BMC Medical Informatics and Decision Making

RESEARCH ARTICLE Open Access

Differentially private release of medical microdata: an efficient and practical approach for preserving informative attribute values

Hyukki Lee and Yon Dohn Chung*

Abstract

Background: Various methods based on *k*-anonymity have been proposed for publishing medical data while preserving privacy. However, the *k*-anonymity property assumes that adversaries possess fixed background knowledge. Although differential privacy overcomes this limitation, it is specialized for aggregated results. Thus, it is difficult to obtain high-quality microdata. To address this issue, we propose a differentially private medical microdata release method featuring high utility.

Methods: We propose a method of anonymizing medical data under differential privacy. To improve data utility, especially by preserving informative attribute values, the proposed method adopts three data perturbation approaches: (1) generalization, (2) suppression, and (3) insertion. The proposed method produces an anonymized dataset that is nearly optimal with regard to utility, while preserving privacy.

Results: The proposed method achieves lower information loss than existing methods. Based on a real-world case study, we prove that the results of data analyses using the original dataset and those obtained using a dataset anonymized via the proposed method are considerably similar.

Conclusions: We propose a novel differentially private anonymization method that preserves informative values for the release of medical data. Through experiments, we show that the utility of medical data that has been anonymized via the proposed method is significantly better than that of existing methods.

Keywords: Medical privacy, Data release, Data anonymization, Differential privacy, Privacy-preserving data publishing

Background

Introduction

In the last few decades, significant volumes of medical data have been collected and stored; consequently, there have been developments in the ability to process these data. Analytics on such stored data can help realize efficient healthcare services. For instance, data mining techniques applied to medical and social media data enable disease monitoring as well as health-based trend analyses. Furthermore, analyzing data of varying natures can help acquire new knowledge and intelligence, explore new hypotheses, and identify hidden patterns [1, 2].

Although possessing medical data benefits the data holders, it is occasionally necessary to release these data. For example, if data holders are not experts in conducting data analyses, they should outsource such analyses to a third-party. However, privacy concerns must take precedence during such a release of data, because the

*Correspondence: ydchung@korea.ac.kr
Department of Computer Science and Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, 02841 Seoul,Republic of Korea

data might include sensitive information, such as the disease statuses of individuals. Several privacy models have been proposed to protect the privacy of individuals. These models can be broadly categorized into two types: (1) k-anonymity and its extensions [3–6] and (2) differential privacy [7].

The concept of k-anonymity was introduced by Sweeney and Samarati [3]. In this model, each record of an individual contained in a released dataset cannot be distinguished from the records of at least k-1 other individuals. k-anonymity can reduce the risk of privacy breaches under certain assumptions; however, various studies have indicated the vulnerability of k-anonymity and proposed stronger privacy models such as l-diversity, t-closeness, and p-sensitive [4–6]. These privacy models are similar to k-anonymity as they guarantee privacy through syntactic conditions; thus, they are termed *syntactic privacy models*. Although syntactic privacy models can effectively protect privacy under certain conditions, they are inherently vulnerable to various attacks [8].

In contrast to syntactic privacy, differential privacy (also known as semantic privacy) provides a more rigorous guarantee of privacy, regardless of the background knowledge of adversaries. Dwork et al. introduced the concept of ϵ-differential privacy [7], which provides a mathematically provable guarantee of protecting the privacy of individuals. The goal of differential privacy is that the output of a query should not be considerably influenced when a single record is added or removed. Differential privacy has emerged as the *de-facto standard* for privacy-preserving data analyses.

Differential privacy typically targets privacy-preserving data mining, which responds to query processing of the data rather than the publishing of microdata. Although some methods for publishing differentially private data based on non-interactive settings have been proposed, these methods focus on aggregated results such as histograms or contingency tables [9, 10]. However, if the domain of informative attributes used for the analysis is large, such as the disease attributes in medical data, it is difficult to create a contingency table. In several real-world data publishing scenarios, releasing microdata is even more suitable due to the flexibility it yields to data analysts. Consequently, in this paper, we propose a method called ***IPA*** (Informative attribute Preserving Anonymization) for publishing medical microdata under differential privacy. This study focuses on the method to perturb a raw dataset to provide differentially private results on a record-by-record basis, while improving data utility by preserving informative attributes.

Motivation

The most commonly used method to achieve differential privacy is the addition of noise to the results. In a previously reported approach, noise was added to a contingency table of the raw dataset under non-interactive settings [9]. This implies that noise is added to every possible combination of the domain values for all attributes, irrespective of the existence of a record that corresponds to each combination in the raw dataset. For instance, suppose that we prepare a differentially private contingency table for the raw medical dataset listed in Table 1. The records are aggregated using all attributes, i.e., *Age, Gender and Disease*, to create a contingency table, which is presented as Table 2. Thereafter, noisy counts are added to every possible combination of the domain values for each attribute to achieve differential privacy, as shown in Table 3. If the dataset features many dimensions and/or the dimensions have large domains, a large amount of noise should be added. This leads to extreme distortion in the data.

Table 1 Original table

Age	Gender	Disease
10	M	Anemia
14	F	Gastritis
19	F	Pneumonia
12	F	Anemia
15	M	Pneumonia

To reduce the information loss caused by noise, generalization-based approaches have been proposed [10]. These approaches generalize original data by converting raw domain values with more general but semantically consistent values; for example, a specific *Age* value of 13 can be generalized into the interval [10-19]. Table 4 presents an example of a generalized contingency table. All the records have been generalized into indistinguishable groups, which are called *equivalent classes*, such as <[10-19], ∗, and Anemia>. Due to this generalization, the number of combinations is reduced; consequently, the total number of noisy counts is decreased.

It should be noted that generalization also distorts data, although the amount of distortion is less than that caused by noise. In particular, when informative attributes are generalized, the quality of data is affected considerably. Previous methods limit the informative attributes used for

Table 2 Contingency table created using Table 1

Age	Gender	Disease	Count
10	M	Anemia	1
12	F	Anemia	1
14	F	Gastritis	1
15	M	Pneumonia	1
19	F	Pneumonia	1

Table 3 Noisy version of contingency table

Age	Gender	Disease	Noisy count
10	M	Anemia	2
10	M	Gastritis	0
10	M	Pneumonia	1
10	F	Anemia	0
...	...	...	...
19	F	Gastritis	1
19	F	Pneumonia	1

analyses to *Class* attributes (i.e., True or False) and do not generalize informative attributes. Therefore, it is difficult to use these methods for publishing medical data, because such data typically involve informative attributes with large domains, such as those of diseases and medications. In this study, we neither generalize the informative attributes nor do we create contingency tables; instead, we publish anonymized microdata with raw informative values.

Contributions

Although several methods for releasing anonymized data have been proposed, a majority of these methods are based on syntactic privacy models [11, 12]. As mentioned above, stronger guarantees of privacy through differential privacy are required to protect the privacy of an individual. Furthermore, some of the previous works on publishing differentially private data are only relevant for classification analyses [13, 14]. In this paper, we propose a data anonymization method based on the differential privacy theory. To the best of our knowledge, this is the first work to propose a differentially private microdata publishing method for informative attributes with large domains. We evaluate the performance of the proposed method in terms of data utility and accuracy, through real-world analyses. The contributions of this study are as follows:

- We design a data anonymization method in which informative attributes remain unperturbed, while still complying with differential privacy. Regardless of the type and domain of the attribute, the raw informative values are preserved.
- We devise an algorithm that identifies useful anonymized datasets. This algorithm provides differentially private and high-utility anonymized datasets.
- We conduct extensive experiments and compare the proposed method with related existing methods. The experimental results prove that the proposed algorithm significantly improves data utility and also provides a rigorous privacy guarantee.

Table 4 Generalized noisy version of contingency table

Age	Gender	Disease	Noisy count
[10 − 19]	*	Anemia	3
[10 − 19]	*	Gastritis	1
[10 − 19]	*	Pneumonia	1

Preliminaries

Differential privacy is a rigorous privacy model that does not involve any assumptions regarding the background knowledge of adversaries. It guarantees that almost no difference will be observed in the output of any query when a single record is added to or removed from the database. Formally, differential privacy is defined as follows:

Definition 1 (**ϵ-differential privacy).** *Assume a mechanism $\mathcal{A}$ that randomizes query outputs and any pair of neighboring databases $\mathcal{D}$ and $\mathcal{D}'$. Then, $\mathcal{A}$ satisfies ϵ-differential privacy if and only if:*

$$Pr\left[\mathcal{A}\left(\mathcal{D}\right) = S\right] \leq exp\left(\epsilon\right) \times Pr\left[\mathcal{A}\left(\mathcal{D}'\right) = S\right] \quad \text{where } \mathcal{S} \in Range(\mathcal{A}). \tag{1}$$

□

We assume that $\mathcal{D}$ and $\mathcal{D}'$ are neighboring databases if they differ in exactly one record. In particular, we can obtain $\mathcal{D}'$ from $\mathcal{D}$ by adding or removing an arbitrary record. If Eq. 1 is satisfied, there is a high probability that $\mathcal{D}$ and $\mathcal{D}'$ produce the same query results. Therefore, even an adversary with maximal background knowledge cannot infer a particular record.

Definition 2 (**Sensitivity).** *For all $\mathcal{D}$ and $\mathcal{D}'$, the sensitivity of the function f is defined as*

$$\Delta f = \max_{\mathcal{D},\mathcal{D}'} \left|\left|f\left(\mathcal{D}\right) - f\left(\mathcal{D}'\right)\right|\right|. \tag{2}$$

□

Sensitivity is the maximal change inflicted on the output, when adding or removing an arbitrary record. Assume that the function f answers count queries over a dataset $\mathcal{D}$. Then, for any neighboring dataset $\mathcal{D}'$, the result from f would differ by at most 1; therefore, the sensitivity of f would be 1.

To satisfy differential privacy, two mechanisms have been proposed: the *Laplace mechanism* and the *exponential mechanism* [7, 15]. The Laplace mechanism adds noise to the output of the function; this noise is sampled from a Laplace distribution. The noise is decided based on the privacy parameter ϵ and the sensitivity of the function Δf.

Theorem 1 (**Laplace mechanism).** *Let $f(\mathcal{D})$ denote an output from the database $\mathcal{D}$. The Laplace mechanism satisfies ϵ-differential privacy if the random noise sampled from the Laplace distribution with mean μ=0 and scale*

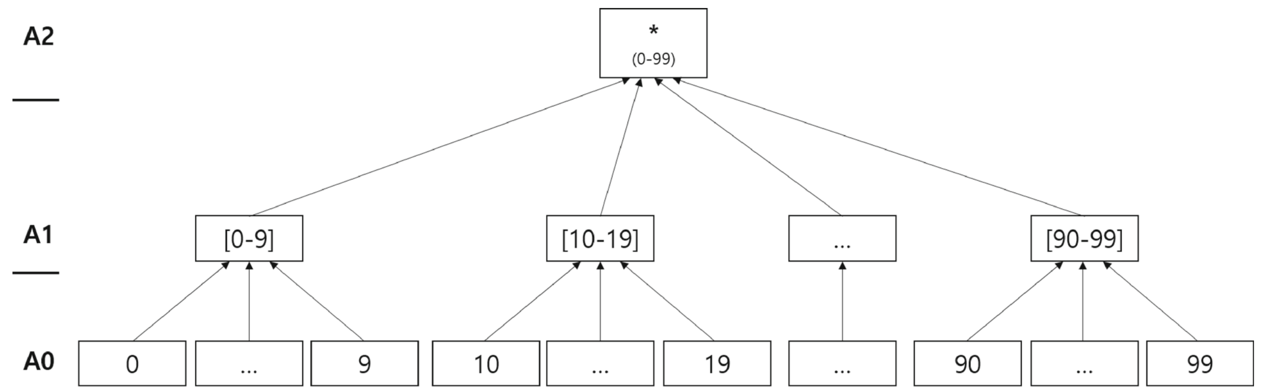


Fig. 1 Taxonomy tree of the *Age* attribute

b=$\Delta f/\epsilon$ *is added to f(*$\mathcal{D}$*).* □

The exponential mechanism is used with maximum utility when the output of the function is an object and not a real value. The aim of this exponential mechanism is to choose the output with the highest score. It assigns scores to possible outputs using a score function. Thereafter, the mechanism randomly selects an output from the possible result set. The likelihood of selection increases exponentially for the outputs with higher scores.

Theorem 2 (**Exponential mechanism).** *Let* $\mathcal{R}$ *be the possible results of the function f. For the score function* $\mathcal{S} : \mathcal{D} \times \mathcal{R} \rightarrow \mathbb{R}$*, a mechanism that outputs* $r \in \mathcal{R}$ *with a probability that is proportional to exp* $\left(\frac{\epsilon\mathcal{S}(\mathcal{D},r)}{2\Delta\mathcal{S}}\right)$ *satisfies* ϵ*-differential privacy, where* $\Delta\mathcal{S}$ *is the sensitivity of* $\mathcal{S}$*.* □

Differential privacy involves two composition properties: sequential composition and parallel composition [16]. Sequential composition is applicable to cases wherein a sequence of computations is performed on a single dataset, whereas parallel composition is applicable to a sequence of computations on disjoint datasets.

Theorem 3 (**Sequential composition).** *Let each function* f_i *provide* ϵ_i*-differential privacy. Thus, sequentially running all functions* f_i *over the dataset* $\mathcal{D}$ *provides* $\left(\sum_i \epsilon_i\right)$*-differential privacy.* □

Theorem 4 (**Parallel composition).** *Let each function* f_i *provide* ϵ_i*-differential privacy. Thus, applying each function over a set of disjoint datasets* $\mathcal{D}_i$ *provides* $\max_i(\epsilon_i)$*-differential privacy.* □

Generalization refers to replacing original values with less specific values. Generalized values are specified by a predefined generalization hierarchy. Figures 1, 2, and 3 present taxonomy trees representing the generalization hierarchies of the attributes *Age*, *Sex*, and *Zip*, respectively. Suppression involves substituting a specific value from the original dataset with a special symbol such as "$*$," which denotes "anything" in the anonymized dataset. In Figures 1, 2, and 3, $*$ is the suppressed value.

When anonymizing datasets, we employ the *full-domain generalization* algorithm [17], which maps the entire domain of an attribute in the initial microdata to a more general domain, based on its domain generalization hierarchy (also known as its taxonomy tree). Taxonomy trees of the attributes are combined to form a multi-attribute generalization hierarchical lattice. Figure 4 depicts an example of such a generalization lattice. Each combination, such as <A1, S0, Z0>, is called a *node*. The notation <A1, S0, Z0> indicates that all values in the *Age* attribute have been generalized using A1 in the taxonomy tree ($\{[0-9], [10-19], ..., [90-99]\}$) and that the *Sex* and *Zipcode* attributes have been generalized using S0 and Z0, respectively, (i.e., they are not generalized). The algorithm generalizes the dataset and measures information loss in the generalized dataset for each node of the lattice. The node with the lowest information loss is returned.

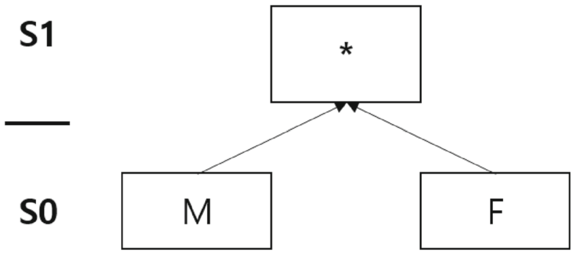


Fig. 2 Taxonomy tree of the *Sex* attribute

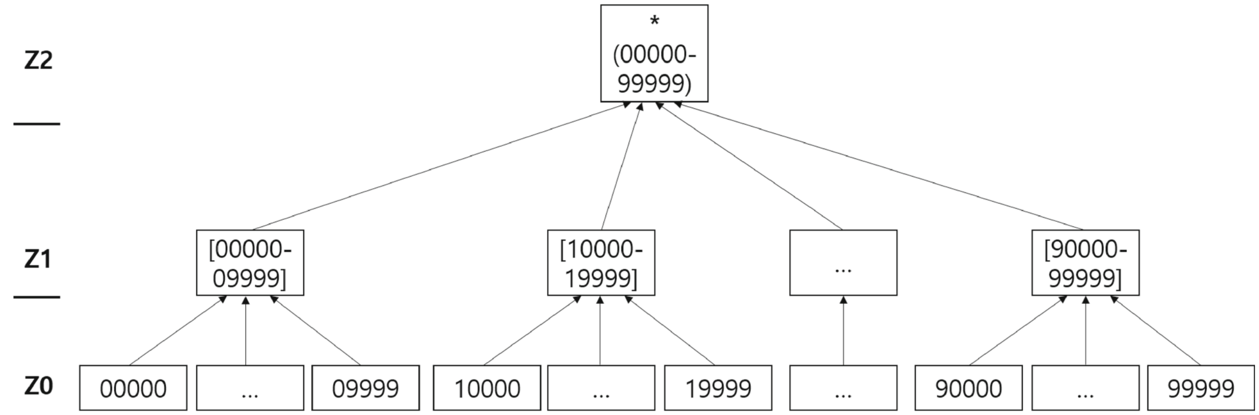


Fig. 3 Taxonomy tree of the *Zip* attribute

Methods

Problem settings

Consider that a data holder possesses a dataset $\mathcal{D}$ that contains multi-dimensional records, and each record belongs to a unique individual. This data holder wants to release an ϵ-differential private version of $\mathcal{D}$ with high data utility. It should be noted that all personal identifiable information, such as *SSNs (Social Security Numbers)*, has already been removed. $\mathcal{D}$ is defined as a set of records, and each record consists of a set of dimension attributes A_{dim}=$\{A_1, ..., A_q\}$ belonging to individuals, such as their age and gender. The A_{dim} attribute values of an individual might be acquired via publicly available data sources such as those on the world wide web and social networking services; thus, adversaries could easily obtain these values. Additionally, $\mathcal{D}$ contains informative large-domain categorical attributes A_{inf} that are used for data analyses. The A_{inf} attribute values are private information, and adversaries cannot obtain these values. Privacy breaches occur if adversaries gain knowledge regarding the A_{inf} values. We assume that each attribute $A_i \in A_{dim}$ has a predefined taxonomy tree.

Basic concepts

In this section, we introduce the overall process of the proposed anonymization method (*IPA*). *IPA* consists of three steps: (1) generating candidates for data perturbation, (2) utility scoring of all candidates, and (3) choosing the result based on the scores. Figure 5 presents the process of *IPA*.

Data perturbation is essential for anonymization, and several data perturbation techniques are available. We adopt three data perturbation methods: *generalization*, *suppression*, and *insertion*; these methods were chosen for specific reasons. Noise insertion is a typical method of

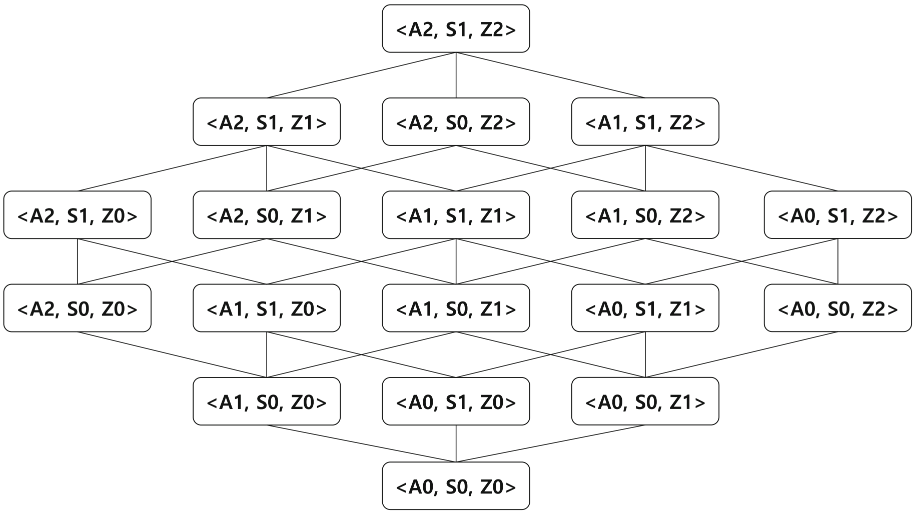
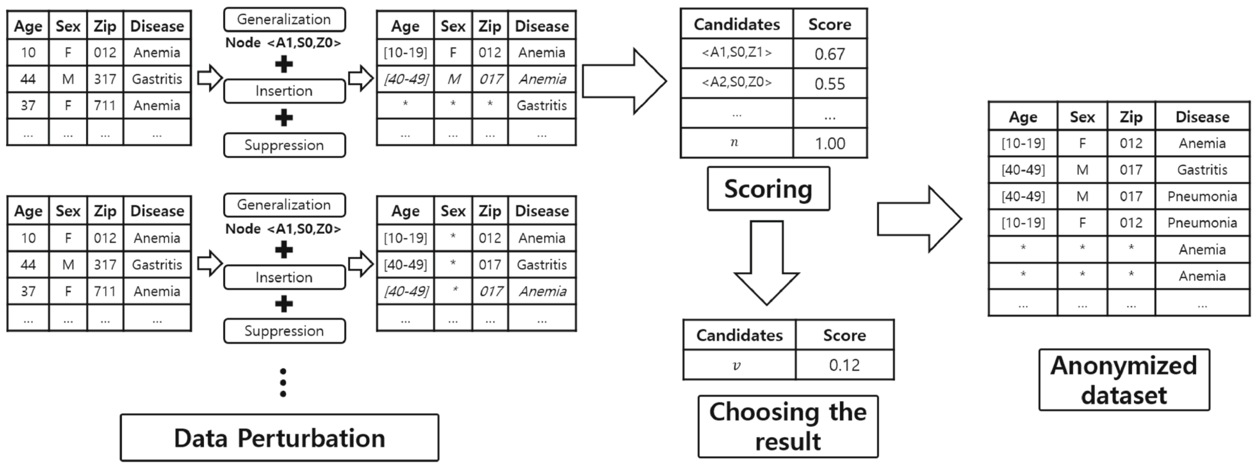


Fig. 4 Generalization hierarchical lattice



Age	Sex	Zip	Disease
10	F	012	Anemia
44	M	317	Gastritis
37	F	711	Anemia
...	...	...	...

Age	Sex	Zip	Disease
[10-19]	F	012	Anemia
[40-49]	*M*	*017*	*Anemia*
*	*	*	Gastritis
...	...	...	...

Age	Sex	Zip	Disease
10	F	012	Anemia
44	M	317	Gastritis
37	F	711	Anemia
...	...	...	...

Age	Sex	Zip	Disease
[10-19]	*	012	Anemia
[40-49]	*	017	Gastritis
[40-49]	*	*017*	*Anemia*
...	...	...	...

Candidates	Score
<A1,S0,Z1>	0.67
<A2,S0,Z0>	0.55
...	...
n	1.00

Candidates	Score
v	0.12

Age	Sex	Zip	Disease
[10-19]	F	012	Anemia
[40-49]	M	017	Gastritis
[40-49]	M	017	Pneumonia
[10-19]	F	012	Pneumonia
*	*	*	Anemia
*	*	*	Anemia
...	...	...	...

Fig. 5 Process of *IPA*

achieving differential privacy; however, the insertion-only approach involves substantial information loss due to the amount of noise. In terms of differential privacy, generalization does not help satisfy the privacy requirement. However, it can be used to improve utility by reducing noise and the domain size. Suppression is applied to equivalent classes containing few records. It helps reduce the number of counterfeit records; its details are described in subsequent sections. As *IPA* employs full-domain generalization, it generates candidates of perturbed datasets for all nodes in the generalization hierarchical lattice. Subsequently, the score of each dataset is measured based on the information loss and a result dataset is then selected. It should be noted that deterministic algorithms cannot satisfy differential privacy. Therefore, we employed the exponential mechanism to choose the node that will be the result dataset. In *IPA*, we allocate the privacy budget over four different parts, i.e., suppression threshold, number of counterfeit records, determining the informative attribute value of a counterfeit record, and choosing an anonymized dataset, which are proved by Theorems 5, 6, 7, and 8, respectively.

Step 1: data perturbation

In *IPA*, all dimension attributes $A_{dim}=\{A_1, ..., A_q\}$ are generalized using a predefined taxonomy tree (line 2 in Algorithm 1). The values of informative attributes (also known as measure attributes) remain unchanged during the generalization phase. The domain of generalized values is determined using the taxonomy tree. For example, Table 5 is an original table with A_{dim} = $\{Age, Gender, Zipcode\}$ and A_{inf} = $\{Disease\}$. Table 6 presents a generalized version of Table 5. As a result of this generalization, the values of attributes $A_1, ..., A_q$ in the same equivalent class become indistinguishable. This implies that the unit of the disjoint dataset has changed from a single record to an equivalent class. According to the parallel composition theorem, adding Laplace noise to each disjoint dataset can achieve differential privacy. Therefore, noise decreases as the number of equivalent classes decreases. When determining the *generalization* boundary, the privacy budget is not allocated. The generalization boundary is typically determined using the predefined taxonomy tree and not through a particular value or by distributing the dataset. Thus, one record

Table 5 Original table

Age	Gender	Zipcode	Disease
17	M	28912	Gastritis
16	M	23512	Pneumonia
13	M	24231	Pneumonia
24	F	31891	Anemia
29	F	34225	Anemia
25	F	37756	Diabetes
67	M	80061	Stroke

Table 6 Generalized table

Age	Gender	Zipcode	Disease
[10 − 19]	M	[20000 − 29999]	Gastritis
[10 − 19]	M	[20000 − 29999]	Pneumonia
[10 − 19]	M	[20000 − 29999]	Pneumonia
[20 − 29]	F	[30000 − 39999]	Anemia
[20 − 29]	F	[30000 − 39999]	Anemia
[20 − 29]	F	[30000 − 39999]	Diabetes
[60 − 69]	M	[80000 − 89999]	Stroke

does not affect the generalization boundaries of other records. Therefore, privacy breaches do not occur when determining the generalization boundary.

Algorithm 1: Data Perturbation Algorithm

```
   Input  : Original data $OriginalData$, Taxonomy trees
            $Taxonomy$, Privacy parameter $\epsilon$, and Suppression
            parameter $t$
   Output: Anonymized data $AnonymizedData$
1  $Inf \leftarrow$ Initialize each informative value in $OriginalData$;
2  /* Generate candidates based on taxonomy tree */
3  $\hat{T} = Generalization(OriginalData, Taxonomy)$;
4  $E \leftarrow$ list of equivalent classes in $\hat{T}$;
5  /* Suppression */
6  for $i = 1$ to $|k|$ do
7      if $|E_i| <= t + Lap\left((t-1)/\epsilon^{suppression}\right)$ then
8          $\hat{T}^{*}.add(Suppression(E_i))$;
9      else
10         $\hat{T}^{*}.add(E_i)$;
11     end
12 end
13 /* Record insertion */
14 $E \leftarrow$ list of equivalent classes in $\hat{T}^{*}$;
15 for $i = 1$ to $|l|$ do
16     $n \leftarrow$ number of records in $E_i$;
17     $n' \leftarrow n + Lap\left(1/\epsilon^{insertion}\right)$;
18     for $j = 1$ to $n'$ do
19         Determine an informative value $v \in Inf$ with
           probability $\left(\frac{\exp\left(\frac{\epsilon^{value}}{2\Delta S} S(E_i, v)\right)}{\sum_{v \in Inf} \exp\left(\frac{\epsilon^{value}}{2\Delta S} S(E_i, v)\right)}\right)$;
20         $E_i.add(v)$;
21     end
22     $AnonymizedData.add(E_i)$;
23 end
24 return $AnonymizedData$
```

In full-domain generalization, a given value is mapped to a pre-determined generalized value (or interval) for all records. Accordingly, an adversary can realize that a specific record is not present in the original dataset if its corresponding equivalent class does not exist in the result dataset. To prevent this type of privacy breach, we adopt the suppression technique (lines 6-12 in Algorithm 1). Suppression implies that all dimension attribute values of a record are substituted with "∗," which can be mapped to all the values in the domain. Because of the suppressed equivalent classes, adversaries will be unable to identify the equivalent class of the suppressed record. For example, in Tables 5 and 7, <[60-69], M, [80000-89999], Stroke> is suppressed to $<*, *, *,$ Stroke>. As the suppressed record is unknown, adversaries cannot identify the suppressed equivalent class from all other equivalent classes, except for the equivalent classes in the table. Furthermore, utility can also be improved via suppression. This is because suppression is performed on the generalized dataset and only a small amount of noise is added, as compared to the addition of noise for every possible equivalent class. We use the hyper-parameter t as the threshold for suppression. If the number of records in an equivalent class is less than or equal to t, the equivalent class is suppressed. For example, if we set $t = 2$, as the equivalent class corresponding to <[60-69], M, [80000-89999]> contains only one record, it is suppressed. All attribute values except the measure attributes are represented as "∗." However, it should be noted that using a fixed threshold value can result in a privacy breach. Assume that there are exactly t records in an equivalent class. Thus, the inclusion or exclusion of one record determines whether or not the equivalent class is suppressed. Accordingly, *IPA* uses the Laplace mechanism to add noise to the threshold value. Let the threshold be t and the Laplace noise be $T \sim Lap\left((t-1)/\epsilon^{suppression}\right)$. Then, the noisy threshold is $t + Lap\left((t-1)/\epsilon^{suppression}\right)$ (line 7), and sensitivity of the suppression threshold is $(t-1)$. More formally, suppression is defined as follows:

Definition 3 **(Suppression).** *Let OT be the original table, GT be the generalized table, t be the suppression threshold,* $\epsilon^{suppression}$ *be the privacy budget, and* $E_i (i = 1, ..., k)$ *be an equivalent class in GT. If* $|E_i| \leq t + Lap\left((t-1)/\epsilon^{suppression}\right)$, E_i *is suppressed.* □

Theorem 5 (Suppression threshold based on Definition 3 achieves $\left(\epsilon^{suppression}\right)$-differential privacy.)**.**

Table 7 Suppressed table

Age	Gender	Zipcode	Disease
[10 − 19]	M	[20000 − 29999]	Gastritis
[10 − 19]	M	[20000 − 29999]	Pneumonia
[10 − 19]	M	[20000 − 29999]	Pneumonia
[20 − 29]	F	[30000 − 39999]	Anemia
[20 − 29]	F	[30000 − 39999]	Anemia
[20 − 29]	F	[30000 − 39999]	Diabetes
*	*	*	Stroke

Table 8 Inserted table

Age	Gender	Zipcode	Disease
[10 − 19]	M	[20000 − 29999]	Gastritis
[10 − 19]	M	[20000 − 29999]	Pneumonia
[10 − 19]	M	[20000 − 29999]	Pneumonia
[10 − 19]	M	[20000 − 29999]	Gastritis
[20 − 29]	F	[30000 − 39999]	Anemia
[20 − 29]	F	[30000 − 39999]	Anemia
[20 − 29]	F	[30000 − 39999]	Diabetes
[20 − 29]	F	[30000 − 39999]	Anemia
*	*	*	Stroke

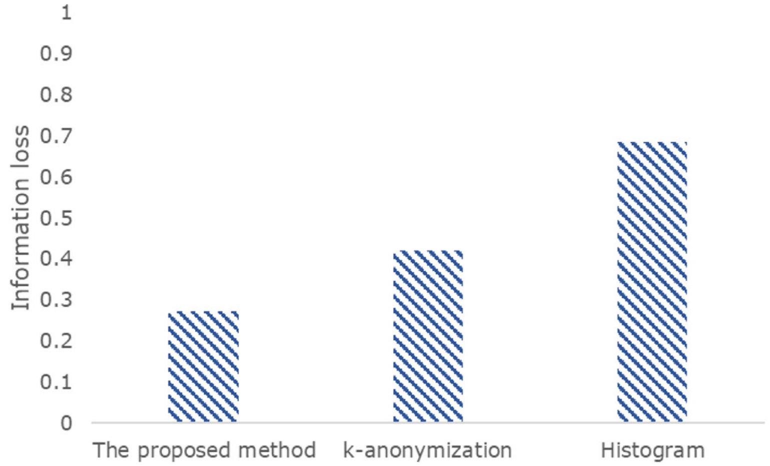


Fig. 6 Comparison of the proposed and previous methods in terms of information loss

Proof Let $(t - 1)$ be the sensitivity of a suppression threshold. Thus, the privacy budget is $\epsilon^{suppression}$, and a differentially private version of the suppression threshold is $t + Lap\left((t-1)/\epsilon^{suppression}\right)$. Based on Theorem 1, adding noise generated using the Laplace distribution $Lap\left((t-1)/\epsilon^{suppression}\right)$ to the suppression threshold achieves $\left(\epsilon^{suppression}\right)$-differential privacy. □

To comply with differential privacy, counterfeit records are inserted into equivalent classes as noise (lines 14-23). Two aspects need to be considered when inserting these counterfeit records. First, the number of counterfeit records to be inserted into each equivalent class needs to be determined. We use the Laplace mechanism to determine the number of counterfeit records to be inserted. Let the number of records in an equivalent class be n and the Laplace noise be $C \sim Lap\left(1/\epsilon^{insertion}\right)$. Thus, the size of an equivalent class, excluding suppressed or empty records, is $n + Lap\left(1/\epsilon^{insertion}\right)$ (lines 16-17).

Theorem 6 (Inserting $n + Lap(1/\epsilon^{insertion})$ counterfeit records achieves $\left(\epsilon^{insertion}\right)$-differential privacy.)**.**

Proof Let the sensitivity of a count query be 1, privacy budget be $\epsilon^{insertion}$, and number of counterfeit records be $n + Lap\left(1/\epsilon^{insertion}\right)$. All equivalent classes have exclusive boundaries determined using Theorems 1 and 4. Thus, adding independently generated counterfeit records from the Laplace distribution $Lap\left(1/\epsilon^{insertion}\right)$ to each equivalent class achieves $\left(\epsilon^{insertion}\right)$-differential privacy. □

Thereafter, we need to determine the informative attribute values of newly inserted records. The smaller the distortion in the informative value ratio of an equivalent class, the better the utility. Therefore, in *IPA*, informative attribute values are determined using the exponential mechanism with the ratio of number of informative values

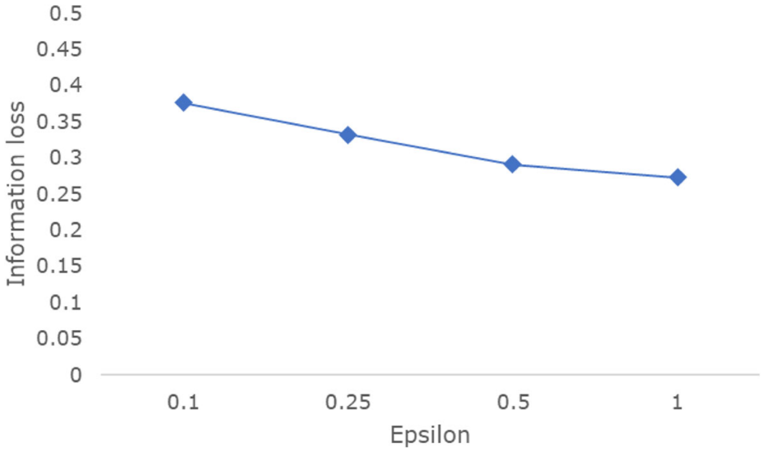
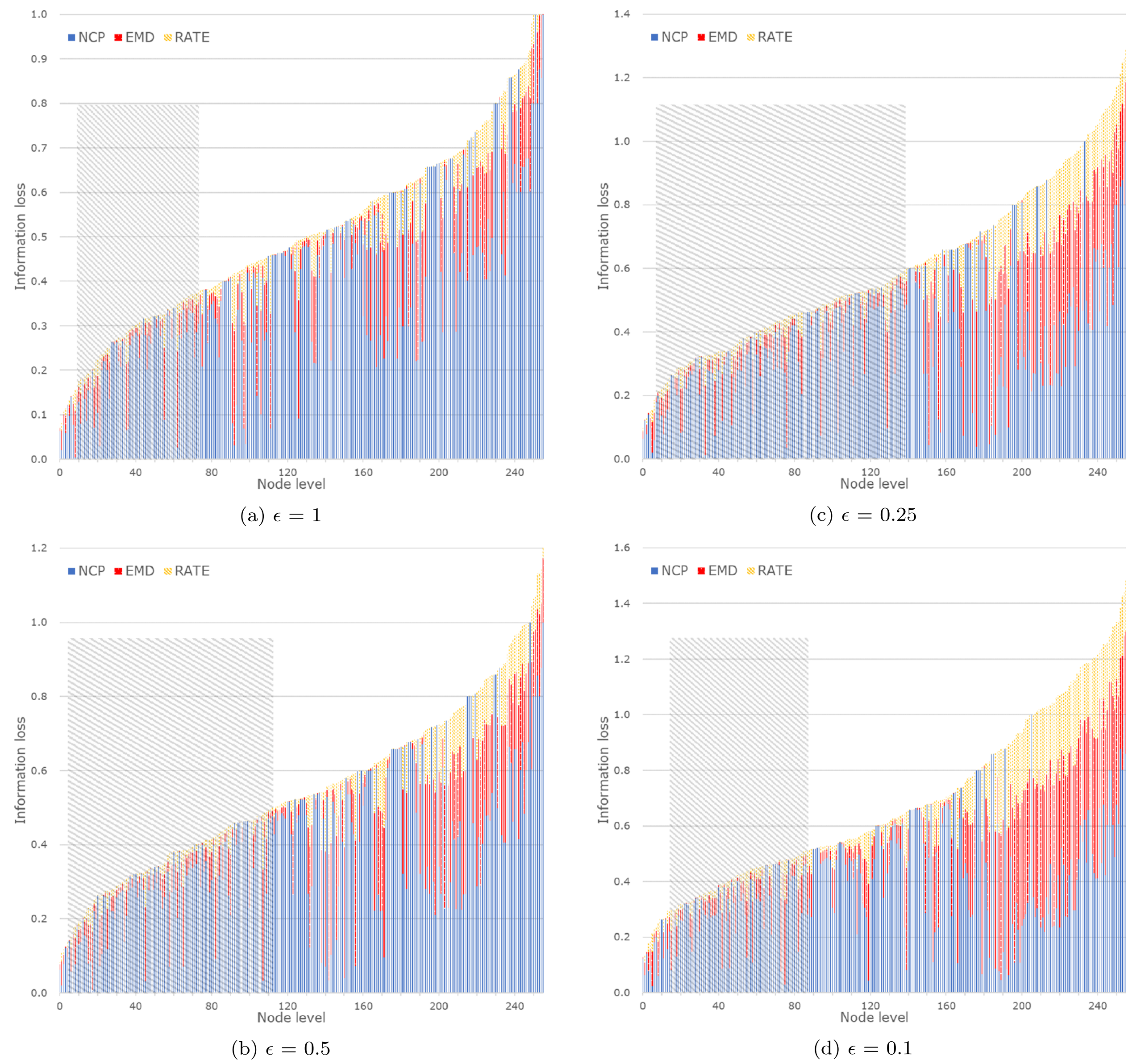


Fig. 7 Information loss with varying ϵ



(a) $\epsilon = 1$

(c) $\epsilon = 0.25$

(b) $\epsilon = 0.5$

(d) $\epsilon = 0.1$

Fig. 8 Information loss of candidate nodes

in an equivalent class. Let $Count_i(v)$ be the number of records that have the informative value v in E_i, where E_i is an equivalent class, Inf be a domain of informative values in $OriginalData$, and Inf_i be a domain of informative values in E_i. $|E_i|$ denotes the number of records in E_i, $|Inf|$ denotes the size of Inf, and $|Inf_i|$ denotes the size of Inf_i. The score function is calculated as follows:

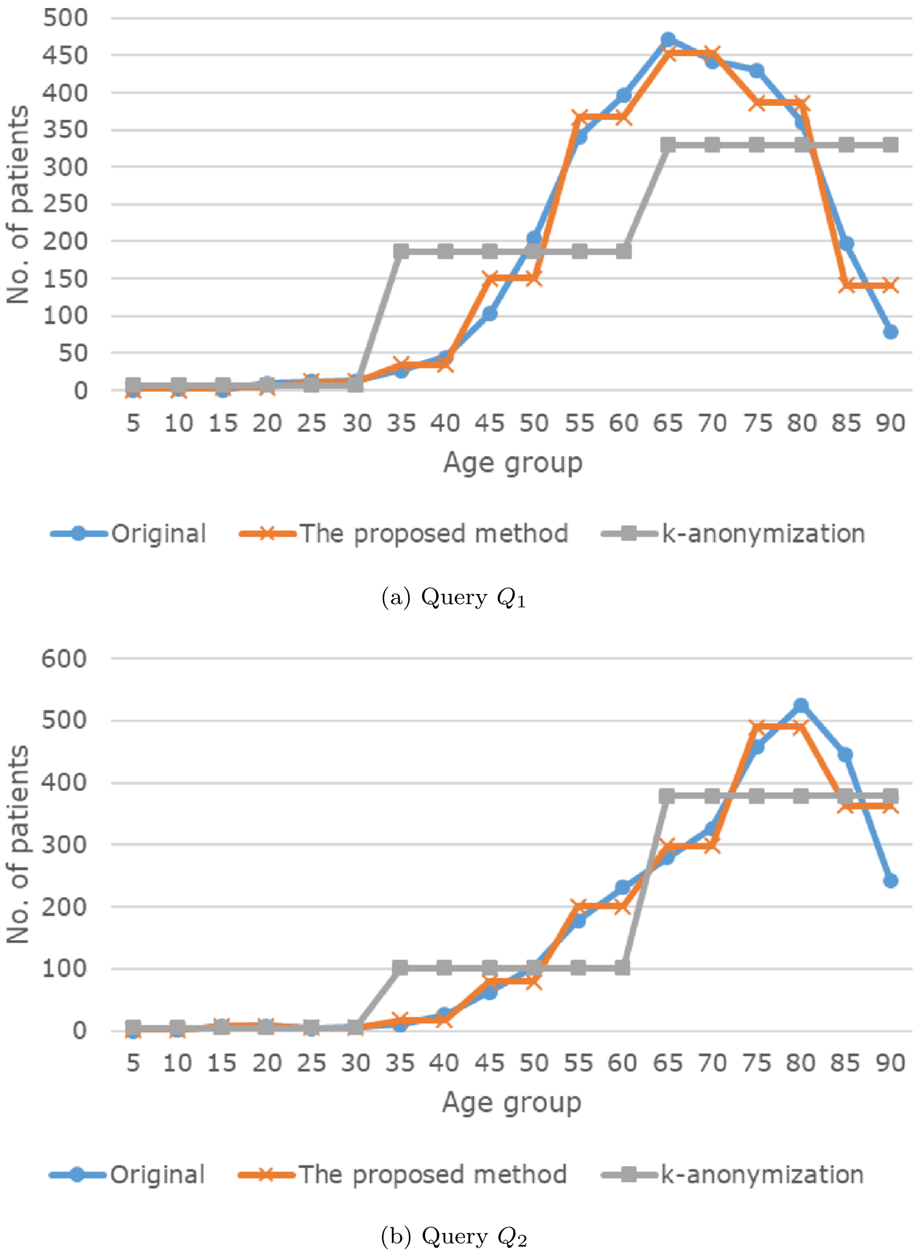
$$S(E_i, v) = \begin{cases} \frac{Count_i(v)}{|E_i|+1} & \text{if } v \text{ exists in } E_i \\ \frac{1}{(|E_i|+1)*(|Inf|-|Inf_i|)} & \text{otherwise} \end{cases} \quad (3)$$

Based on the scores of all candidates for the informative values, the exponential mechanism selects a candidate v with the following probability (line 19):

$$\frac{\exp\left(\frac{\epsilon^{value}}{2\Delta S} S(E_i, v)\right)}{\sum_{v \in Inf} \exp\left(\frac{\epsilon^{value}}{2\Delta S} S(E_i, v)\right)} \quad (4)$$

An example is presented in Table 8. Two records have been inserted: $<[10-19]$, M, $[20000-29999]$, *Gastritis* $>$ (Row 4) and $<[20-29]$, F, $[30000-39999]$, *Anemia* $>$ (Row 8).

Theorem 7 (Determining informative attribute values for inserted records based on Eq. 4 achieves (ϵ^{value})-differential privacy.).



(a) Query Q_1

(b) Query Q_2

Fig. 9 Results of the analysis queries

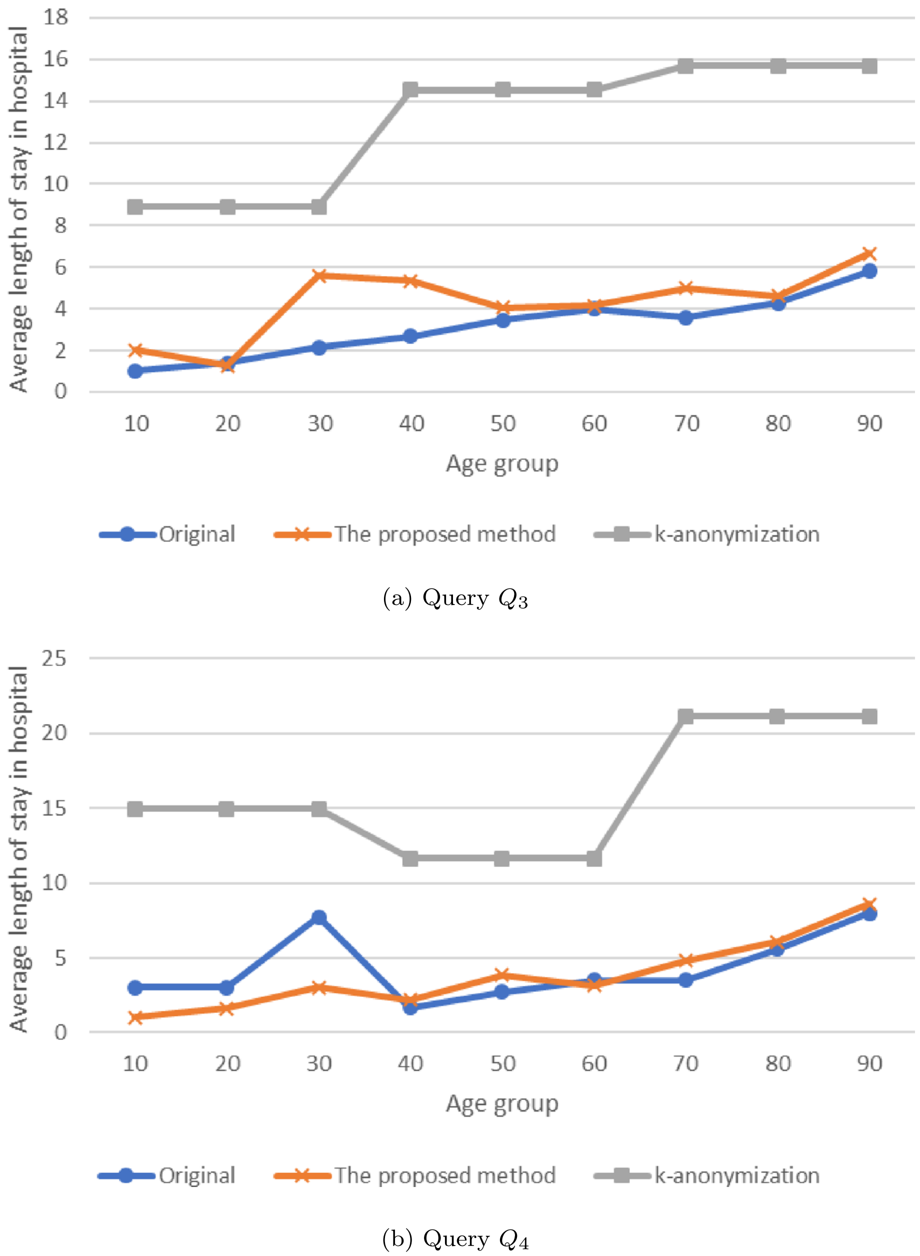
Proof Let *Inf* be the set of candidate values from which an informative attribute value is to be chosen. The *IPA* method selects a value $v \in Inf$ with the probability given in Eq. 4, where $S(E_i, Inf)$ is a score function and ΔS is the sensitivity of function S. Based on Theorem 2, choosing an informative value with a probability proportional to $\exp\left(\frac{\epsilon^{value}}{2\Delta S}\right)$ satisfies (ϵ^{value})-differential privacy. □

Step 2: scoring all candidates

We employ the information loss caused by data perturbation as a score function. In *IPA*, there are three factors that cause information loss.

The first factor is generalization. To measure the information loss caused by generalization, we introduce the concept of the NCP (Normalized Certainty Penalty) [18]. Let v be a value, $|v|$ be the number of leaf nodes covered by v corresponding to the generalization hierarchy, and $|\mathcal{L}|$ be the total number of leaf nodes in the generalization hierarchy. Then, the NCP of a value is defined as follows:

$$NCP_{value}(v) = \begin{cases} 0, & |v| = 1 (v \text{ is leaf}) \\ \frac{|v|}{|\mathcal{L}|}, & otherwise \end{cases} \tag{5}$$



(a) Query Q_3

(b) Query Q_4

Fig. 10 Results of the analysis queries

$$NCP(\hat{\mathcal{D}}) = \frac{\sum_{\forall r \in \hat{\mathcal{D}}} \sum_{\forall \mathcal{A}_{dim} \in \hat{\mathcal{D}}} NCP_{value}(v)}{|\hat{\mathcal{D}}|} \tag{6}$$

The value of NCP lies between 0 (i.e., minimum generalization) and 1 (i.e., maximum generalization). Therefore, the sensitivity of $\Delta NCP\left(\hat{\mathcal{D}}\right)$ is 1.

The second factor involves the distortion caused by inserted records. To measure this distortion, we employ the EMD (Earth Movers's Distance) measure, which evaluates the dissimilarity between two multi-dimensional distributions [5]. For two distributions of the original and anonymized datasets, i.e., $P_{\mathcal{D}} = (p_1, p_2, ..., p_m)$ and $Q_{\hat{\mathcal{D}}} = (q_1, q_2, ..., q_m)$, respectively, the EMD is defined as follows:

$$EMD\left[P_{\mathcal{D}},\ Q_{\hat{\mathcal{D}}}\right] = \frac{1}{2}\sum_{k=1}^{m} \left|p_k - q_k\right| \tag{7}$$

The EMD of two completely different equivalent classes is at most 1. Thus, the sensitivity of the EMD $\Delta EMD\left[P_{\mathcal{D}},\ Q_{\hat{\mathcal{D}}}\right]$ is 1.

Finally, the third factor in loss is the proportion of counterfeit records in equivalent classes, which can be defined as follows:

$$Rate_{class}(E_i) = \frac{\left|Counterfeit_i\right|}{|E_i|} \tag{8}$$

where $Counterfeit_i|$ denotes the number of counterfeit records inserted into E_i, and the sensitivity $\Delta Rate_{class}(E_i)$ is 1. *Rate* of the anonymized dataset $\hat{\mathcal{D}}$ is defined as follows:

$$Rate(\hat{\mathcal{D}}) = \frac{\sum_{\forall E_i \in \hat{\mathcal{D}}} Rate_{class}(E_i)}{\mathit{The\ number\ of\ equivalent\ classes}} \tag{9}$$

We use the sum of these three metrics to determine the total information loss.

$$IL(\hat{\mathcal{D}}) = NCP\left(\hat{\mathcal{D}}\right) + EMD\left[P_{\mathcal{D}},\ Q_{\hat{\mathcal{D}}}\right] + Rate\left(\hat{\mathcal{D}}\right) \tag{10}$$

As sensitivity of each metric is 1, the sensitivity of information loss $\Delta IL\left(\hat{\mathcal{D}}\right)$ is 3.

Step 3: choosing the result

In this section, we discuss the method of choosing a result from the set of candidates. Furthermore, we prove that *IPA* is differentially private.

We first measure the score of all candidates and then choose a result. To assign a high score to the dataset with low information loss, the score function u is calculated as follows:

$$u\left(\hat{\mathcal{D}}\right) = \left(3 - IL\left(\hat{\mathcal{D}}\right)\right) \tag{11}$$

Let $Candidates_i$ be the set of candidate anonymized datasets; thus, the result is selected using probability:

$$\frac{\exp\left(\frac{\epsilon^{candidates}}{2\Delta u} u\left(\hat{\mathcal{D}}\right)\right)}{\sum_{result \in Candidates_i} \exp\left(\frac{\epsilon^{candidates}}{2\Delta u} u\left(\hat{\mathcal{D}}\right)\right)} \tag{12}$$

Algorithm 2 illustrates the algorithm for choosing a result node. The algorithm begins with the creation of the hierarchical generalization lattice (line 1). Thereafter, the algorithm perturbs the original dataset for each node and calculates information loss (lines 2-5). After perturbing the dataset, a result is determined (line 7). The source code for Algorithms 1 and 2 is publicly available at GitHub [19].

Theorem 8 (Choosing an anonymized dataset according to Algorithm 2 achieves $\left(\epsilon^{candidates}\right)$-differential privacy.).

Proof Let $Candidates_i$ be the set of candidate datasets from which a single anonymized dataset is chosen. *IPA* selects the dataset $result \in Candidates_i$ using the probability in Eq. 12, where $u\left(\hat{\mathcal{D}}\right)$ is a score function and Δu is the sensitivity of the function u. Based on Theorem 2, choosing an informative value with a probability proportional to $\exp\left(\frac{\epsilon^{candidates}}{2\Delta u}\right)$ achieves $\left(\epsilon^{candidates}\right)$-differential privacy. □

Algorithm 2: Algorithm for Choosing a Result Node

```
Input : Original data $OriginalData$, Generalization
        lattice $Lattice$, Privacy parameter $\epsilon$, and
        Suppression parameter $t$
Output: Anonymized data $result$
1 $Candidates_i \leftarrow$
  $list\ of\ anonymized\ results\ and\ information\ loss$
2 for each node $o_i \in Lattice$ do
3     $temp =$
      $DataPerturbation(OriginalData, Lattice_{o_i}, \epsilon, t)$; //
      Algorithm 1
4     $Candidates.add(temp, u(temp))$;
5 end
6 Determine a $result \in Candidates_i$ with probability
  $\left(\frac{\exp\left(\frac{\epsilon^{candidates}}{2\Delta u} u\left(\hat{\mathcal{D}}\right)\right)}{\sum_{result \in Candidates_i} \exp\left(\frac{\epsilon^{candidates}}{2\Delta u} u\left(\hat{\mathcal{D}}\right)\right)}\right)$;
7 return $result$
```

Thus, we have proven that each part of *IPA* guarantees differential privacy. These parts run on the same dataset; therefore, according to Theorem 3, *IPA* achieves $\left(\epsilon^{suppression} + \epsilon^{insertion} + \epsilon^{value} + \epsilon^{candidates}\right)$-differential privacy.

Theorem 9 (*IPA* achieves $\left(\epsilon^{suppression} + \epsilon^{insertion} + \epsilon^{value} + \epsilon^{candidates}\right)$-differential privacy.).

Proof IPA consists of four parts: (1) determining the suppression threshold, (2) adding noisy records, (3) choosing an informative value, and (4) choosing a node. We showed that each operation is differentially private on its own. As these operations run on the same dataset, based on Theorem 3, *IPA* achieves $\left(\epsilon^{suppression} + \epsilon^{insertion} + \epsilon^{value} + \epsilon^{candidates}\right)$-differential privacy. □

Results and discussion

In this section, we present the experimental evaluation of *IPA* with respect to the utility of the output data and real-world analyses. For this evaluation, we use the NPS (National Patients Sample) dataset from HIRA (Health Insurance Review and Assessment which is a service in Korea) [20]. The NPS dataset consists of EHRs(Electronic Health Records) sampled from 3% sampled Korean people, in 2011. We analyze 1,361,000 records with 6 attributes: *Age, Sex, Length of stay in hospital, Location Surgery status,* and *Disease.* We consider the disease attribute as the informative attribute.

Table 9 Result of query Q_1

Age group	Original	The proposed method	k-anonymization
5	1.0	1	6.3
10	2.0	1	6.3
15	1.0	4	6.3
20	9.0	4	6.3
25	12.0	12	6.3
30	13.0	12	6.3
35	27.0	34.5	186.2
40	44.0	34.5	186.2
45	104.0	150.5	186.2
50	205.0	150.5	186.2
55	341.0	367.0	186.2
60	396.0	367.0	186.2
65	472.0	452.5	329.8
70	442.0	452.5	329.8
75	430.0	386.5	329.8
80	360.0	386.5	329.8
85	197.0	141.5	329.8
90	78.0	141.5	329.8

Data utility

We measure the amount of distortion in the anonymized dataset in comparison with its raw version. We compare the proposed method with k-anonymization [17] and differentially private histogram methods [10]. In medical privacy settings, epsilon is typically set as 0.1-2 [14, 21, 22]. According to previous studies, 10-anonymity can be achieved when epsilon is equal to 1 [23]. Therefore, we set the parameter values as $\epsilon = 1$ and $k = 10$. Figure 6 illustrates the information loss of anonymized datasets, where ϵ is 1 and $\epsilon^{suppression}$, $\epsilon^{insertion}$, ϵ^{value}, and $\epsilon^{candidates}$ are 0.1, 0.3, 0.3, and 0.3, respectively. The information loss of *IPA*, k-anonymization, and the histogram are 0.28, 0.43, and 0.69, respectively, as shown in the figure. For each experiment, we executed 10 runs and averaged the results of all the runs. *IPA* achieves lower information loss than the other methods, while guaranteeing more rigorous privacy.

Figure 7 illustrates the information loss while varying the privacy budget ϵ. As expected, the information loss tends to decrease when ϵ increases. Figures 8, 9, and 10 provide the details. The proportions of *NCP*, *EMD*, and *Rate* in total information loss are represented by blue, red, and yellow lines, respectively. The x-axis denotes the node level in the hierarchical generalization lattice, and the area shaded with gray blocks represents the range from which experimental results are selected. For example, in Fig. 8a, the average information loss is 0.28, and the range is 0.16 to 0.38. As ϵ decreases, the proportions of *EMD* and *Rate* become larger than that of *NCP*, the gray block area increases, and the overall information loss increases. The

Table 10 Result of query Q_2

Age group	Original	The proposed method	k-anonymization
5	0.0	1.0	4.5
10	2.0	1.0	4.5
15	7.0	8.0	4.5
20	8.0	8.0	4.5
25	4.0	4.5	4.5
30	6.0	4.5	4.5
35	10.0	17.5	101.7
40	26.0	17.5	101.7
45	63.0	79.5	101.7
50	102.0	79.5	101.7
55	178.0	201.0	101.7
60	231.0	201.0	101.7
65	279.0	298.0	379.2
70	326.0	298.0	379.2
75	457.0	489.5	379.2
80	525.0	489.5	379.2
85	445.0	363.5	379.2
90	243.0	363.5	379.2

Table 11 Result of query Q_3

Age group	Original	The proposed method	k-anonymization
10	1.0	2.2	8.9
20	1.4	1.4	8.9
30	2.1	5.7	8.9
40	2.7	5.5	14.5
50	3.4	4.2	14.5
60	4.0	4.1	14.5
70	3.6	4.8	15.7
80	4.3	4.5	15.7
90	5.8	6.5	15.7

Table 12 Result of query Q_4

Age group	Original	The proposed method	k-anonymization
10	3.0	1.0	15.0
20	3.0	1.4	15.0
30	7.7	2.8	15.0
40	1.7	2.4	11.7
50	2.7	3.6	11.7
60	3.5	3.2	11.7
70	3.5	4.7	21.1
80	5.6	6.1	21.1
90	8.0	8.5	21.1

range in Fig. 8d is narrower than that in Fig. 8c because lower level nodes are not selected by the score function as the overall information loss increases.

Real-world analysis

We present a real-world analysis to illustrate the usefulness of *IPA*. We compare the results of *IPA* with those of the original dataset and of k-anonymity, using aggregation queries. The queries used for data analysis are as follows:

- Q_1: SELECT FLOOR(Age/5)*5 AS AgeGroup, COUNT(*) FROM *NPS dataset* WHERE Sex = 'M' and *Surgery status* = 'N' and *Disease* = '*stroke*' GROUP BY FLOOR(Age/5)*5
- Q_2: SELECT FLOOR(Age/5)*5 AS AgeGroup, COUNT(*) FROM *NPS dataset* WHERE Sex = 'F' and *Surgery status* = 'N' and *Disease* = '*stroke*' GROUP BY FLOOR(Age/5)*5
- Q_3: SELECT FLOOR(Age/5)*5 AS AgeGroup, AVG(*Length of stay in hospital*) AS Average length of stay in hospital FROM *NPS dataset* WHERE Sex = 'M' and *Surgery status* = 'N' and *Disease* = '*stroke*' GROUP BY FLOOR(Age/5)*5
- Q_4: SELECT FLOOR(Age/5)*5 AS AgeGroup, AVG(*Length of stay in hospital*) AS Average length of stay in hospital FROM *NPS dataset* WHERE Sex = 'F' and *Surgery status* = 'N' and *Disease* = '*stroke*' GROUP BY FLOOR(Age/5)*5

Q_1 and Q_2 represent the number of *stroke* patients for each age group (0-4, 5-9,...,86-90). Q_3 and Q_4 represent the average duration of stay in a hospital.

Figures 9 and 10 and Tables 9, 10, 11, and 12 present the results of the analysis queries. In Fig. 9, the x-axis represents the age group (which corresponds to the first projection column of Q_1 and Q_2) and the y-axis represents the number of *stroke* patients (which corresponds to the second projection column of Q_1 and Q_2). In Fig. 10, the x-axis represents the age group (which corresponds to the first projection column of Q_3 and Q_4) and the y-axis represents the average duration of stay in a hospital for *stroke* patients (which corresponds to the second projection column of Q_3 and Q_4). In each figure and table, the results of *IPA* are more similar to those of the original data, compared to the results of k-anonymity.

Conclusions

Publishing anonymized microdata bestows additional flexibility to data recipients, as compared to providing sampled data or answers to specific queries. Considering this, we proposed a differentially private medical microdata releasing method that preserves measure attribute values; this proposed method is called *IPA*. To achieve this notion of privacy, we adopt differential privacy, which does not make any assumptions regarding the background knowledge of adversaries. To improve utility while preserving privacy, *IPA* employs three data perturbation methods: generalization, insertion, and suppression. *IPA* generalizes attribute values, except for measure attributes, to reduce the number of counterfeit records. Thereafter, it adds noisy records to achieve differential privacy; it also suppresses equivalent classes to avoid the addition of counterfeit records to empty equivalent classes. Through the results of our experiments, we demonstrated that *IPA* can reduce noise with an appropriate level of generalization. In addition, an experimental evaluation of a real-world data analysis proved that *IPA* can reduce information loss and also improve the utility of medical microdata published via differential private methods.

Abbreviations

IPA: Informative attribute preserving anonymization; SSN: Social security number; NCP: Normalized certainty penalty; EMD: Earth mover's distance; NPS: National patients sample; HIRA: Health insurance review and assessment; EHR: Electronic health record

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00269, A research on safe and convenient big data processing methods). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019H1D8A2105513).

Authors' contributions

HL designed the study, performed the analysis, and drafted the manuscript. YDC reviewed the manuscript, contributed to the discussion, and assisted with and supervised the design of the study.

Availability of data and materials

The datasets generated and analyzed in the current study are available in the HIRA repository [20].

Ethics approval and consent to participate

This study used open data from HIRA, which exempted the study from ethical approval according to the Korea National Institute for Bioethics Policy [24]. The Korean bioethics and safety act exempts ethical approval when exploiting existing data. Therefore, this study did not require ethical approval and consent from patients.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 24 December 2019 Accepted: 26 June 2020
Published online: 08 July 2020

References

1. Ren J-J, Sun T, He Y, Zhang Y. A statistical analysis of vaccine-adverse event data. BMC Med Inform Decis Mak. 2019;19(1):101.
2. Jing X, Emerson M, Masters D, Brooks M, Buskirk J, Abukamail N, Liu C, Cimino JJ, Shubrook J, De Lacalle S, et al. A visual interactive analytic tool for filtering and summarizing large health data sets coded with hierarchical terminologies (VIADS). BMC Med Inform Decis Mak. 2019;19(1):31.
3. Sweeney L. Int J Uncertain, Fuzziness Knowl-Based Syst. 2002;10(05): 557–70.

4. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov Data (TKDD). 2007;1(1):3.
5. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering. IEEE Computer Society; 2007. p. 106–15.
6. Truta TM, Vinay B. Privacy protection: p-sensitive k-anonymity property. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06). IEEE; 2006. p. 94.
7. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography Conference. Springer; 2006. p. 265–84.
8. Ganta SR, Kasiviswanathan SP, Smith A. Composition attacks and auxiliary information in data privacy. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2008. p. 265–73.
9. Mohammed N, Chen R, Fung B, Yu PS. Differentially private data release for data mining. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2011. p. 493–501.
10. Li H, Xiong L, Jiang X, Liu J. Differentially private histogram publication for dynamic datasets: an adaptive sampling approach. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM; 2015. p. 1001–10.
11. Lee H, Kim S, Kim JW, Chung YD. Utility-preserving anonymization for health data publishing. BMC Med Inform Decis Mak. 2017;17(1):104.
12. Xu Y, Ma T, Tang M, Tian W. A survey of privacy preserving data publishing using generalization and suppression. Appl Math Inf Sci. 2014;8(3):1103.
13. Xu C, Ren J, Zhang Y, Qin Z, Ren K. DPPro: Differentially private high-dimensional data release via random projection. IEEE Trans Inf Forensics Secur. 2017;12(12):3081–93.
14. Al-Hussaeni K, Fung BC, Iqbal F, Liu J, Hung PC. Differentially private multidimensional data publishing. Knowl Inf Syst. 2018;56(3):717–52.
15. McSherry F, Talwar K. Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). IEEE; 2007. p. 94–103.
16. McSherry F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. ACM; 2009. p. 19–30.
17. LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. ACM; 2005. p. 49–60.
18. Xu J, Wang W, Pei J, Wang X, Shi B, Fu AW-C. Utility-based anonymization using local recoding. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2006. p. 785–90.
19. Informative Attribute Preserving Anonymization Method. 2020. Available at https://github.com/hyukki-db/IPA. Accessed 30 Apr 2019.
20. Health Insurance Review and Assessment Service in Korea. 2012. Available at http://opendata.hira.or.kr. Accessed 1 Dec 2019.
21. Mohammed N, Jiang X, Chen R, Fung BC, Ohno-Machado L. Privacy-preserving heterogeneous health data sharing. J Am Med Inform Assoc. 2013;20(3):462–9.
22. 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.
23. Li N, Qardaji WH, Su D. Provably private data anonymization: Or, k-anonymity meets differential privacy. CoRR, abs/1101.2604. 2011;49:55.
24. Korea National Institute for Bioethics Policy. Available at http://irb.or.kr/menu02/commonDeliberation.aspx. Accessed 23 June 2019.

Publisher's Note

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