International Conference on Data Engineering (ICDE 2006), p. 24 (2006)

[14] Netherlands, S.: u-argus user's manual, `http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf`

[15] Qamar, N., Faber, J., Ledru, Y., Liu, Z.: Automated reviewing of healthcare security policies. In: 2nd International Symposium on Foundations of Health Information Engineering and Systems (FHIES), pp. 176–193 (2012)

[16] Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: PODS, p. 188 (1998)

[17] Sweeney, L.: Simple demographics often identify people uniquely. Pittsburgh: Carnegie Mellon University, Data Privacy Working Paper 3, 50–59 (2000)

[18] Sweeney, L.: Computational disclosure control - a primer on data privacy protection. Technical report, Massachusetts Institute of Technology (2001)

[19] Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 557–570 (2002)

[20] Templ, M.: Statistical disclosure control for microdata using the r-package sdcmicro. Transactions on Data Privacy 1(2), 67–85 (2008)

[21] Templ, M., Meindl, B.: The anonymisation of the cvts2 and income tax dataset. an approach using r-package sdcmicro (2007)

[22] Xiao, X., Wang, G., Gehrke, J.: Interactive anonymization of sensitive data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2009), pp. 1051–1054 (2009)