(ε, k)-Randomized Anonymization: ε-Differentially Private Data Sharing with k-Anonymity

Akito Yamamoto[1][a], Eizen Kimura[2][b] and Tetsuo Shibuya[1][c]
[1]*Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan*
[2]*Medical School of Ehime University, Ehime, Japan*

Keywords: Differential Privacy, Randomized Response, k-Anonymity, Data Sharing.

Abstract: As the amount of biomedical and healthcare data increases, data mining for medicine becomes more and more important for health improvement. At the same time, privacy concerns in data utilization have also been growing. The key concepts for privacy protection are k-anonymity and differential privacy, but k-anonymity alone cannot protect personal presence information, and differential privacy alone would leak the identity. To promote data sharing throughout the world, universal methods to release the entire data while satisfying both concepts are required, but such a method does not yet exist. Therefore, we propose a novel privacy-preserving method, (ε, k)-Randomized Anonymization. In this paper, we first present two methods that compose the Randomized Anonymization method. They perform k-anonymization and randomized response in sequence and have adequate randomness and high privacy guarantees, respectively. Then, we show the algorithm for (ε, k)-Randomized Anonymization, which can provide highly accurate outputs with both k-anonymity and differential privacy. In addition, we describe the analysis procedures for each method using an inverse matrix and expectation-maximization (EM) algorithm. In the experiments, we used real data to evaluate our methods' anonymity, privacy level, and accuracy. Furthermore, we show several examples of analysis results to demonstrate high utility of the proposed methods.

1 INTRODUCTION

With the recent increase in health awareness, the volume of biomedical data has grown, and data mining for medicine and healthcare has gained importance (Rakesh Kumar et al., 2019; Wu et al., 2021). At the same time, privacy concerns in releasing data have been recognized (Hlávka, 2020; Su et al., 2021), and the development and discussion of privacy-preserving methods of personal information contained in datasets are now one of essential research topics. Furthermore, general data sharing methods in compliance with European Union's General Data Protection Regulation (GDPR) (European Commission, 2016) and other regulations are required to promote the utilization of medical data on a global basis in the future.

The two most important concepts to publish data while protecting privacy are k-anonymity (Sweeney, 2002) and differential privacy (Dwork, 2006). k-anonymity can prevent identity disclosure and is widely used for healthcare data sharing (Emam and Dankar, 2008; Lee et al., 2017). Differential privacy is a framework to protect information on individuals' participation and has increasingly been applied to genomic data and other health research (Aziz et al., 2019; Ficek et al., 2021). These two concepts focus on different aspects of data privacy and need to complement each other: k-anonymity alone would reveal the presence of an individual, while differential privacy alone would leak the identity. In fact, it is reported that for a data application system to be GDPR-compliant, we need to satisfy differential privacy in addition to anonymity (Cummings and Desai, 2018).

In this study, we propose new privacy-preserving methods for data sharing that satisfy both k-anonymity and ε-differential privacy. Our methods differs from existing methods (Li et al., 2012; Holohan et al., 2017; Tsou et al., 2021) in that they do not assume data sampling and can release all the information in the original data. In particular, the contributions of this study are as follows:

1. We employ the randomized response as a mecha-

[a] https://orcid.org/0000-0002-3769-3352
[b] https://orcid.org/0000-0002-0690-8568
[c] https://orcid.org/0000-0003-1514-5766

Yamamoto, A., Kimura, E. and Shibuya, T.
(, k)-Randomized Anonymization: -Differentially Private Data Sharing with k-Anonymity.
DOI: 10.5220/0011665600003414
In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 5: HEALTHINF*, pages 287-297
ISBN: 978-989-758-631-6; ISSN: 2184-4305

nism to satisfy ε-differential privacy, and by combining it with k-anonymization, we propose three methods to release data while protecting both identity and presence of individuals. The first two methods perform k-anonymization and randomized response in sequence. While these algorithms are straightforward, each has its strength: adequate randomness and high privacy guarantees. Then, based on these two algorithms, we propose a novel privacy-preserving data sharing method with both of their advantages, (ε, k)-Randomized Anonymization. We theoretically guarantee that all of these three algorithms satisfy ε-differential privacy. In addition, we describe the analysis procedures using an inverse matrix and an expectation-maximization (EM) algorithm.

2. In the experiments, we evaluated each method's anonymity, privacy level, and output accuracy by using real data. The first method, k-anonymization followed by randomized response, reduces the anonymity of the output, but provides the highest accuracy. The second method, randomized response followed by k-anonymization, can achieve stronger privacy guarantees, although the accuracy is lower due to the high randomness. (ε, k)-Randomized Anonymization, which combines the above two methods, can provide both high privacy assurance and accuracy without compromising anonymity. These results indicate that (ε, k)-Randomized Anonymization is a novel privacy-preserving method for data sharing that can achieve high utility while satisfying both k-anonymity and ε-differential privacy.

In Section 2, we briefly review the related work. In Section 3, we describe the preliminary definitions for this study. In Section 4, we propose new privacy-preserving data sharing methods that satisfy both k-anonymity and ε-differential privacy. In Section 5, we evaluate our proposed methods by experiments using real data. In Section 6, we summarize our study with future direction. Python codes of our methods are available at https://github.com/ay0408/Randomized-Anonymization.

2 RELATED WORK

The most prominent concept for privacy-preserving data sharing is k-anonymity (Sweeney, 2002). This aims to prevent the identity disclosure of a targeted individual in a dataset. However, there is still a risk of identifying whether the target is in the dataset depending on the adversary's prior knowledge (Li et al., 2012). In contrast to k-anonymity, differential privacy (Dwork, 2006) is a framework to protect information about the presence of the target. In the concept of differential privacy, we can guarantee data privacy no matter what information the adversary knows. However, should the information that the dataset contains the target be leaked, the individual's identity could be revealed.

In this situation, it is desirable to develop new data sharing methods that can protect both the identity and presence of individuals, and there have been several studies (Li et al., 2012; Holohan et al., 2017; Tsou et al., 2021) to connect k-anonymity and differential privacy. First, Li *et al.* showed that k-anonymization can achieve differential privacy when preceded by random sampling (Li et al., 2012). Subsequently, Tsou *et al.* presented an anonymization method that satisfies k-anonymity and differential privacy by applying KD-tree in addition to random sampling (Tsou et al., 2021). These studies assume that we sample the data before releasing them. As yet, no method can release all the information in the original data while satisfying anonymity and privacy. Other methods proposed by Holohan *et al.* consider varying the approach depending on data types and apply k-anonymization to attribute data and differentially private methods to numerical data (Holohan et al., 2017), but the dataset as a whole is not completely privacy-preserving. This paper proposes novel methods to publish all the data while achieving both k-anonymity and differential privacy without sampling. Although our methods do not protect against attribute disclosure, they may achieve stronger privacy guarantees by combining with the existing work on the relationship between t-closeness and differential privacy (Domingo-Ferrer and Soria-Comas, 2015).

3 PRELIMINARIES

3.1 k-Anonymity

The concept of k-anonymity (Sweeney, 2002) was proposed for privacy-preserving microdata sharing. The k-anonymity requires that each tuple value of quasi-identifier (QI) attributes appears at least k times in a dataset, so that even if adversaries know a tuple value of a particular individual's QIs, they cannot uniquely identify the exact record of the individual. The following is the definition of k-anonymity.

Definition 1. (*k-anonymity (Sweeney, 2002)*)
Let T be a table and QI_T be the quasi-identifiers associated with it. T satisfies k-anonymity if and only if each tuple value in $T[QI_T]$ appears at least k times in $T[QI_T]$.

The k-anonymity protects against identity disclosure, but does not against a risk of identifying the presence of the individual in the dataset. Therefore, when we also aim to protect the information of the individual's participation, a stronger privacy guarantee is required.

3.2 ε-Differential Privacy

Differential privacy (Dwork, 2006) was developed in the field of cryptography as a framework that allows statistical analysis of databases while preserving personal data in the database from adversaries. Unlike k-anonymity, differential privacy can protect the information about whether the individual is in the dataset or not. The idea of differential privacy is based on the fact that it should be almost impossible to distinguish between two *neighboring* datasets differing in just one record. The privacy level in differential privacy is evaluated by the parameter $\varepsilon > 0$. Smaller ε values achieve stronger privacy guarantees but reduce the utility of the data. The following is the definition of ε-differential privacy.

Definition 2. (ε-*Differential Privacy (Dwork, 2006)*) *A randomized mechanism M satisfies* ε-*differential privacy if, for any $S \subset$ range M and all neighboring datasets D and D',* $\Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S]$.

One main mechanism to satisfy ε-differential privacy is the Laplace mechanism (Dwork et al., 2006), which only adds a random noise according to the *sensitivity* of the function to the original data and outputs private values. The definition of the *sensitivity* is as follows.

Definition 3. (*Sensitivity for the Laplace Mechanism (Dwork et al., 2006)*)
The sensitivity of a function $f : \mathcal{D}^M \to \mathbb{R}^d$ is

$$\Delta f = \max_{D,D'} ||f(D) - f(D')||_1,$$

where $D, D' \in \mathcal{D}^M$ are neighboring datasets.

Releasing $f(D) + b$ satisfies ε-differential privacy when b is random noise derived from a Laplace distribution with mean 0 and scale $\frac{\Delta f}{\varepsilon}$ (Dwork et al., 2006).

The Laplace mechanism is highly practical when the *sensitivity* is small, as in the case of histogram publication (Meng et al., 2017). However, if the *sensitivity* is large compared to the original value or the dataset consists of discrete values, the outputs may be less accurate. Therefore, in this study, we consider to satisfy ε-differential privacy by the technique of randomized response, which randomly perturbs each individual's attribute values.

3.2.1 Randomized Response

Randomized response was first introduced by Warner (Warner, 1965) to encourage survey participants to answer sensitive questions truthfully. This mechanism was shown to be differentially private (Dwork and Roth, 2014) and has been well used for hypothesis testing (Gaboardi and Rogers, 2018) and crowdsourcing (Erlingsson et al., 2014). In the following, we describe the randomized response approach in the case where all the participants in a dataset is divided into $m\,(\geq 2)$ mutually exclusive and exhaustive classes.

The randomized response with m classes follows an $m \times m$ distortion matrix:

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix},$$

where $p_{uv} = \Pr[x' = u | x = v]$ $(u, v \in \{1, 2, \cdots, m\})$ denotes the probability that the randomized output is u when the real class of the participant is v. Here, the sum of probabilities of each column is 1. When the following inequality holds:

$$\varepsilon \geq \max_{u=1,2,\cdots,m} \frac{\max_{v=1,2,\cdots,m} p_{uv}}{\min_{v=1,2,\cdots,m} p_{uv}},$$

the randomized response satisfies ε-differential privacy (Wang et al., 2016).

4 METHODS

In this study, we propose new privacy-preserving methods for medical data sharing that satisfy both k-anonymity and ε-differential privacy. First, we present two algorithms that apply the k-anonymization method and randomized response in sequence, and discuss the advantages and disadvantages of each. Then, we propose a novel method, (ε, k)-Randomized Anonymization, based on the first two algorithms. This method is expected to achieve both high accuracy and strong privacy assurance of the output. Also, we theoretically prove that each algorithm satisfies differential privacy. Furthermore, we describe the procedures for conducting statistical analysis using the published data.

4.1 k-Anonymization → Randomized Response

First, we present an algorithm that performs k-anonymization on the original table followed by the randomized response.

Algorithm 1: k-Anonymization → Randomized Response.

Input: A table with QIs, privacy parameters k and ε.
Output: A k-anonymized and ε-differentially private table.

1: Group input tuples of QIs into clusters, s.t., each cluster has at least k tuples.
2: Let m be the number of clusters and c_i be the cluster to which each tuple i belongs.
3: Construct an $m \times m$ distortion matrix $\mathbf{P}$, and perform the randomized response for each c_i according to $\mathbf{P}$. Let $\tilde{c}_i$ be the randomized output from c_i.
4: For each tuple i, replace the QI values with the representative values of $\tilde{c}_i$.

The distortion matrix $\mathbf{P}$ in Algorithm 1 satisfies the following equation to maximize the sum of the diagonal components (Wang et al., 2016):

$$\mathbf{P}_{uv} = \begin{cases} \frac{e^{\varepsilon}}{e^{\varepsilon}+m-1} & (u = v) \\ \frac{1}{e^{\varepsilon}+m-1} & (u \neq v) \end{cases}.$$

We provide a privacy guarantee of Algorithm 1 by Theorem 1.

Theorem 1. *Algorithm 1 Satisfies ε-differential Privacy.*

Proof. Let T be the output table from Algorithm 1, and g, g' be input tables differing in a tuple of one individual. We let $\mathcal{A}$ be the mechanism represented by Algorithm 1 and show

$$\Pr[\mathcal{A}(g) = T] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(g') = T].$$

Here, we denote the procedure of applying the randomized response to a cluster c by a function RR. Let c_i be the cluster containing tuple i, then the following equation holds:

$$\begin{aligned} & \Pr[\mathcal{A}(g) = T] \\ = \; & \Pr[RR(c_{g_0}) = c_{T_0}] \cdot \Pr[RR(c_{g_1}) = c_{T_1}] \\ & \cdots \Pr[RR(c_{g_{|T|-1}}) = c_{T_{|T|-1}}] \;\; (\neq 0). \end{aligned}$$

Suppose that the j-th tuple is different in g and g', then

$$\frac{\Pr[\mathcal{A}(g) = T]}{\Pr[\mathcal{A}(g') = T]} = \frac{\Pr[RR(c_{g_j}) = c_{T_j}]}{\Pr[RR(c_{g'_j}) = c_{T_j}]}. \quad (1)$$

Since the elements of the distortion matrix in Algorithm 1 are $\frac{e^{\varepsilon}}{e^{\varepsilon}+m-1}$ or $\frac{1}{e^{\varepsilon}+m-1}$, we can show $(1) \leq e^{\varepsilon}$. Therefore, Algorithm 1 satisfies ε-differential privacy. □

When analyzing the data using the output from this algorithm, we first create a vector $d \in \mathbb{N}^m$ representing the number of elements in each cluster. Then, we recover the original distribution by $\tilde{d} = \mathbf{P}^{-1} d$. Here, the elements of $\mathbf{P}^{-1}$ are as follows:

$$(\mathbf{P}^{-1})_{uv} = \begin{cases} \frac{e^{\varepsilon}+m-2}{e^{\varepsilon}-1} & (u = v) \\ \frac{-1}{e^{\varepsilon}-1} & (u \neq v) \end{cases}.$$

After that, change the negative elements of $\tilde{d}$ to 0, and finally, calculate $\tilde{d} \times \frac{||d||_1}{||\tilde{d}||_1}$ so that the sum of elements of $\tilde{d}$ becomes equal to that of d.

In addition, we can also use an expectation-maximization (EM) algorithm to reconstruct $\tilde{d}$ following some existing studies (Fanti et al., 2016; Ye et al., 2019). Unlike the above procedure using $\mathbf{P}^{-1}$, the EM algorithm has an advantage that it does not output negative numbers. The detailed procedure is as follows:

i. **Initialization:**
Let s be the number of individuals in the dataset. Create $x \in \mathbb{R}^{s \times m}$ s.t.

$$x_{h,i} = \begin{cases} 1 & \left(\sum_{j=0}^{i-1} d_j \leq h < \sum_{j=0}^{i} d_j\right) \\ 0 & (\text{otherwise}) \end{cases}$$

Set $\theta_0^0 = \theta_1^0 = \cdots = \theta_{m-1}^0 = \frac{1}{m}$. (This is a uniform distribution and m is the number of clusters.)

ii. **e-Step:**
For any individual h ($0 \leq h < s$) and any cluster i ($0 \leq i < m$),

$$\begin{aligned} \theta_{h,i}^k &= \Pr[z_{h,i} = 1 | x_{h,i}] \\ &= \frac{\Pr[x_{h,i} | z_{h,i} = 1] \cdot \Pr[z_{h,i} = 1]}{\sum_{j=0}^{m-1} \Pr[x_{h,i} | z_{h,j} = 1] \cdot \Pr[z_{h,j} = 1]} \\ &= \frac{\Pr[x_{h,i} | z_{h,i} = 1] \cdot \theta_i^{k-1}}{\sum_{j=0}^{m-1} \Pr[x_{h,i} | z_{h,j} = 1] \cdot \theta_j^{k-1}}. \end{aligned}$$

iii. **m-Step:**

$$\theta_i^k = \frac{1}{s} \sum_{h=0}^{s-1} \theta_{h,i}^k$$

iv. Repeat steps ii and iii until $\sum_i \left| \theta_i^k - \theta_i^{k-1} \right| < \delta$ for some $\delta > 0$, then calculate $\tilde{d} = s \cdot \theta^k$.

Here, $z_{h,i}$ in the EM algorithm is unobserved data and satisfies the following equations:

$$\begin{aligned} \Pr[x_{h,i} | z_{h,i}] &= \begin{cases} \frac{e^{\varepsilon}}{e^{\varepsilon}+m-1} & (x_{h,i} = 1) \\ \frac{m-1}{e^{\varepsilon}+m-1} & (x_{h,i} = 0) \end{cases} \\ \Pr[x_{h,i} | z_{h,j}] &= \begin{cases} \frac{1}{e^{\varepsilon}+m-1} & (x_{h,i} = 1) \\ \frac{e^{\varepsilon}+m-2}{e^{\varepsilon}+m-1} & (x_{h,i} = 0) \end{cases} \quad (i \neq j). \end{aligned}$$

The computational complexity of the E-Step is $O(sm^2)$, so when the dataset size is large or the anonymity parameter k is small, it could take a much longer time than when using $\mathbf{P}^{-1}$.

Algorithm 1 is the first method for privacy-preserving medical data sharing that satisfies both anonymity and differential privacy. One drawback of this method is that the output table does not strictly satisfy k-anonymity. For a large k, anonymity of the dataset is expected to be little compromised, but for small values of k, more accurate algorithm is desired.

4.2 Randomized Response → k-Anonymization

The next algorithm is to perform randomized response first, then k-anonymization. Unlike Algorithm 1, the output exactly satisfies k-anonymity.

Algorithm 2: Randomized Response → k-Anonymization.

Input: A table with QIs, privacy parameters k and ε.
Output: A k-anonymized and ε-differentially private table.

1: Let X be the set of possible tuples of QIs and n be the size of X.
2: Let $t_i \in X$ be the i-th tuple value.
3: Construct an $n \times n$ distortion matrix $\mathbf{P}$, and perform the randomized response for each t_i according to $\mathbf{P}$.
4: Group the randomized tuples into clusters, s.t., each cluster has at least k tuples.
5: Let c_i be the cluster to which each tuple i belongs.
6: For each tuple i, replace the QI values with the representative values of c_i.

Similar to Algorithm 1, the distribution matrix $\mathbf{P}$ satisfies the following equation:

$$\mathbf{P}_{uv} = \begin{cases} \frac{e^{\varepsilon}}{e^{\varepsilon}+n-1} & (u = v) \\ \frac{1}{e^{\varepsilon}+n-1} & (u \neq v) \end{cases}.$$

Here, for privacy assurance of Algorithm 2 and data analysis using the output, we consider the probability of the i-th tuple belonging to cluster c_i in the output table. We let $\hat{c}_i$ be the cluster to which tuple i should belong based on the input data, and r_j be the number of possible tuple values that cluster j can contain. Then, the probability that $\hat{c}_i$ changes to c_i is as follows:

$$\Pr[\hat{c}_i \to c_i] = \begin{cases} \frac{e^{\varepsilon}+(r_{c_i}-1)}{e^{\varepsilon}+n-1} & (\hat{c}_i = c_i) \\ \frac{r_{c_i}}{e^{\varepsilon}+n-1} & (\hat{c}_i \neq c_i) \end{cases}.$$

Using this probability, the privacy guarantee of Algorithm 2 is shown by Theorem 2.

Theorem 2. *Algorithm 2 Satisfies* ε*-differential Privacy.*

Proof. Similar to the proof of Theorem 1, we consider the following equation:

$$\begin{aligned} & \Pr[\mathcal{A}(g) = T] \\ = \; & \Pr[c_{g_0} \to c_{T_0}] \cdot \Pr[c_{g_1} \to c_{T_1}] \\ & \cdots \Pr[c_{g_{|T|-1}} \to c_{T_{|T|-1}}]. \end{aligned}$$

When the j-th tuple is different in g and g',

$$\frac{\Pr[\mathcal{A}(g) = T]}{\Pr[\mathcal{A}(g') = T]} = \frac{\Pr[c_{g_j} \to c_{T_j}]}{\Pr[c_{g'_j} \to c_{T_j}]}. \tag{2}$$

Here, using the probability above, the following inequalities hold:

$$(2) \leq \frac{e^{\varepsilon} + (r_{c_{g_j}} - 1)}{r_{c_{g_j}}} = 1 + \frac{e^{\varepsilon} - 1}{r_{c_{g_j}}} \leq e^{\varepsilon}.$$

Therefore, Algorithm 2 satisfies ε-differential privacy. In particular, when $\min_j r_{c_{g_j}} \geq 2$, this algorithm achieves a truly higher privacy guarantee than ε. □

When analyzing the data based on the output from Algorithm 2, we consider the following $m \times m$ matrix $\mathbf{Q}$:

$$\mathbf{Q} = \frac{1}{e^{\varepsilon}+n-1} \begin{pmatrix} e^{\varepsilon}+r_0-1 & r_0 & \cdots & r_0 \\ r_1 & e^{\varepsilon}+r_1-1 & \cdots & r_1 \\ \vdots & \vdots & \ddots & \vdots \\ r_{m-1} & r_{m-1} & \cdots & e^{\varepsilon}+r_{m-1}-1 \end{pmatrix},$$

where m is the number of clusters, and note that $\sum_i r_i = n$. Similar to the case of Algorithm 1, we first create a vector $d \in \mathbb{N}^m$ representing the number of elements in each cluster, then reconstruct the original distribution using $\mathbf{Q}^{-1}$. The elements of $\mathbf{Q}^{-1}$ are as follows:

$$(\mathbf{Q}^{-1})_{uv} = \begin{cases} \frac{e^{\varepsilon}+n-r_u-1}{e^{\varepsilon}-1} & (u = v) \\ \frac{-r_u}{e^{\varepsilon}-1} & (u \neq v) \end{cases}.$$

When using the EM algorithm, we can follow the same procedure as Algorithm 1 and the unobserved data $z_{h,i}$ satisfies the following equations:

$$\Pr[x_{h,i}|z_{h,i}] = \begin{cases} \frac{e^{\varepsilon}+r_i-1}{e^{\varepsilon}+n-1} & (x_{h,i} = 1) \\ \frac{n-r_i}{e^{\varepsilon}+n-1} & (x_{h,i} = 0) \end{cases}$$

$$\Pr[x_{h,i}|z_{h,j}] = \begin{cases} \frac{r_i}{e^{\varepsilon}+n-1} & (x_{h,i} = 1) \\ \frac{e^{\varepsilon}+n-r_i-1}{e^{\varepsilon}+n-1} & (x_{h,i} = 0) \end{cases} \quad (i \neq j).$$

Algorithm 2 can guarantee that the output table is k-anonymized and, moreover, achieve a truly stronger privacy guarantee than ε. Therefore, Algorithm 2 is superior to Algorithm 1 in terms of anonymity and privacy protection, but when n is much larger than the number of clusters, the accuracy is expected to decrease because the randomness of the output increases.

4.3 (ε, k)-Randomized Anonymization

Finally, we propose a novel method that combines the moderate randomness of Algorithm 1 and high anonymity and privacy guarantees of Algorithm 2. In this method, we perform $k'(< k)$-anonymization first, and then apply the randomized response to the anonymized data. After that, we perform k-anonymization on the randomized tuples. The detailed algorithm is shown in Algorithm 3.

Algorithm 3: (ε, k)-Randomized Anonymization.

Input: A table with QIs, privacy parameters k', k and ε.

Output: A k-anonymized and ε-differentially private table.

1: Group input tuples of QIs into clusters, s.t., each cluster has at least k' tuples.
2: Let m' be the number of clusters and c'_i be the cluster to which each tuple i belongs.
3: Construct an $m' \times m'$ distortion matrix $\mathbf{P}$, and perform the randomized response for each c'_i according to $\mathbf{P}$. Let $\tilde{c'}_i$ be the randomized output from c'_i.
4: For each tuple i, replace the QI values with the representative values of $\tilde{c'}_i$.
5: Group the randomized tuples into clusters, s.t. each cluster has at least k tuples.
6: Let c_i be the cluster to which each tuple i belongs.
7: For each tuple i, replace the QI values with the representative values of c_i.

The distribution matrix $\mathbf{P}$ in Algorithm 3 satisfies the following equation:

$$\mathbf{P}_{uv} = \begin{cases} \frac{e^{\varepsilon}}{e^{\varepsilon}+m'-1} & (u = v) \\ \frac{1}{e^{\varepsilon}+m'-1} & (u \neq v) \end{cases}.$$

Similar to the case of Algorithm 2, the probability that $\hat{c}_i$ changes to c_i through Algorithm 3 can be expressed as follows:

$$\Pr[\hat{c}_i \to c_i] = \begin{cases} \frac{e^{\varepsilon}+(r'_{c_i}-1)}{e^{\varepsilon}+m'-1} & (\hat{c}_i = c_i) \\ \frac{r'_{c_i}}{e^{\varepsilon}+m'-1} & (\hat{c}_i \neq c_i) \end{cases},$$

where r'_j is the number of possible randomized tuple values that cluster j can contain. Using this probability, we can show the privacy guarantee of Algorithm 3 by Theorem 3.

Theorem 3. *Algorithm 3 satisfies ε-differential privacy.*

Proof. Similar to the proof of Theorem 2, the following inequalities hold:

$$\frac{\Pr[\mathcal{A}(g) = T]}{\Pr[\mathcal{A}(g') = T]} \leq \frac{e^{\varepsilon} + (r'_{c_{g_j}} - 1)}{r'_{c_{g_j}}} = 1 + \frac{e^{\varepsilon} - 1}{r'_{c_{g_j}}} \leq e^{\varepsilon}.$$

Therefore, Algorithm 3 satisfies ε-differential privacy and when $\min_j r'_{c_{g_j}} \geq 2$, the privacy guarantee is truly higher than ε. □

When analyzing the data based on the output from Algorithm 3, we consider an $m \times m$ matrix $\mathbf{Q}'$ whose elements are as follows:

$$\mathbf{Q'}_{uv} = \begin{cases} \frac{e^{\varepsilon}+r'_u-1}{e^{\varepsilon}+m'-1} & (u = v) \\ \frac{r'_u}{e^{\varepsilon}+m'-1} & (u \neq v) \end{cases}.$$

Then, we can analyze the data based on the recovered distribution $\tilde{d} = \mathbf{Q'}^{-1} d$, like in the previous cases.
If we use the EM algorithm, consider the unobserved data $z_{h,i}$ satisfying the following equations:

$$\Pr[x_{h,i}|z_{h,i}] = \begin{cases} \frac{e^{\varepsilon}+r'_i-1}{e^{\varepsilon}+m'-1} & (x_{h,i} = 1) \\ \frac{m'-r'_i}{e^{\varepsilon}+m'-1} & (x_{h,i} = 0) \end{cases}$$

$$\Pr[x_{h,i}|z_{h,j}] = \begin{cases} \frac{r_i}{e^{\varepsilon}+m'-1} & (x_{h,i} = 1) \\ \frac{e^{\varepsilon}+m'-r'_i-1}{e^{\varepsilon}+m'-1} & (x_{h,i} = 0) \end{cases} \quad (i \neq j).$$

5 EXPERIMENTS AND DISCUSSION

In the experiments, we used the data provided by Japan Medical Association Medical Information Management Organization (Control Number: 2021-3). Medical ethical approval was obtained from Anonymized Medical Data Provision Review Board of Japan Medical Association Medical Information Management Organization.

Using the provided data, we first examined the characteristics of our proposed methods, including the anonymity of Algorithm 1 and the privacy level of Algorithms 2 and 3. Then, we conducted an age distribution analysis for a disease and measured the difference between the original data and the analysis results obtained from our methods to show their utility. Furthermore, we present several examples of analysis results using our methods and demonstrate that the results are roughly identical to the original data while satisfying both k-anonymity and ε-differential privacy.

5.1 Data Description

We used the data on diseases in this experiment. The data size is $1,512,673$. The data contains four at-

tributes that can be regarded as QIs: Medical Institution Code, Consultation Date, Sex, and Age. Medical Institution Code is from 1 to 38, Consultation Date is from April 1, 2020 to April 30, 2021, Sex is M or F and Age is from 0 to 105. We considered performing k-anonymization based on these attributes. In the following, we show the k-anonymization method for the experiment.

5.1.1 k-Anonymization

First, we represent the set of QIs of each individual as a single integer score. Note that the score and the set of QIs are in one-to-one correspondence. Then, we can satisfy k-anonymity by grouping the individuals whose scores are close to each other. Finally, by replacing each individual's score with a representative value of the group, a k-anonymized table can be output. The detailed algorithm is shown in Algorithm 4.

This k-anonymization method does not mask any of the QI attributes. Therefore, analyses for all the QI attributes can be performed in the similar way, and the representative values of each cluster can be calculated easily for our proposed methods. However, the score calculation in Algorithm 4 may result in poor accuracy of the QI information in the lower bits (i.e., c_i or d_i) because individuals with different values tend to be clustered in the same group. Should the analysis purpose and the use of the data be known in detail, a better k-anonymization method would exist, so this is one important problem for the future work.

Algorithm 4: k-Anonymization method for the experiment.

Input: A table with QIs (Medical Institution Code, Consultation Date, Sex, and Age) and privacy parameter k.

Output: A k-anonymized table.

1: Let C, D, S, and A be the number of possible values of Medical Institution Code, Consultation Date, Sex, and Age, respectively.
2: Rewrite each value of QIs as a natural number in the range of $[0, C-1]$, $[0, D-1]$, $[0, S-1]$, and $[0, A-1]$, respectively.
3: Denote the set of QIs of the individual i by the following SCORE_i:
$\mathrm{SCORE}_i = c_i + C \cdot d_i + C \cdot D \cdot s_i + C \cdot D \cdot S \cdot a_i$,
where c_i, d_i, s_i, and a_i are the QIs of individual i.
4: Let n_j be the number of individuals with a score j.
5: Partition the distribution of scores so that each group contains k or more individuals by the following procedure:
$m = 0, t = 0$
Let p be a vector representing the maximum score in each group.
for j **in** $0, 1, \ldots, C \cdot D \cdot S \cdot A - 1$ **do**
$m = m + n_j$
if $m \geq k$ **do**
$p_t = j; t = t + 1; m = 0$
if $m < k$ **do**
$p_{t-1} = C \cdot D \cdot S \cdot A - 1$
6: Replace SCORE_i of each individual i with the representative value of the group in which the score is included. When the score is in the group g, the replaced SCORE_i can be calculated as follows:
$\mathrm{SCORE}_i = \lfloor \frac{p_{g-1} + p_g}{2} \rfloor$.

5.2 Results

In this subsection, we evaluate the utility of the proposed methods in terms of anonymity, privacy level, and accuracy of the analysis results.

5.2.1 Anonymity of Algorithm 1

Algorithm 1 applies randomized response after k-anonymization, so the anonymity of the output table becomes less than or equal to k. In this experiment, we varied the values of k (from 10 to 100) and ε (from 1 to 20), the inputs to Algorithm 1, to measure the output anonymity. The results are plotted in Figure 1.

The results show that the output anonymity is roughly proportional to the input k, and the rate of decrease in anonymity is almost independent of k. Regarding the effect of ε, a larger ε preserves higher anonymity. However, as ε value increases, the privacy guarantee under differential privacy decreases and the information about the presence of individuals is more likely to be revealed. Therefore, Algorithm 1 has a strong trade-off between anonymity and privacy level, requiring a larger ε for stronger anonymity and a smaller ε for a stronger privacy assurance.

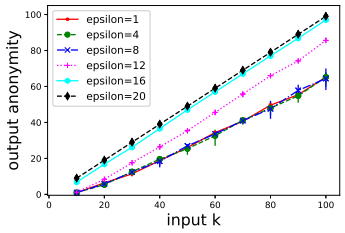

Figure 1: Anonymity of the output by Algorithm 1 when varying the input k and ε.

5.2.2 Privacy Level of Algorithms 2 and 3

Unlike Algorithm 1, the output tables from Algorithms 2 and 3 are truly k-anonymized, and stronger privacy guarantees than the input ε can be provided. In this experiment, we varied the values of ε (from 1 to 20) and k (from 10 to 200) and measured the output privacy level. The results are shown in Figure 2.

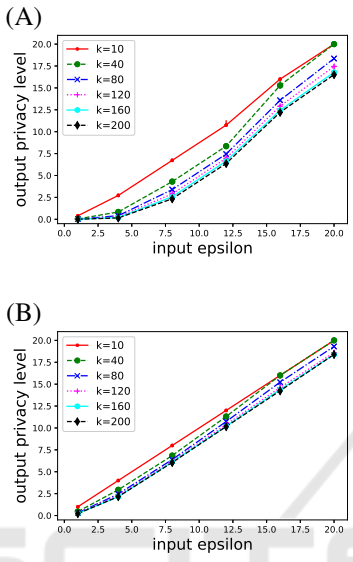

Figure 2: Privacy level of the output by Algorithms 2 (A) and 3 (B) when varying the input ε and k.

As for Algorithm 2, a smaller value of ε makes the privacy assurance stronger. However, the added perturbation also increases, which may lead to poor output accuracy. As for Algorithm 3, the privacy levels are relatively stable for any ε compared to Algorithm 2, and the accuracy of Algorithm 3 is expected to be higher than that of Algorithm 2. Regarding the effect of k, a higher k can provide a stronger privacy guarantee. Therefore, when using Algorithms 2 and 3, we can enhance both anonymity and privacy guarantee by increasing k.

5.2.3 Accuracy

Next, to evaluate the output accuracy from each algorithm, we compared the original data and the analysis results by our methods for the age distribution of those diagnosed with gastritis. Here, we use an inverse matrix to analyze data for quick execution. Because the data size used in this study was large (1,512,673) and the number of QIs was small (4), we can set relatively large values as the anonymity parameter k. In this experiment, we considered four cases of $k = 100, 400, 800$, and $1,600$. To measure the difference between the original age distribution and the analysis results from our methods, we use the KL divergence (Kullback and Leibler, 1951). The definition is as follows.

Definition 4. (*KL Divergence (Kullback and Leibler, 1951)*)
For discrete probability distributions p and q defined on the same probability space X, the KL divergence is defined by $D_{\mathrm{KL}}(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$.

In this experiment, we let q be the original distribution. A smaller KL divergence indicates that p and q are closer together. The results are plotted in Figure 3.

The results in all the cases show a similar trend of accuracy when varying the value of ε. When the value of k increased, the input values were easily perturbed by grouping in k-anonymization, resulting in lower accuracy.

Then, we will discuss each algorithm. Algorithm 2 (k-RR) has too strong a privacy guarantee when ε is small as we showed in Figure 2, so the output accuracy was poor. Algorithm 1 (RR-k) provided the highest accuracy because the output privacy level does not change from ε, but we should note that anonymity is reduced. Algorithm 3 (RA) also maintains high accuracy without compromising anonymity and with stronger privacy guarantees than ε. These results indicate that our Algorithm 3, (ε, k)-Randomized Anonymization, can provide high-quality results in all aspects of anonymity, privacy guarantee, and output accuracy.

5.3 Examples

As examples of analysis results, we examined the distributions of those diagnosed with gastritis. Here, we performed both the analysis using an inverse matrix and EM algorithm. The values of k and ε were set to 800 and 16, respectively. We analyzed the age distribution and consultation month, and the results are plotted in Figures 4 and 5.

These results show that using either an inverse matrix or EM algorithm, the analysis results are equivalent to the original distribution, indicating high utility of our methods.

6 CONCLUSION

In this study, we proposed new privacy-preserving methods for data sharing that satisfy both k-anonymity and ε-differential privacy. Our methods have the advantage that they do not assume data sampling and can release all the information in

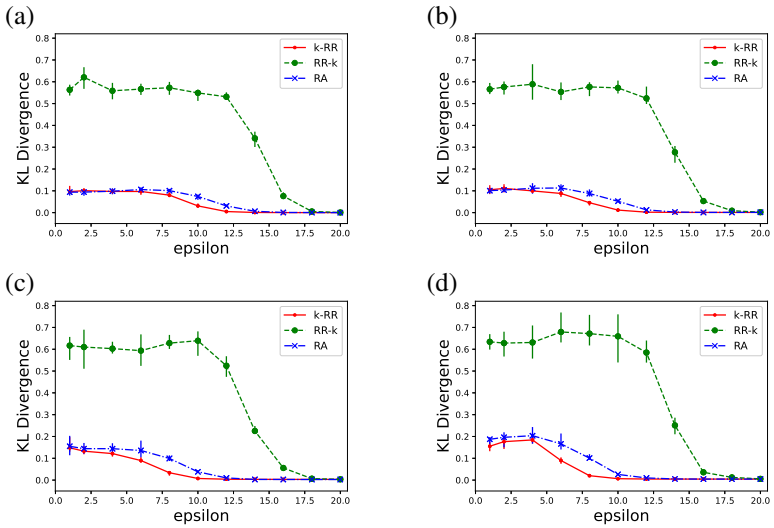

Figure 3: KL Divergence between the original age distribution and the analysis results from our methods when (a) $k = 100$, (b) $k = 400$, (c) $k = 800$, and (d) $k = 1{,}600$. k-RR (red, solid), RR-k (green, dashed), and RA (blue, dash-dot) represent our Algorithms 1, 2, and 3, respectively.

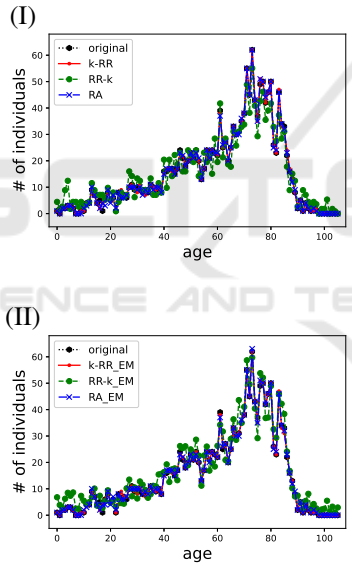

Figure 4: Age distributions of those diagnosed with gastritis when using an inverse matrix (I) and EM algorithm (II).

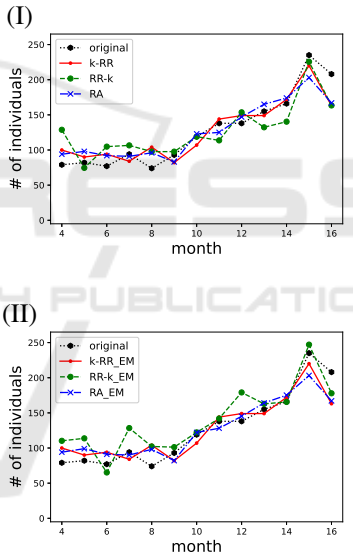

Figure 5: Consultation month distributions (from April 2020 to April 2021) of those diagnosed with gastritis when using an inverse matrix (I) and EM algorithm (II).

the original data. In particular, our third method, (ε, k)-Randomized Anonymization, is a novel method that can achieve a stronger privacy guarantee than ε while truly satisfying k-anonymity. Also, the experiments using real data show that (ε, k)-Randomized Anonymization can provide highly accurate results close to the original data. Not only data sharing methods, we also described two analysis procedures: one using an inverse matrix and the other using an EM algorithm.

An important future work is the development of k-anonymization methods suited for integration with the randomized response technique. The optimized methods may strongly depend on the data usage and analysis purposes, so we plan to explore this problem continuously. Furthermore, combination with the concept of t-closeness and the use of RAPPOR (Erlingsson et al., 2014) instead of randomized response will also be beneficial. We hope that this study will help in free sharing of biomedical and healthcare data throughout the world in the future.

ACKNOWLEDGEMENTS

This research was supported by Health Labour Sciences Research Grant 21AC1001.

REFERENCES

Aziz, M. M. A., Sadat, M. N., Alhadidi, D., Wang, S., Jiang, X., Brown, C. L., and Mohammed, N. (2019). Privacy-preserving techniques of genomic data-a survey. *Brief Bioinform.*, 20(3):887–895.

Cummings, R. and Desai, D. (2018). The role of differential privacy in GDPR compliance.

Domingo-Ferrer, J. and Soria-Comas, J. (2015). From t-closeness to differential privacy and vice versa in data anonymization. *Knowl.-Based Sys.*, 74:151–158.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I., editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Halevi, S. and Rabin, T., editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

Emam, K. E. and Dankar, F. K. (2008). Protecting privacy using k-anonymity. *J. Am. Med. Inform. Assoc.*, 15(5):627–637.

Erlingsson, U., Pihur, V., and Korolova, A. (2014). RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 1054–1067, New York, NY, USA. Association for Computing Machinery.

European Commission (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Fanti, G., Pihur, V., and Erlingsson, U. (2016). Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies (PoPETS)*, issue 3, 2016.

Ficek, J., Wang, W., Chen, H., Dagne, G., and Daley, E. (2021). Differential privacy in health research: A scoping review. *Journal of the American Medical Informatics Association*, 28(10):2269–2276.

Gaboardi, M. and Rogers, R. (2018). Local private hypothesis testing: Chi-square tests. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1626–1635. PMLR.

Hlávka, J. P. (2020). Chapter 10 - security, privacy, and information-sharing aspects of healthcare artificial intelligence. In Bohr, A. and Memarzadeh, K., editors, *Artificial Intelligence in Healthcare*, pages 235–270. Academic Press.

Holohan, N., Antonatos, S., Braghin, S., and Aonghusa, P. M. (2017). (k,ε)-anonymity: k-anonymity with ε-differential privacy. *arXiv: Cryptography and Security*.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Lee, H., Kim, S., Kim, J. W., and Chung, Y. D. (2017). Utility-preserving anonymization for health data publishing. *BMC Med. Inform. Decis. Mak.*, 17:104.

Li, N., Qardaji, W., and Su, D. (2012). On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '12, page 32–33, New York, NY, USA. Association for Computing Machinery.

Meng, X., Li, H., and Cui, J. (2017). Different strategies for differentially private histogram publication. *J. Commun. Inf. Netw.*, 2:68–77.

Rakesh Kumar, S., Gayathri, N., Muthuramalingam, S., Balamurugan, B., Ramesh, C., and Nallakaruppan, M. (2019). Chapter 13 - medical big data mining and processing in e-healthcare. In Balas, V. E., Son, L. H., Jha, S., Khari, M., and Kumar, R., editors, *Internet of Things in Biomedical Engineering*, pages 323–339. Academic Press.

Su, J., Cao, Y., Chen, Y., Liu, Y., and Song, J. (2021). Privacy protection of medical data in social network. *BMC Med. Inform. Decis. Mak.*, 21:286.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570.

Tsou, Y.-T., Alraja, M. N., Chen, L.-S., Chang, Y.-H., Hu, Y.-L., Huang, Y., Yu, C.-M., and Tsai, P.-Y. (2021). (k,ε,δ)-anonymization: Privacy-preserving data release based on k-anonymity and differential privacy. *Serv. Oriented Comput. Appl.*, 15(3):175–185.

Wang, Y., Wu, X., and Hu, D. (2016). Using randomized response for differential privacy preserving data collection. In Palpanas, T. and Stefanidis, K., editors, *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops 2016, Bordeaux, France, March 15, 2016*, volume 1558 of *CEUR Workshop Proceedings*.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Wu, W.-T., Li, Y.-J., Feng, A.-Z., Li, L., Huang, T., Xu, A.-D., and Lyu, J. (2021). Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Military Med. Res.*, 8:44.

Ye, Y., Zhang, M., Feng, D., Li, H., and Chi, J. (2019). Multiple privacy regimes mechanism for local differential

privacy. In Li, G., Yang, J., Gama, J., Natwichai, J., and Tong, Y., editors, *Database Systems for Advanced Applications*, pages 247–263, Cham. Springer International Publishing.