Privacy-preserving heterogeneous health data sharing

Noman Mohammed,[1] Xiaoqian Jiang,[2] Rui Chen,[1] Benjamin C M Fung,[1] Lucila Ohno-Machado[2]

▶ Additional appendices are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001027).

[1]Department of Computer Science and Software Engineering, Concordia University, Montreal, Quebec, Canada
[2]Division of Biomedical Informatics, University of California, San Diego, California, USA

Correspondence to
Noman Mohammed, Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd West, Montreal, QC H3G 1M8, Canada; no_moham@cse.concordia.ca

Received 20 April 2012
Revised 13 April 2012
Accepted 1 November 2012

To cite: Mohammed N, Jiang X, Chen R, *et al*. *J Am Med Inform Assoc* Published Online First: 12 December 2012 doi:10.1136/amiajnl-2012-001027

ABSTRACT

Objective Privacy-preserving data publishing addresses the problem of disclosing sensitive data when mining for useful information. Among existing privacy models, *ε-differential privacy* provides one of the strongest privacy guarantees and makes no assumptions about an adversary's background knowledge. All existing solutions that ensure ε-differential privacy handle the problem of disclosing relational and set-valued data in a privacy-preserving manner separately. In this paper, we propose an algorithm that considers both relational and set-valued data in differentially private disclosure of healthcare data.

Methods The proposed approach makes a simple yet fundamental switch in differentially private algorithm design: instead of listing all possible records (ie, a contingency table) for noise addition, records are generalized before noise addition. The algorithm first generalizes the raw data in a probabilistic way, and then adds noise to guarantee ε-differential privacy.

Results We showed that the disclosed data could be used effectively to build a decision tree induction classifier. Experimental results demonstrated that the proposed algorithm is scalable and performs better than existing solutions for classification analysis.

Limitation The resulting utility may degrade when the output domain size is very large, making it potentially inappropriate to generate synthetic data for large health databases.

Conclusions Unlike existing techniques, the proposed algorithm allows the disclosure of health data containing both relational and set-valued data in a differentially private manner, and can retain essential information for discriminative analysis.

INTRODUCTION

With the wide deployment of electronic health record systems, health data are being collected at an unprecedented rate. The need for sharing health data among multiple parties has become evident in several applications ,[1] such as decision support, policy development, and data mining. Meanwhile, major concerns have been raised about individual privacy in health data sharing. The current practice of privacy protection primarily relies on policies and guidelines, for example, the Health Insurance Portability and Accountability Act (HIPAA)[2] in the USA. HIPAA defines two approaches to achieve de-identification: the first is *Expert Determination*, which requires that an expert certify that the re-identification risk inherent in the data is sufficiently low; the second is *Safe Harbor*, which requires the removal and suppression of a list of attributes.[3] *Safe Harbor* requires data disclosers to follow a checklist[4] to remove specific information to de-identify the records.

However, there are numerous controversies on both sides of the privacy debate regarding these HIPAA privacy rules.[5] Some think that the protections provided in the de-identified data are not sufficient.[6] Others contend that these privacy safeguards hamper biomedical research, and that observing them may preclude meaningful studies of medical data that depend on suppressed attributes, for example, fine-grained epidemiology studies in areas with fewer than 20 000 residents or geriatric studies requiring detailed ages in those over 89.[3] There are concerns that privacy rules will erode the efficiencies that computerized health records may create, and in some cases, interfere with law enforcement.[5] Recently, the Institute of Medicine Committee on Health Research and the Privacy of Health Information concluded that the privacy rules do not adequately safeguard privacy and also significantly impede high-quality research.[7] The result is that patients' health records are not well protected at the same time that researchers cannot effectively use them for discoveries.[8] Technical efforts are highly encouraged to make published health data both privacy-preserving and useful.[9]

Anonymizing health data is a challenging task due to inherent *heterogeneity*. Modern health data are typically composed of different types, for example relational data (eg, demographics) and set-valued data (eg, diagnostic codes and laboratory tests). In relational data (eg, gender, age, body mass index), records contain only one value for each attribute. On the other hand, set-valued data (eg, diagnostic codes and laboratory tests) contain one or more values (cells) for each attribute. For example, the attribute-value {1*, 2*} of the *diagnostic code* contains two separate cells: {1*} and {2*}. For many medical problems, different types of data need to be published *simultaneously* so that the correlation between different data types can be preserved. Such an emerging heterogeneous data-publishing scenario, however, is seldom addressed in the existing literature on privacy technology. Current techniques primarily focus on a single type of data[10] and therefore are unable to thwart privacy attacks caused by inferences involving different data types. In this article, we propose an algorithm so that heterogeneous health data can be published yet retain essential information for supporting data mining tasks in a differentially private manner. The following real-life scenario further illustrates the privacy threats resulting from heterogeneous health data sharing.

Table 1 Raw patient data

ID	Sex	Age	Diagnostic code	Class
1	Male	34	11, 12, 21, 22	Y
2	Female	65	12, 22	N
3	Male	38	12	N
4	Female	33	11, 12	Y
5	Female	18	12	Y
6	Male	37	11	N
7	Male	32	11, 12, 21, 22	Y
8	Female	25	12, 21, 22	N

Example 1 Consider the raw patient data in table 1 (the attribute ID is just for the purposes of illustration). Each row in the table represents information from a patient.

The attributes *Sex*, *Age*, and *Diagnostic code* are categorical, numerical, and set-valued, respectively. Suppose that the data owner needs to release table 1 for the purpose of classification analysis on the class attribute, which has two values, Y and N, indicating whether or not the patient is deceased. If a record in the table is too specific such that not many patients can match it, releasing the data may lead to the re-identification of a patient. For example, Loukides *et al*[11] demonstrated that for the International Classification of Diseases (ICD), Ninth Revision (ICD-9) codes (or 'diagnostic codes' for brevity), one source of set-valued data could be used by an adversary for linkage to patients' identities. Needless to say, the knowledge of both relational and set-valued data about a victim makes the privacy attack easier for an adversary. Suppose that the adversary knows that the target patient is female and her diagnostic codes contain {11}. Then, record #4 can be uniquely identified, since she is the only *Female* with diagnostic codes {11,12} in the raw data. Thus, identifying her record results in disclosure that she also has {12}. Note that we do not make any assumption about the adversary's background knowledge. An adversary may have partial or full information about the set-valued data and can try to use any background knowledge to identify the victim.

To prevent such linking attacks, a number of partition-based privacy models have been proposed.[12–16] However, recent research has indicated that these models are vulnerable to various privacy attacks[17–20] and provide insufficient privacy protection. In this article, we employ *differential privacy*,[21] a privacy model that provides provable privacy guarantees and that is, by definition, immune against all aforementioned attacks. Differential privacy makes no assumption about an adversary's background knowledge. A differentially private mechanism ensures that the probability of any output (released data) is almost equally likely from all nearly identical input data sets and thus guarantees that all outputs are insensitive to any single individual's data. In other words, an individual's privacy is not at risk because of inclusion in the disclosed data set.

Motivation

Existing algorithms that provide differential privacy guarantees are based on two approaches: *interactive* and *non-interactive*. In an interactive framework, a data miner can pose aggregate queries through a private mechanism, and a database owner answers these queries in response. Most of the proposed methods for ensuring differential privacy are based on an interactive framework.[22–26] In a non-interactive framework the database owner first anonymizes the raw data and then releases the anonymized version for public use. In this article, we adopt the non-interactive framework as it has a number of advantages for data mining.[10] Current techniques that adopt the non-interactive approach publish contingency tables or marginals of the raw data.[27–30] The general structure of these approaches is to first derive a frequency matrix of the raw data over the database domain.

Table 2 Contingency table

Job	Age	Count
Professional	(18–40)	3
Professional	(40–65)	1
Artist	(18–40)	4
Artist	(40–65)	0

For example, table 2 shows the contingency table of table 3. After that, noise is added to each count to satisfy the privacy requirement. Finally, the noisy frequency matrix is published. However, this approach is not suitable for high-dimensional data with a large domain because when the added noise is relatively large compared to the count, the utility of the data is significantly destroyed. We also confirm this point in the 'Experimental description' section. Our proposed solution instead first probabilistically generates a generalized contingency table and then adds noise to the counts. For example, table 4 is a generalized contingency table of table 3. Thus the count of each partition is typically much larger than the added noise.

Contributions

We propose a novel technique for publishing heterogeneous health data that provides an *ε-differential privacy* guarantee. While protecting privacy is a critical element in data publishing, it is equally important to preserve the utility of the published data, since this is the primary reason for data release. Taking the decision tree induction classifier as an example, we show that our sanitization algorithm can be effectively tailored for preserving information in data mining tasks. The contributions of this article are:

1. To our knowledge, a *differentially private* data disclosure algorithm that simultaneously handles both relational and set-valued data has not been previously developed. The proposed differentially private data algorithm is based on a generalization technique and preserves information for classification analysis. Previous work[31] suggests that deterministic generalization techniques cannot be used to achieve ε-differential privacy, as they depend heavily on the data to be disseminated. Yet, we show that

Table 3 Sample data table

Job	Age
Engineer	34
Lawyer	50
Engineer	38
Lawyer	33
Dancer	20
Writer	37
Writer	32
Dancer	25

Table 4 Generalized contingency table

Job	Age	Count
Engineer	[18–40)	2
Engineer	[40–65)	0
Lawyer	[18–40)	1
Lawyer	[40–65)	1
Dancer	[18–40)	2
Dancer	[40–65)	0
Writer	[18–40)	2
Writer	[40–65)	0

differentially private data can be released through the addition of uncertainty in the generalization procedure.

2. The proposed algorithm can also handle numerical attributes. Unlike existing methods,[30] it does not require the numerical attributes to be pre-discretized. The algorithm adaptively determines the split points for numerical attributes and partitions the data based on the workload, while guaranteeing ε-differential privacy. This is an essential requirement for obtaining accurate classification, as we show in the 'Discussion' section. Moreover, the algorithm is computationally efficient.
3. It is well acknowledged that ε-differential privacy provides a strong privacy guarantee. However, the utility of data disclosed by *differentially private* algorithms has received much less study. Does an interactive approach offer better data mining results than a non-interactive approach? Does differentially private data disclosure provide less utility than disclosure based on k-anonymous data? Experimental results demonstrate that our algorithm outperforms the recently proposed differentially private interactive algorithm for building a classifier[26] and the *top-down specialization* (*TDS*) approach[32] that publishes k-anonymous data for classification analysis.

This article is organized as follows. The 'Preliminaries' section provides an overview of the generalization technique and presents the problem statement. Our anonymization algorithm is explained in 'The algorithm' section. In the 'Experimental description' section, we experimentally evaluate the performance of our solution, and we summarize our main findings in the 'Discussion' section.

PRELIMINARIES

In this section, we introduce the notion of generalization in the context of data publishing, followed by a problem statement.

Generalization

Let $D = \{r_1, \ldots, r_n\}$ be a multiset of records, where each record r_i represents the information of an individual with d attributes $A = \{A_1, \ldots, A_d\}$. We represent the data set D in a tabular form and use the terms 'data set' and 'data table' interchangeably. We assume that each attribute A_i has a finite domain, denoted by $\Omega(A_i)$. The domain of D is defined as $\Omega(D) = \Omega(A_1)x, \ldots x\Omega(A_d)$. To generalize a data set D, we replace the value of an attribute with a more general value. The exact general value is determined according to the attribute partition.

Definition 2.1 *(Partition)* The partitions $P(A_i)$ of a numerical attribute are the intervals $\langle I_1, I_2, \ldots, I_k \rangle$ in $\Omega(A_i)$ such that

$$\bigcup_{j=1}^{k} I_j = \Omega(A_i)$$

For categorical and set-valued attributes, partitions are defined by a set of nodes from the taxonomy tree such that it covers the whole tree, and each leaf node belongs to exactly one partition.

For example, Any_{sex} is the general value of *Female* according to the taxonomy tree of *Sex* in figure 1. Similarly, *age* 23 and 11 can be represented by the interval $[18-40)$ and the code 1*, respectively. For numerical attributes, these intervals are determined adaptively from the entire data.

Definition 2.2 (Generalization) Generalization is defined by a function $\Phi = \{\phi_1, \phi_2, \ldots \phi_d\}$, where $\phi_i : v \rightarrow p$ maps each value $v \in \Omega(A_i)$ to a $p \in p(A_i)$.

Clearly, given a data set D over a set of attributes $A = \{A_1, \ldots, A_d\}$, many alternative generalization functions are feasible. Each generalization function partitions the attribute domains differently. To satisfy ε-differential privacy, the algorithm must determine a generalization function that is insensitive to the underlying data. More formally, for any two data sets D and D′, where $|D \Delta D'| = 1$ (ie, they differ on at most one record), the algorithm must ensure that the ratio of $Pr[Ag(D) = \Phi]$ and $Pr[Ag(D') = \Phi]$ is bounded, where $Ag(\cdot)$ is a randomized algorithm (see online supplementary appendix A1). One naive solution satisfying ε-differential privacy is to have a fixed generalization function, irrespective of the input data set (ie, by definition *zero*-differentially private but useless). However, the proper choice of a generalization function is crucial since the data mining result varies significantly for different choices of partitioning. In the 'Experimental description' section, we present an efficient algorithm for determining an adaptive partitioning technique for classification analysis while guaranteeing ε-differential privacy. Online supplementary appendix A1 presents an overview of ε-differential privacy and the core mechanisms to achieve ε-differential privacy.

Problem statement

Suppose a data owner wants to release a de-identified data table $D^{\wedge}(A_{\downarrow}1^{\uparrow}pr, \ldots A_{\downarrow}d^{\uparrow}pr, A^{\uparrow}cls)$ where the symbols A_*^{pr} and A^{cls} correspond to predictor attributes and the class attribute, respectively, for release to the public for classification analysis. The attributes in D are classified into three categories: (1) an *identifier* A_i attribute that explicitly identifies an individual, such as *SSN* (social security number), and *Name*. These attributes are removed before releasing the data as per the HIPAA Privacy Rule; (2) a *class* attribute A^{cls} that contains the class value; the goal of the data miner is to build a classifier to accurately predict the value of this attribute; and (3) a set of d *predictor* attributes $A^{pr} = \{A_1^{pr}, \ldots, A_d^{pr}\}$, whose values are used to predict the binary label of the class attribute.

Figure 1 Taxonomy tree of attributes.

We require the class attribute to be categorical, and the predictor attribute can be categorical, numerical, or set-valued. Further, we assume that for each categorical or set-valued attribute A_i^{pr}, a taxonomy tree is provided. The taxonomy tree of an attribute A_i^{pr} specifies the hierarchy among the values. Our problem statement can be written as: given a data table D and the privacy parameter ε, our objective is to generate a de-identified data table $\widehat{D}$ such that $\widehat{D}$ (1) satisfies ε-differential privacy, and (2) preserves as much information as possible for classification analysis.

THE ALGORITHM

In this section, we present an overview of our *Diff*erentially private algorithm based on *Gen*eralization (*DiffGen*). We elaborate the key steps, and prove that the algorithm is ε-differentially private in online supplementary appendix A2. In addition, we present the implementation details and analyze the complexity of the algorithm in online supplementary appendix A3.

Algorithm 1 DiffGen

Input: Raw data set D, privacy budget ε, and number of specializations h Output: Generalized data set $\widehat{D}$

1: Initialize every value in D to the topmost value (see figure 2 for details);
2: Initialize Cut_i to include the topmost value (see figure 2 for details);
3: Set a privacy budget for specification of predictors $\varepsilon' \leftarrow \frac{\varepsilon'}{2(|A_n^{pr}|+2h)}$;
4: Determine the split value for each $v_n \in \cup Cut_i$ with probability $\alpha \exp\left(\frac{\varepsilon'}{2\Delta u} u(D, v_n)\right)$;
5: Compute the score for each candidate $\forall v \in \cup Cut_i$ (see online supplementary appendix A2 for details);
6: for $i = 1$ to h do
7: Select $v \in \cup Cut_i$ with probability $\alpha \exp\left(\frac{\varepsilon'}{2\Delta u} u(D, v)\right)$;
8: Specialize v on D and update $\cup Cut_i$;
9: Determine the split value for each new $v_n \in \cup Cut_i$ with probability $\alpha \exp\left(\frac{\varepsilon'}{2\Delta u} u(D, v_n)\right)$;
10: Update score for $v \in \cup Cut_i$;
11: end for
12: return each group with count $\left(c + Lap\left(\frac{2}{\varepsilon}\right)\right)$, where $Lap(\cdot)$ denotes the probability density function of Laplacian distribution.

Algorithm 1 first generalizes the predictor attributes A^{pr} and divides the raw data into several equivalence groups, where all the records within a group have the same attribute values. Then, the algorithm publishes the noisy counts of the groups. The general idea is to sanitize the raw data by a sequence of specializations, starting from the topmost general state as shown in figure 2. A *specialization*, written as $v \rightarrow child(v)$, where $child(v)$ denotes the set of child values of v, replaces the parent value v with a child value. The specialization process can be viewed as pushing the 'cut' of each taxonomy tree downwards. A *cut* of the taxonomy tree for an attribute A_i^{pr}, denoted by Cut_i, contains exactly one value on each root-to-leaf path. The value of the set-valued attribute of a record can be generalized to a *cut* if *every* item in the record can be generalized to a node in the cut and *every* node in the cut generalizes some items in the record. For example, the value $\{21, 22\}$ can be generalized to the hierarchy cuts $\{2*\}$ and $\{**\}$, but *not* $\{1*, 2*\}$. Figure 2 shows a solution cut indicated by the dashed curve representing table 5, which corresponds to de-identified data to be disseminated.

Initially, *DiffGen* creates a single partition by generalizing all values in A^{pr} to the topmost value in their taxonomy trees (line 1). The Cut_i contains the topmost value for each attribute A_i^{pr} (line 2). The specialization starts from the topmost cut and pushes down the cut iteratively by specializing some value in the current cut. At each iteration, *DiffGen* uses an *exponential mechanism*[33] (see online supplementary appendix A1) to select a candidate $v \in \cup Cut_i$ for specialization (line 7). Candidates are selected based on their score values (see online supplementary appendix A2), and different heuristics (eg, *information gain* and *max frequency*) can be used to determine the score of the candidates. Then, the algorithm specializes v and updates $\cup Cut_i$ (line 8). As taxonomy trees for the numerical attributes are not given, *DiffGen* again uses the exponential mechanism to determine the split value dynamically for each numerical candidate $V^n \in \cup Cut_i$ (lines 4 and 9). *DiffGen* specializesv by recursively distributing the records from the parent partition into disjoint child partitions with more specific values based on the taxonomy tree. For set-valued attributes, the algorithm computes the noisy count of each child partition to determine whether it is empty or not. Only 'non-empty' partitions are considered for further split in the next iteration. We provide additional details for candidate selection and the split value determination steps in online supplementary appendix A2. *DiffGen* also calculates the score for each new candidate due to the specialization (line 10). The algorithm terminates after a given number of specializations. Finally, the algorithm adds Laplace noise (see online supplementary appendix A1) to each equivalence group of the leaf partition to construct the sanitized data table $\widehat{D}$. We use the following example to facilitate understanding of how to use score functions, which are based on heuristics (eg, information gain and max frequency), for specification.

Example 2 Consider table 1 with ε=1 and h=2. Initially, the algorithm creates one root partition containing all the records that are generalized to $\langle Any_{\downarrow}Sex, [18-65), ** \rangle$. $\cup Cut_i$ includes $\{Any_{sex}, [18-65), **\}$. To find the first specialization among the candidates in $\cup Cut_i$, we compute the scores of (Any_{sex}), *[18–65)*, and $**$. We show how to compute the information gain *(InfoGain)* and maximum frequency *(Max)* scores of Any_{sex} for the specialization $Any_{sex} \rightarrow \{Male, Female\}$. Details of these two utility functions are discussed in online supplementary appendix 2.

1. Information gain:

$$H_{Anysex}(Table) = -\frac{4}{8} \times \frac{\log_2 4}{8} - \frac{4}{8} \times \frac{\log_2 4}{8} = 1$$

$$H_{Male}(Table) = -\frac{2}{4} \times \frac{\log_2 2}{4} - \frac{2}{4} \times \frac{\log_2 2}{4} = 1$$

$$H_{Female}(Table) = -\frac{2}{4} \times \frac{\log_2 2}{4} - \frac{2}{4} \times \frac{\log_2 2}{4} = 1$$

$$InfoGain(Table, Any_{sex}) = 1 - \left(\frac{4}{8} \times 1 + \frac{4}{8} \times 1\right) = 0$$

2. Max:

$$Max(Table, Any_{sex}) = 2 + 2 = 4$$

.

Let the first specialization be $** \rightarrow \{1*, 2*\}$. The algorithm then creates three child partitions with the child values $\{1*\}$,

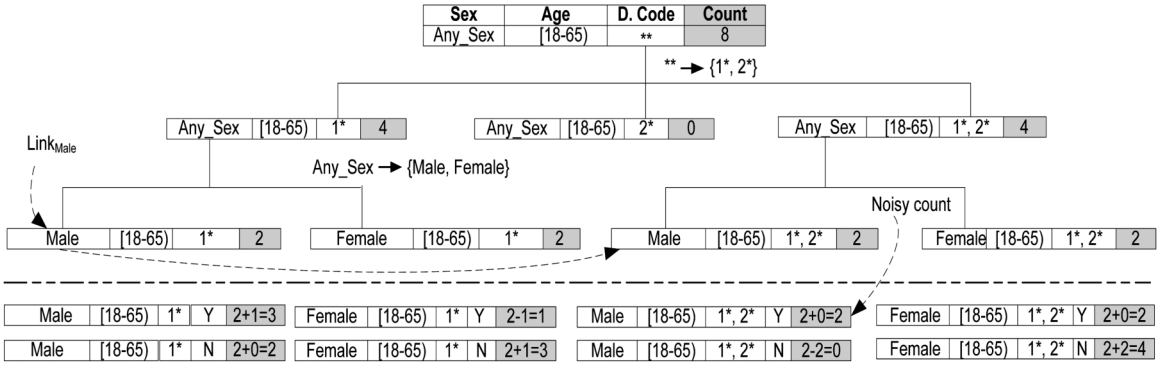

Figure 2 Tree for partitioning records. A randomized mechanism was deployed for specializing predictors in a top-down manner (using half of the privacy budget). At leaf nodes, random noise is added to the count of elements using the second half of the privacy budget to ensure overall ε-*differentially private* outputs.

{∗∗}, and {1∗, 2∗}, respectively, by replacing the node {∗∗} with different combinations of its children, leading r_3, r_4, r_5, and r_6 to the child partition {1∗} and r_1, r_2, r_7, and r_8 to the child partition {1∗, 2∗}. Suppose that the noisy counts indicate that these two child partitions are 'non-empty'. Then further splits are needed for them. There is no need to explore the child partition {2∗} any more, as it is considered 'empty'. $\cup Cut_i$ is updated to $\{Any_{sex}[18-65), 1*, 2*\}$. Suppose that the next specialization is $_{sex} \rightarrow \{Male, Female\}$, which creates further specialized partitions. Finally, the algorithm outputs the equivalence groups of each leaf partition along with their noisy counts as shown in figure 2 under the dotted line.

Please refer to online supplementary appendix A2 for privacy analysis of *DiffGen* and to online supplementary appendix A3 for implementation details.

EXPERIMENTAL DESCRIPTION

In this section our objectives are to study the impact of enforcing differential privacy on the data quality in terms of classification accuracy (CA), and to evaluate the scalability of the proposed algorithm for handling large data sets. We also compare *DiffGen* with *DiffP-C4.5*,[26] a differentially private interactive algorithm for building a classifier, and with the *TDS* approach[32] that publishes k-anonymous data for classification analysis. All experiments were conducted on an Intel Core i7 2.7GHz PC with 12GB RAM.

We employed two real-life data sets: *MIMIC* and *Adult*. We retrieved the *MIMIC* data set from the Multi-parameter Intelligent Monitoring in Intensive Care II research database,[34] which contains over 36 000 intensive care unit episodes. Specifically, we picked eight features (ie, marital status, gender, ethnicity, payment description, religion description, admission type, admission source, and ICD-9 code) and a target variable (ie, mortality). Of the eight features, the first seven are categorical attributes while the last one is a set-valued attribute. The publicly available *Adult*[35] data set is the 1994 US census data set that has been widely used for testing many sanitization algorithms. *Adult* has 45 222 census records with six numerical attributes, eight categorical attributes, and a binary *class* column representing two income levels, ≤US$50 K or >US$50 K. Please refer to Fung *et al*[32] for the description of attributes.

To evaluate the impact on classification quality, we divided the data into training and testing sets. First, we applied our algorithm to sanitize the training set and determine the $\cup Cut_i$. Then, the same $\cup Cut_i$ was applied to the testing set to produce a generalized test set. Next, we built a classifier on the sanitized training set and measured the CA on the generalized records of the test set. For classification models, we used the well-known C4.5 classifier.[36] For each experiment, we executed 10 runs and averaged the results over the runs.

Table 5 Differentially private disclosed data (ε=1, h=2)

Sex	Age	Diagnostic code	Class	Count
Male	[18–65)	1*	Y	3
Male	[18–65)	1*	N	2
Female	[18–65)	1*	Y	1
Female	[18–65)	1*	N	3
Male	[18–65)	1*, 2*	Y	2
Male	[18–65)	1*, 2*	N	0
Female	[18–65)	1*, 2*	Y	2
Female	[18–65)	1*, 2*	N	4

MIMIC data set

We applied *DiffGen* to the *MIMIC* data set for both utility functions (ie, Max and InfoGain). Figure 3 shows the CA for Max and InfoGain, where the privacy budget ε=0.1, 0.25, 0.5, 1, and the number of specializations h=5. We used two thirds of the records to build the classifier and measured the accuracy on the remaining third of the records. Both utility functions have similar performance, where CA spans from 86% to 89% under different privacy budgets. The experimental result suggests that the proposed algorithm can achieve good CA on heterogeneous health data. We could not directly compare our method with others for the *MIMIC* data set because we are not aware of an approach that can sanitize heterogeneous data while ensuring ε-differential privacy.

Adult data set

To better visualize the cost and benefit of our approach, we provide additional measures: *baseline accuracy* (*BA*) is the CA measured on the raw data without sanitization. BA−CA represents the cost in terms of classification quality for achieving a given ε-differential privacy requirement. At the other extreme, we measure *lower-bound accuracy* (*LA*), which is the accuracy on the raw data with all attributes (except for the *class* attribute) removed. CA−LA represents the benefit of our method over the naive non-disclosure approach.

Figure 4A depicts the CA for the utility function Max, where the privacy budget ε=0.1, 0.25, 0.5, 1, and the number of

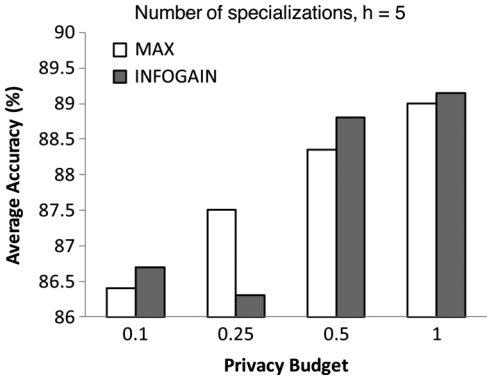

Figure 3 Classification accuracy for the *MIMIC* data set using *DiffGen* based on two scoring functions: information gain (INFOGAIN) and maximum frequency (MAX).

specializations $4 \leq h \leq 16$. The BA and LA are 85.3% and 75.5%, respectively, as shown in the figure by the dotted lines. For $\varepsilon=1$ and $h=10$, BA−CA is around 3% and CA−LA is 6.74%. For $\varepsilon=0.5$, BA−CA spans from 3.57% to 4.8%, and CA−LA spans from 5% to 6.23%. However, when ε decreases to 0.1, CA quickly decreases to about 78% (highest point), the cost increases to about 7%, and the benefit decreases to about 3%. These results suggest that for an acceptable privacy budget such as 1, the cost for achieving ε-differential privacy is small, while the benefit of our method over the naive method is large. Figure 4B depicts the CA for the utility function InfoGain. The performance of InfoGain is not as good as that of Max because the difference between the scores of a good and a bad attribute is much smaller for InfoGain as compared to Max. Therefore, the exponential mechanism does not work as effectively in the case of InfoGain as it does for Max.

Figure 5A shows the CA of *DiffGen*, *DiffP-C4.5*, and *TDS*. For *DiffGen*, we use utility function Max and fix the number of specializations $h=15$. *DiffP-C4.5* also uses the *Adult* data set and all the results of the *DiffP-C4.5* are taken from the paper by Friedman and Schuster.[26] For *TDS* we fixed the anonymity threshold $k=5$ and conducted the experiment ourselves. Following the same setting,[26] we executed 10 runs of 10-fold cross-validation to measure the CA.

The accuracy of *DiffGen* is clearly better than that of *DiffP-C4.5* for privacy budgets $\varepsilon \leq 2$. Note that the privacy budget should be typically smaller than 1.[21] Even for a higher budget, the accuracy of *DiffGen* is comparable to that of *DiffP-C4.5*. The major advantage of our algorithm is that we publish data and the data miner has much better flexibility to perform the required data analysis. On the other hand, in *DiffP-C4.5* the classifier is built through interactive queries; therefore, the database has to be permanently shut down to satisfy the privacy requirement after generating only *one* classifier. The experimental result also shows that *DiffGen* performs better than *TDS*. For a higher anonymity threshold k, the accuracy of *TDS* will be lower. One advantage of *DiffGen* is that, unlike *TDS*, it does not need to ensure that every equivalence group contains k records; therefore, *DiffGen* is able to provide more detailed information than *TDS*. This result demonstrates for the first time that, if designed properly, a differentially private algorithm can provide better utility than a partition-based approach.

All previous experiments can finish the sanitization process within 30 s. We further study the scalability of our algorithm over large data sets. We generate different data sets of different sizes by randomly adding records to the *Adult* data set. For each original record r, we create $\alpha-1$ variations of the record by replacing some of the attribute values randomly from the same domain. Here α is the blowup scale and thus the total number of records is $\alpha \times 45\,222$ after adding random records. Figure 5B depicts the runtime for 200 000 to 1 million records for $h=15$ and $\varepsilon=1$.

Summary

We observed two general trends from the experiments. First, the privacy budget has a direct impact on the CA. A higher budget results in better accuracy since it ensures better attribute partitioning and lowers the magnitude of noise that is added to the count of each equivalence group. Second, the CA initially increases with the increase in the number of specializations, but decreases after a certain threshold. This is an interesting observation. The number of equivalence groups increases quite rapidly with an increase in the number of specializations, resulting in a smaller count per group. Up to a certain threshold it has a positive impact due to more precise values; however, the influence of the Laplace noise gets stronger as the number of specializations grows. Note that if the noise is as large as the count, then the disclosed data are useless. This confirms that listing all possible combinations of values (ie, contingency table) and then adding noise to their counts is not a good approach for high-dimensional data since the noise will be as big as the count. Since this is a non-interactive approach, the data owner

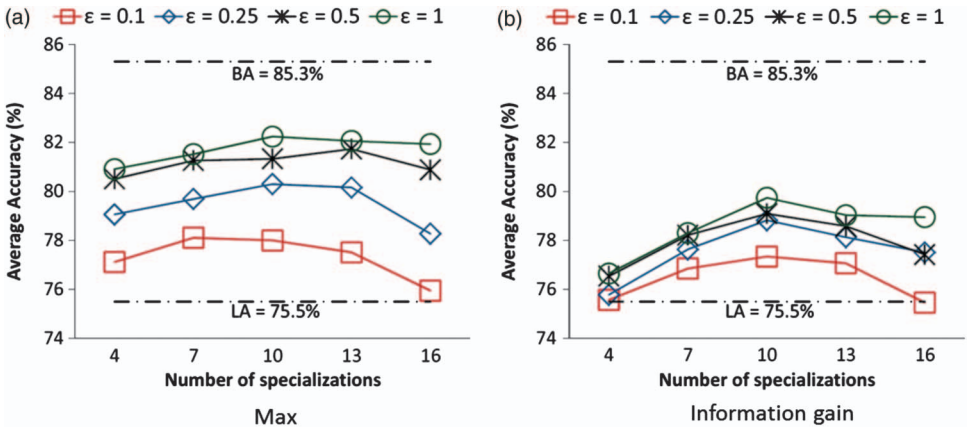

Figure 4 Classification accuracy for the *Adult* data set. BA, baseline accuracy; LA, lower-bound accuracy.

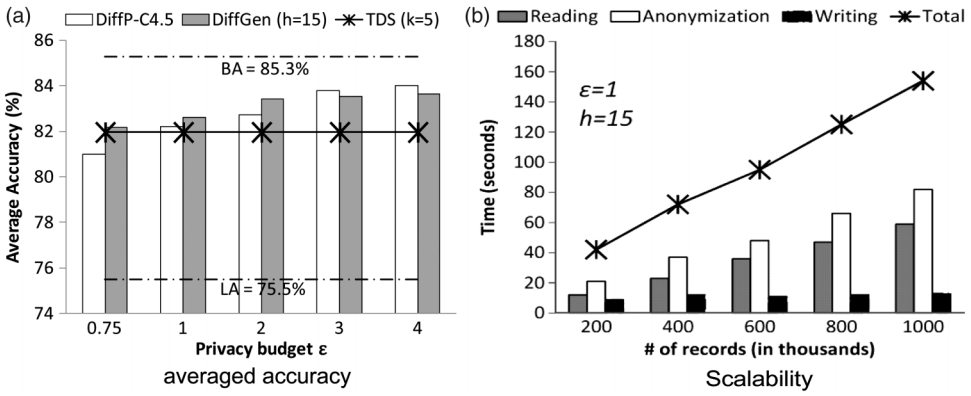

Figure 5 Comparison of *DiffGen* with *DiffP-C4.5* and top-down specialization (*TDS*) algorithms. (A) Evaluation of averaged accuracy, where the bottom and topmost lines stand for the worst case (ie, all records generalized to one super record) and the optimal case (ie, no record is generalized at all), respectively; (B) evaluation in terms of reading, anonymization, and writing time of all three algorithms.

can try different values of h to find the threshold and then release the sanitized data. Determining a good value of h adaptively, given the data set and the privacy budget, is an interesting approach for future work that we plan to investigate.

DISCUSSION

Does the algorithm yield a globally optimal solution? Can the algorithm be easily modified to anonymize sequential data? Are noise addition and generalization-based techniques desirable for the users of medical data sets? In this section, we provide answers to these questions.

Globally optimal

The proposed algorithm does not yield an optimal *solution cut*, rather it is suboptimal. The algorithm uses an exponential mechanism which probabilistically chooses a candidate with a high score. Thus, it is possible that a different solution cut may provide better utility. However, it is important to note that maximizing the overall sum of the *Score* for specializations in the training data does not guarantee having the lowest classification error in the testing data.

Sequential data

While set-valued data (not considering the order) are useful for many data analysis tasks, we acknowledge that in some other analysis tasks the order of items could provide extra useful information and may pose new re-identification risks. The proposed anonymization algorithm can be extended to handle sequential data as well. In order to anonymize sequential data, we need to preprocess each item in the sequence by its order and then consider the item–order pair as the new 'item'. Then, these new item sets could be used as an input to our algorithm, which will then prevent re-identification attacks based on the order of items. However, this is not currently implemented in the paper.

Usefulness of our approach

One of the inputs of our algorithm is the number of generalizations. Data users could specify the desired degree of generalizations by setting a proper value for h. Concerning the negative impact of noise (ie, on data utility) added to satisfy differential privacy, we expect the sanitized data from our algorithm to be useful for a few tasks (eg, classification as illustrated in this paper), but not all data analysis tasks, especially those that focus on attribute values of individuals. This is an inherent limitation of differential privacy methodology, and it is the price we pay to achieve a provable privacy guarantee. We acknowledge this limitation and will seek for better solutions in future work.

In summary, the generalization technique used in *DiffGen* might result in loss of information, which leads to a tradeoff problem between data utility and privacy protection, like many other privacy enhancement algorithms. The generalization is also context dependent. For example, if we want to generalize diseases (ie, asthma with respiratory disease and psoriasis with dermatological disease), resultant outputs might be adequate to answer some scientific questions (ie, classification tasks determined upfront) but may be insufficient to answer others, (eg, immunologic diseases, which could encompass both asthma and psoriasis). This is a fundamental limitation of generalization techniques and the type of lumping must be used carefully.

CONCLUSION

This paper presents a new anonymization algorithm that achieves ε-differential privacy and supports effective classification analysis for heterogeneous health data. The proposed solution connects the classical generalization technique with output perturbation to effectively sanitize raw data. Experimental results suggest that the proposed solution provides better utility than a pure differentially private interactive approach or an approach to simply produce k-anonymous data.

Acknowledgements We thank Dr Shuang Wang for helping with the experiments and useful discussions.

Contributors The authors are ranked according to their contributions. The authorship credit is based on meeting all the following criteria: (1) substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; (2) drafting the article or revising it critically for important intellectual content; and (3) final approval of the version to be published.

Funding BCMF, NM, and RC are supported in part by Discovery Grants (356065-2008) and Canada Graduate Scholarships from the Natural Sciences and Engineering Research Council of Canada (NSERC). XJ and LO-M were funded in part by NIH grants 1K99LM 011392-01, R01LM009520, U54 HL108460, R01HS019913, and UL1RR031980.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 O'Keefe CM. Privacy and the use of health data—reducing disclosure risk. *Electronic J Health Informatics* 2008;3:e5:1–e5:9.
2 Standards for privacy of individually identifiable health information. Final Rule, 45 CFR parts 160 and 164. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf (accessed 20 Feb 2012).

3 Benitez K, Loukides G, Malin BA. Beyond safe harbor: automatic discovery of health information de-identification policy alternatives. *The 1st ACM International Health Informatics Symposium*; ACM, 2010:163–72.
4 Madsen E, Masys DR, Miller RA. HIPAA Possumus. *J Am Med Inform Assoc* 2003;10:294.
5 Baumer D, Earp JB, Payton FC. Privacy of medical records: IT implications of HIPAA. *ACM Comput Soc (SIGCAS)* 2000;30:40–7.
6 McGraw D. Why the HIPAA privacy rules would not adequately protect personal health records: Center for Democracy and Technology (CDT) brief. http://www.cdt.org/brief/why-hipaa-privacy-rules-would-not-adequately-protect-personal-health-records (accessed 20 Feb 2012).
7 Institute of Medicine. *Beyond the HIPAA privacy rule: enhancing privacy, improving health through research*. Washington (DC): National Academies Press (US); 2009.
8 The HIPAA Privacy Rule: Lacks Patient Benefit, Impedes Research Growth: Association of Academic Health Centers (AAHC) brief, Washington, DC. http://www.aahcdc.org/Resources/ReportsAndPublications/IssueBriefs/View/ArticleId/65/The-HIPAA-Privacy-Rule-Lacks-Patient-Benefit-Impedes-Research-Growth.aspx (accessed 2 Oct 2011).
9 McGraw D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. *J Am Med Inform Assoc* 2012. In press. doi:10.1136/amiajnl-2012-000936
10 Fung BCM, Wang K, Chen R, *et al*. Privacy-Preserving Data Publishing: A survey of recent developments. *ACM Comput Surveys* 2010;42:1–53.
11 Loukides G, Denny J, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. *J Am Med Inform Assoc* 2010;17:322–7.
12 Sweeney L. *k*-anonymity: A model for protecting privacy. *Int J Uncertainty Fuzziness Knowledge Based Syst* 2002;10:557–70.
13 Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. *Process of The Acm Sigact Sigmod Sigart Symposium on Principles of Database Systems (PODS)*; 1998:1–13.
14 Machanavajjhala A, Kifer D, Gehrke J, *et al*. l-diversity: privacy beyond k-anonymity. *ACM Trans Knowledge Discov Data (TKDD)* 2007;1:3-es.
15 Wong RCW, Li J, Fu AWC, *et al*. (a,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*; Philadelphia: ACM, 2006:754–9.
16 Martin DJ, Kifer D, Machanavajjhala A, *et al*. Worst-case background knowledge for privacy-preserving data publishing. *IEEE 23rd International Conference on Data Engineering*; IEEE, 2007:126–35.
17 Wong RCW, Fu AWC, Wang K, *et al*. Minimality attack in privacy preserving data publishing. Proceedings of the 33rd International Conference on Very Large Data Bases; Vienna, Austria, 2007:543–54.
18 Ganta S, Kasiviswanathan S, Smith A. Composition attacks and auxiliary information in data privacy. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM, 2008:265–73.
19 Kifer D. Attacks on privacy and Definetti's theorem. Proceedings of the 35th SIGMOD International Conference on Management of Data; ACM, 2009:127–38.
20 Wong RCW, Fu AWC, Wang K, *et al*. Can the utility of anonymized data be used for privacy breaches? *ACM Trans Knowledge Discov Data* 2011;5:1–24.
21 Dwork C. Differential privacy. *Automata, Languages Programming*. 33rd International Colloquium on Automata, Languages and Programming (ICALP). Springer Verlag, 2006;4052:1–12.
22 Dwork C, McSherry F, Nissim K, *et al*. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography* 2006;3876:265–84.
23 Dinur I, Nissim K. Revealing information while preserving privacy. Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. New York, NY, USA: ACM, 2003:202–10.
24 Roth A, Roughgarden T. Interactive privacy via the median mechanism. Proceedings of the 42nd ACM symposium on Theory of computing. New York, NY, 2010:765–74.
25 Bhaskar R, Laxman S, Smith A, *et al*. Discovering frequent patterns in sensitive data. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, 2010:503–10.
26 Friedman A, Schuster A. Data mining with differential privacy. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*; New York, NY, USA: ACM Press, 2010:493–502.
27 Bhaskar B, Chaudhuri K, Dwork C, Kale S, McSherry F, Talwar K. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. New York, NY, USA: ACM, 2007:273–82.
28 Hay M, Rastogi V, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency. Very Large Database (VLDB) Endowment 2010;3 (1-2):1021–32.
29 Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference (TCC), 2006:265–84.
30 Xiao X, Wang G, Gehrke J. Differential Privacy via Wavelet Transforms. Knowledge and Data Engineering. *IEEE Transactions on*; 2011;23:1200–14.
31 Machanavajjhala A, Gehrke J, Gotz M. Data publishing against realistic adversaries. *Proceedings of the VLDB Endowment*; 2009;2:790–801.
32 Fung BCM, Wang K, Yu PS. Anonymizing classification data for privacy preservation. *IEEE Trans Knowledge Data Eng* 2007;19:711–25.
33 McSherry F, Talwar K. Mechanism design via differential privacy. *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*; IEEE, 2007:94–103.
34 Saeed M, Villarroel M, Reisner AT, *et al*. Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. *Crit Care Med* 2011;39:952–60.
35 Frank A, Asuncion A. UCI Machine Learning Repository, 2010.
36 Quinlan JR. *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann, 1993.

Privacy-preserving heterogeneous health data sharing

Noman Mohammed, Xiaoqian Jiang, Rui Chen, et al.

J Am Med Inform Assoc published online December 13, 2012
doi: 10.1136/amiajnl-2012-001027