# Supplements: Privacy-Optimized Randomized Response for Sharing Multi-Attribute Data

In this document, we first review related studies and detailed the differences between them and this study. Then, we provide concrete descriptions of our methods, along with some proofs. For preliminaries and generalized methods, please refer to our main paper. Furthermore, we show additional experimental results on analysis example and discussion.

## S1. Related Work

### A. Differentially Private Methods for Attribute Data

While attribute data play an important role in crowdsourcing and medical data mining, privacy protection in their publication is needed. In particular, various differentially private methods have been proposed for analyzing attribute data [18], and their use for hypothesis testing [6], for example, has also been considered. The most common mechanisms include the Laplace mechanism [3] and randomized response [17], and various variants have been proposed [1], [4], [9], [10], [15]. Among these, the randomized response has been shown to be of highest utility in several existing studies [6], [16], and this study aims to further deepen this mechanism. We concentrate on publishing categorical data itself, which does not require encoding the input data into another format with no specific purpose of analysis; therefore, mechanisms with encoding or hashing are not addressed in this study. In particular, major differences from RAPPOR-related studies [1], [4], [12], [15] are that we envision situations in which not only the summary results of analysis, but also each individual's attribute information, can be observed, and that we can set the respective privacy level to each attribute information.

The most basic randomized response is one that protects the single attribute information of each individual. In this regard, several studies have provided optimal methods in terms of utility [7], [16], with the smallest error in the perturbed data. However, there is a lack of studies on methods for data with multiple attributes, such as medical and healthcare data. Moreover, it has been pointed out that compliance with the EU's GDPR requires that these data be published while satisfying differential privacy in addition to anonymity properties [2]; therefore, developing the randomized response for sharing multi-attribute data is an urgent issue.

### B. Randomized Response for Sharing Multi-Attribute Data

The existing method for sharing multi-attribute data uses the Kronecker product and perturbs each attribute information independently and sequentially [11], [16]. However, the achieved privacy level for the entire dataset is far from optimal and cannot provide sufficient privacy assurance. In fact, a recent study has proposed methods for $2 \times 2$ and $3 \times 2$ tabular data that can achieve stronger privacy guarantees by perturbing the two-attribute information collectively [19], but there is still no method for other multi-attribute data. Therefore, in this study, we propose a privacy-optimized randomized response for sharing general $k$-attribute data that can achieve the strongest privacy level for the entire dataset, when the privacy level for each attribute information is given. This, conversely, indicates that when the privacy level for the entire dataset is given, we can distribute more privacy budget to each attribute information, improving accuracy in actual data analysis.

## S2. Methods

In this study, we first present a procedure for constructing a distortion matrix for the privacy-optimized randomized response for sharing multi-attribute data. This task can be regarded as a linear programming problem with respect to matrix elements, and we can formulate it as a minimization problem. Then, we propose an efficient heuristic method using an inductive algorithm. In this document, we describe the methods for three-attribute data in detail. The proofs (skipped in our main paper) are provided in the next section.

### A. Formulation as Linear Programming Problem

We describe a linear programming problem for obtaining the optimal matrix for three-attribute data. When the number of possible values for the three attributes is $a_1$, $a_2$, and $a_3$, respectively, the size of the distortion matrix $\mathbf{P}$ for the entire data is $a_1a_2a_3 \times a_1a_2a_3$. The possible values for the elements of $\mathbf{P}$ are the following eight, each representing the probability of events that hold the respective condition:

$X_0$ : Input and output data are exactly the same;
$X_1$ : Only the first attribute value differs;
$X_2$ : Only the second attribute value differs;
$X_3$ : Only the third attribute value differs;
$X_4$ : The first and second attribute values differ;
$X_5$ : The first and third attribute values differ;
$X_6$ : The second and third attribute values differ;
$X_7$ : All attribute values are different.

In each column and row of the matrix, these values appear 1, $a_1 - 1$, $a_2 - 1$, $a_3 - 1$, $(a_1 - 1)(a_2 - 1)$, $(a_1 - 1)(a_3 - 1)$,

$(a_2-1)(a_3-1)$, and $(a_1-1)(a_2-1)(a_3-1)$ times, respectively. Note that the total value of these times is indeed $a_1a_2a_3$. Regarding the inequality relations among these elements, to output closer data to the true value with a higher probability, the following should hold:

$$\begin{aligned}&X_0 \geq X_1,\ X_0 \geq X_2,\ X_0 \geq X_3,\\ &X_1 \geq X_4, X_5,\ X_2 \geq X_4, X_6,\ X_3 \geq X_5, X_6,\\ &X_4, X_5, X_6 \geq X_7.\end{aligned}$$

Then, the privacy level satisfied from $\mathbf{P}$ is calculated as

$$\epsilon = \ln \frac{X_0}{X_7}.$$

Because we aim to minimize this value while maintaining the privacy level of each attribute information, we can consider minimizing $x_0$, where $x_j = X_j/X_7$ for $j = 0, 1, \ldots, 7$. Here, $x_7 = X_7/X_7 = 1$, and there are seven variables involved in the linear programming problem, from $x_0$ to $x_6$. Based on the above discussion, when the privacy level of each attribute information is $\epsilon_1$, $\epsilon_2$, and $\epsilon_3$, respectively, the following three equations hold:

$$\begin{aligned}&\frac{x_0 + (a_2-1)x_2 + (a_3-1)x_3 + (a_2-1)(a_3-1)x_6}{x_1 + (a_2-1)x_4 + (a_3-1)x_5 + (a_2-1)(a_3-1)} = e^{\epsilon_1},\\ &\frac{x_0 + (a_1-1)x_1 + (a_3-1)x_3 + (a_1-1)(a_3-1)x_5}{x_2 + (a_1-1)x_4 + (a_3-1)x_6 + (a_1-1)(a_3-1)} = e^{\epsilon_2},\\ &\frac{x_0 + (a_1-1)x_1 + (a_2-1)x_2 + (a_1-1)(a_2-1)x_4}{x_3 + (a_1-1)x_5 + (a_2-1)x_6 + (a_1-1)(a_2-1)} = e^{\epsilon_3}.\end{aligned}$$

In summary, the linear programming problem for finding the optimized matrix for three-attribute data can be formulated as follows:

$$\begin{aligned}\min_{\mathbf{x}} \quad & (1,0,0,0,0,0,0)\cdot\mathbf{x}\\ \text{s.t.} \quad & \mathbf{A}_{ub}\cdot\mathbf{x} \leq \mathbf{b}_{ub},\\ & \mathbf{A}_{eq}\cdot\mathbf{x} = \mathbf{b}_{eq},\end{aligned}$$

where

$$\mathbf{x}^\mathsf{T} = (x_0, x_1, x_2, x_3, x_4, x_5, x_6),$$

$$\mathbf{A}_{ub} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0\\ -1 & 0 & 1 & 0 & 0 & 0 & 0\\ -1 & 0 & 0 & 1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0 & 1 & 0 & 0\\ 0 & -1 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & -1 & 0 & 1 & 0 & 0\\ 0 & 0 & -1 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & -1 & 0 & 1 & 0\\ 0 & 0 & 0 & -1 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & -1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & -1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix},$$

$$\mathbf{b}_{ub}^\mathsf{T} = (0,0,0,0,0,0,0,0,0,-1,-1,-1),$$

$$\mathbf{A}_{eq}^\mathsf{T} = \begin{pmatrix} 1 & 1 & 1\\ -e^{\epsilon_1} & a_1-1 & a_1-1\\ a_2-1 & -e^{\epsilon_2} & a_2-1\\ a_3-1 & a_3-1 & -e^{\epsilon_3}\\ -(a_2-1)e^{\epsilon_1} & -(a_1-1)e^{\epsilon_2} & (a_1-1)(a_2-1)\\ -(a_3-1)e^{\epsilon_1} & (a_1-1)(a_3-1) & -(a_1-1)e^{\epsilon_3}\\ (a_2-1)(a_3-1) & -(a_3-1)e^{\epsilon_2} & -(a_2-1)e^{\epsilon_3} \end{pmatrix},$$

$$\mathbf{b}_{eq} = \begin{pmatrix} (a_2-1)(a_3-1)e^{\epsilon_1}\\ (a_1-1)(a_3-1)e^{\epsilon_2}\\ (a_1-1)(a_2-1)e^{\epsilon_3} \end{pmatrix}.$$

By generalizing the above discussion, we describe a linear programming problem for obtaining the privacy-optimized distortion matrix for $k$-attribute data in our main paper.

### *B. Heuristic Method*

Here, we describe the procedure to find a near-optimal solution for three-attribute data using the optimal solution for two-attribute data.

Let $\epsilon_1$, $\epsilon_2$, and $\epsilon_3$ be the privacy levels and $a_1$, $a_2$, and $a_3$ be the numbers of possible attribute values. Here, we consider the following five values that appear in the distortion matrix:

$$\begin{aligned}X_{3,0} &: \text{Input and output data are exactly the same;}\\ X_{3,1} &: \text{Only the first attrubute value differs;}\\ X_{3,2} &: \text{Only the second attrubute value differs;}\\ X_{3,3} &: \text{Only the third attrubute value differs;}\\ X_{3,4} &: \text{More than one attribute values differ,}\end{aligned}$$

by assuming all probabilities of events that hold the condition where two or more attribute values differ to be equal. Then, there are four variables to consider:

$$x_{3,0} = \frac{X_{3,0}}{X_{3,4}},\ x_{3,1} = \frac{X_{3,1}}{X_{3,4}},\ x_{3,2} = \frac{X_{3,2}}{X_{3,4}},\ x_{3,3} = \frac{X_{3,3}}{X_{3,4}},$$

and the values of $e^{\epsilon_1}$, $e^{\epsilon_2}$, and $e^{\epsilon_3}$ are calculated as

$$\begin{aligned}&\frac{x_{3,0} + (a_2-1)x_{3,2} + (a_3-1)x_{3,3} + a_2a_3 - (1 + a_2 - 1 + a_3 - 1)}{x_{3,1} + a_2a_3 - 1},\\ &\frac{x_{3,0} + (a_1-1)x_{3,1} + (a_3-1)x_{3,3} + a_1a_3 - (1 + a_1 - 1 + a_3 - 1)}{x_{3,2} + a_1a_3 - 1},\\ &\frac{x_{3,0} + (a_1-1)x_{3,1} + (a_2-1)x_{3,2} + a_1a_2 - (1 + a_1 - 1 + a_2 - 1)}{x_{3,3} + a_1a_2 - 1},\end{aligned}$$

respectively. Among these, $e^{\epsilon_1}$ and $e^{\epsilon_2}$ can also be calculated using $x_{2,0}$, $x_{2,1}$, and $x_{2,2}$; therefore, the following relations should hold:

$$\begin{cases} \frac{x_{2,0}+(a_2-1)x_{2,2}}{x_{2,1}+a_2-1} = \frac{x_{3,0}+(a_2-1)x_{3,2}+(a_3-1)x_{3,3}+a_2a_3-(1+a_2-1+a_3-1)}{x_{3,1}+a_2a_3-1}\\ \frac{x_{2,0}+(a_1-1)x_{2,1}}{x_{2,2}+a_1-1} = \frac{x_{3,0}+(a_1-1)x_{3,1}+(a_3-1)x_{3,3}+a_1a_3-(1+a_1-1+a_3-1)}{x_{3,2}+a_1a_3-1} \end{cases} \quad (1)$$

Here, if

$$\begin{cases} x_{3,0} + (a_3 - 1)x_{3,3} = a_3 \cdot x_{2,0} \\ x_{3,1} + a_3 - 1 = a_3 \cdot x_{2,1} \\ x_{3,2} + a_3 - 1 = a_3 \cdot x_{2,2} \end{cases}, \tag{2}$$

the relations in (1) hold. The proof is provided below.

*Proof.* When the relations in (2) hold,

$$\begin{aligned} & x_{3,0} + (a_2 - 1)x_{3,2} + (a_3 - 1)x_{3,3} \\ & \quad + a_2 a_3 - (1 + a_2 - 1 + a_3 - 1) \\ = \; & a_3 \cdot x_{2,0} + (a_2 - 1)\,(a_3 \cdot x_{2,2} - (a_3 - 1)) \\ & \quad + a_2 a_3 - a_2 - a_3 + 1 \\ = \; & a_3 \cdot x_{2,0} + a_3 \cdot (a_2 - 1)x_{2,2} \\ & \quad - (a_2 - 1)(a_3 - 1) + (a_2 - 1)(a_3 - 1) \\ = \; & a_3\,(x_{2,0} + (a_2 - 1)x_{2,2}) \end{aligned}$$

and

$$\begin{aligned} x_{3,1} + a_2 a_3 - 1 &= a_3 \cdot x_{2,1} - (a_3 - 1) + a_2 a_3 - 1 \\ &= a_3(x_{2,1} + a_2 - 1). \end{aligned}$$

Therefore, for the first equality in (1), the right-hand side is

$$\frac{a_3\,(x_{2,0} + (a_2 - 1)x_{2,2})}{a_3(x_{2,1} + a_2 - 1)} = \frac{x_{2,0} + (a_2 - 1)x_{2,2}}{x_{2,1} + a_2 - 1},$$

which is identical to the left-hand side. Similarly, for the second equality, the right-hand side equals the left-hand side when the relations in (2) hold. □

Therefore, for the values from $x_{3,0}$ to $x_{3,3}$ to satisfy the equality relations for privacy levels, we first obtain $x_{3,1}$ and $x_{3,2}$ as

$$\begin{aligned} x_{3,1} &= a_3 \cdot x_{2,1} - a_3 + 1 \\ \text{and} \quad x_{3,2} &= a_3 \cdot x_{2,2} - a_3 + 1 \end{aligned}$$

from the second and third equality relations in (2). Thereafter, using these values, we can solve the following simultaneous linear equations:

$$\begin{cases} x_{3,0} + (a_3 - 1)x_{3,3} = a_3 \cdot x_{2,0} \\ \frac{x_{3,0} + (a_1 - 1)x_{3,1} + (a_2 - 1)x_{3,2} + a_1 a_2 - (1 + a_1 - 1 + a_2 - 1)}{x_{3,3} + a_1 a_2 - 1} = e^{\epsilon_3} \end{cases}$$

and obtain the values of $x_{3,0}$ and $x_{3,3}$.

From the above procedure, all values from $x_{3,0}$ to $x_{3,3}$ can be derived using the values from $x_{2,0}$ to $x_{2,2}$, the optimal solution for the two-attribute case. By constructing a distortion matrix from these values, a randomized response mechanism for three-attribute data that satisfies a near-optimal privacy guarantee for the entire dataset while maintaining the privacy level for each attribute information is expected to be provided.

By generalizing the method for three-attribute data, we can construct a method for $k$-attribute data and present it in our main paper.

## S3. PROOFS

*1) Optimal Mechanism for $m \times n$ Data:* We provide the proof that the solution in our main paper is optimal; that is, the obtained value of $x_{2,0}$ is the minimum under

$$\begin{cases} \frac{x_{2,0} + (n-1) \cdot x_{2,2}}{x_{2,1} + (n-1)} = e^{\epsilon_1} \\ \frac{x_{2,0} + (m-1) \cdot x_{2,1}}{x_{2,2} + (m-1)} = e^{\epsilon_2} \end{cases} \\ \text{and} \quad x_{2,0} \geq x_{2,1}, x_{2,2} \geq 1. \tag{3}$$

*Proof.*

$$\begin{aligned} (3) &\iff \begin{cases} x_{2,0} + (n-1) \cdot x_{2,2} = e^{\epsilon_1} \cdot x_{2,1} + (n-1) \cdot e^{\epsilon_1} \\ x_{2,0} + (m-1) \cdot x_{2,1} = e^{\epsilon_2} \cdot x_{2,2} + (m-1) \cdot e^{\epsilon_2} \end{cases} \\ &\iff \begin{cases} x_{2,1} = \frac{e^{\epsilon_2} + n - 1}{e^{\epsilon_1} + m - 1} \cdot x_{2,2} + \frac{(m-1) \cdot e^{\epsilon_2} - (n-1) \cdot e^{\epsilon_1}}{e^{\epsilon_1} + m - 1} \\ x_{2,1} = \frac{e^{\epsilon_2}}{m-1} \cdot x_{2,2} + e^{\epsilon_2} - \frac{x_{2,0}}{m-1} \end{cases}. \end{aligned}$$

When $\frac{e^{\epsilon_2}}{m-1} \geq \frac{e^{\epsilon_2} + n - 1}{e^{\epsilon_1} + m - 1} \iff e^{\epsilon_1 + \epsilon_2} \geq (m-1)(n-1)$, the values of $x_{2,1}$ and $x_{2,2}$ decrease as $x_{2,0}$ decreases. Because $x_{2,1}, x_{2,2} \geq 1$, $x_{2,0}$ is minimized when $x_{2,1} = 1$ or $x_{2,2} = 1$. When $x_{2,1} = 1$,

$$x_{2,2} = \frac{n \cdot e^{\epsilon_1} - (m-1)(e^{\epsilon_2} - 1)}{e^{\epsilon_2} + n - 1}.$$

Therefore, when

$$\frac{n \cdot e^{\epsilon_1} - (m-1)(e^{\epsilon_2} - 1)}{e^{\epsilon_2} + n - 1} \geq 1 \iff n(e^{\epsilon_1} - 1) \geq m(e^{\epsilon_2} - 1),$$

we can obtain the optimal solution for case (I). $x_{2,0} \geq x_{2,2}$ also holds because

$$x_{2,0} - x_{2,2} = \frac{n(e^{\epsilon_2} - 1)(e^{\epsilon_1} + m - 1)}{e^{\epsilon_2} + n - 1} \geq 0.$$

When $n(e^{\epsilon_1} - 1) < m(e^{\epsilon_2} - 1)$, $x_{2,0}$ is minimized when $x_{2,2} = 1$, and we can obtain the solution for case (II). As in the above case, $x_{2,0} \geq x_{2,1}$ also certainly holds.

When $e^{\epsilon_1 + \epsilon_2} < (m-1)(n-1)$, $x_{2,1}$ and $x_{2,2}$ increase as $x_{2,0}$ decreases. Because $x_{2,1}, x_{2,2} \leq x_{2,0}$, $x_{2,0}$ is minimized when $x_{2,1} = x_{2,0}$ or $x_{2,2} = x_{2,0}$. When $x_{2,1} = x_{2,0}$, $(x_{2,0}, x_{2,2})$ is

$$\left( \frac{(n-1)(e^{\epsilon_1} + m - 1)e^{\epsilon_2}}{-e^{\epsilon_1 + \epsilon_2} + e^{\epsilon_2} + m(n-1)}, \frac{m(n-1)e^{\epsilon_1} + (m-1)(e^{\epsilon_1} - 1)e^{\epsilon_2}}{-e^{\epsilon_1 + \epsilon_2} + e^{\epsilon_2} + m(n-1)} \right).$$

When $x_{2,2} = x_{2,0}$, $(x_{2,0}, x_{2,1})$ is

$$\left( \frac{(m-1)e^{\epsilon_1}(e^{\epsilon_2} + n - 1)}{-e^{\epsilon_1 + \epsilon_2} + e^{\epsilon_1} + (m-1)n}, \frac{(m-1)ne^{\epsilon_2} + (n-1)e^{\epsilon_1}(e^{\epsilon_2} - 1)}{-e^{\epsilon_1 + \epsilon_2} + e^{\epsilon_1} + (m-1)n} \right).$$

Given $x_{2,0} \geq x_{2,1}, x_{2,2}$, we can obtain the optimal solutions for cases (III) and (IV). In the both cases, $x_{2,1}, x_{2,2} \geq 1$ certainly holds as in cases (I) and (II). □

*2) Inductive Method for $k$-Attribute Data:*
If

$$\begin{cases} x_{k,0} + (a_k - 1)x_{k,k} = a_k \cdot x_{k-1,0} \\ x_{k,j} + a_k - 1 = a_k \cdot x_{k-1,j} \quad (j = 1, 2, \ldots, k-1) \end{cases}, \tag{4}$$

the relation

$$\frac{x_{k-1,0}+A(k-1,j)+B(k-1,j)}{x_{k-1,j}+C(k-1,j)} = \frac{x_{k,0}+A(k,j)+B(k,j)}{x_{k,j}+C(k,j)}. \quad (5)$$

holds, where

$$A(k,j) := \sum_{i=1}^{k}(a_i-1)x_{k,i}-(a_j-1)x_{k,j},$$

$$B(k,j) := \frac{\prod_{i=1}^{k}a_i}{a_j}-\left(\sum_{i=1}^{k}a_i-a_j-k+2\right),$$

$$\text{and} \quad C(k,j) := \frac{\prod_{i=1}^{k}a_i}{a_j}-1.$$

We provide the proof below.

*Proof.* When the relations in (4) hold,

$$\begin{aligned}
& x_{k,0}+A(k,j)+B(k,j)\\
= \;& x_{k,0}+(a_k-1)x_{k,k}+\sum_{i=1}^{k-1}(a_i-1)x_{k,i}-(a_j-1)x_{k,j}\\
& +\frac{\prod_{i=1}^{k}a_i}{a_j}-\left(\sum_{i=1}^{k}a_i-a_j-k+2\right)\\
= \;& a_k\cdot x_{k-1,0}+\sum_{i=1}^{k-1}(a_i-1)(a_k\cdot x_{k-1,i}-(a_k-1))\\
& -(a_j-1)(a_k\cdot x_{k-1,j}-(a_k-1))\\
& +a_k\cdot\frac{\prod_{i=1}^{k-1}a_i}{a_j}-\left(a_k+\sum_{i=1}^{k-1}a_i\right)+a_j+k-2\\
= \;& a_k\cdot x_{k-1,0}+a_k\cdot\sum_{i=1}^{k-1}(a_i-1)x_{k-1,i}-a_k\cdot\sum_{i=1}^{k-1}(a_i-1)\\
& +\sum_{i=1}^{k-1}(a_i-1)-a_k\cdot(a_j-1)x_{k-1,j}+(a_k-1)(a_j-1)\\
& +a_k\cdot\frac{\prod_{i=1}^{k-1}a_i}{a_j}-a_k-\sum_{i=1}^{k-1}a_i+a_j+k-2\\
= \;& a_k\cdot x_{k-1,0}+a_k\cdot\sum_{i=1}^{k-1}(a_i-1)x_{k-1,i}-a_k\cdot\sum_{i=1}^{k-1}a_i\\
& +a_k\cdot(k-1)-(k-1)-a_k\cdot(a_j-1)x_{k-1,j}\\
& +a_k\cdot(a_j-1)-(a_j-1)+a_k\cdot\frac{\prod_{i=1}^{k-1}a_i}{a_j}\\
& -a_k+a_j+k-2\\
= \;& a_k\cdot\left(x_{k-1,0}+\sum_{i=1}^{k-1}(a_i-1)x_{k-1,i}-(a_j-1)x_{k-1,j}\right.\\
& \qquad \left.+\frac{\prod_{i=1}^{k-1}a_i}{a_j}-\left(\sum_{i=1}^{k-1}a_i-a_j-k+3\right)\right)\\
= \;& a_k\cdot\big(x_{k-1,0}+A(k-1,j)+B(k-1,j)\big)
\end{aligned}$$

and

$$\begin{aligned}
x_{k,j}+C(k,j) &= a_k\cdot x_{k-1,j}-(a_k-1)+\frac{\prod_{i=1}^{k}a_i}{a_j}-1\\
&= a_k\cdot x_{k-1,j}-a_k+a_k\cdot\frac{\prod_{i=1}^{k-1}a_i}{a_j}\\
&= a_k\cdot\left(x_{k-1,j}+\frac{\prod_{i=1}^{k-1}a_i}{a_j}-1\right)\\
&= a_k\cdot\big(x_{k-1,j}+C(k-1,j)\big).
\end{aligned}$$

Therefore, the right-hand side of (5) equals the left-hand side; that is, the relation (5) holds. □

## S4. Experiments and Discussion

Here, we provide an analysis example using genome statistics and demonstrate the utility of our method.

### A. Analysis Example

In this experiment, to verify the utility of our methods on datasets with a large $k$, we focused on our heuristic method that can be performed in $\mathcal{O}(k^2)$ time and compared its performance with that of the existing method.

In genomic statistical analysis, various statistical tests are often conducted to investigate the relationship between marker loci, such as SNPs, and diseases. The need to protect privacy in the publication of statistics obtained from these tests has been pointed out by several studies [8], [13], [14], and the application of differential privacy has recently been well considered [5], [19]. Here, we used the most common test, the $\chi^2$-test, as an example to show our method's utility when the privacy level for the entire dataset is fixed.

Using the information on each SNP $i$, the following $2\times 2$ contingency table can be constructed:

| | | Disease Status | | Total |
|---|---|---|---|---|
| | | 0 | 1 | |
| Allele | 0 | $A_i$ | $B_i$ | $A_i+B_i$ |
| | 1 | $C_i$ | $D_i$ | $C_i+D_i$ |
| Total | | $A_i+C_i$ | $B_i+D_i$ | $2N$ |

where $N$ is the number of individuals. The $\chi^2$-statistic obtained from the table is

$$\frac{2N\cdot(A_iD_i-B_iC_i)^2}{(A_i+B_i)(C_i+D_i)(A_i+C_i)(B_i+D_i)}.$$

Considering each allele information and disease status as an attribute value, we can employ the randomized response where the number of possible attribute values $(=a_i)$ is 4 as in the existing study [19]. Then, to share the entire dataset containing such attribute information on $k$ SNPs in total, we can construct a distortion matrix for $k$-attribute data using each privacy level $\epsilon_i$ for the information on SNP $i$ and $a_i=4$. Here, if the privacy level of the entire dataset is fixed, the value that can be set as $\epsilon_i$ is expected to be larger with our method than with the existing method. Consequently, the accuracy of the $\chi^2$-statistic calculated using the perturbed information is also expected to be higher.

In the following, we considered the case where the ratio among $\epsilon_i$ $(i = 1, 2, \ldots, k)$ is fixed, and measured the average difference between the original $\chi^2$-statistic and differentially private statistic after randomized response, while varying the value of $k$. Data for evaluation were generated by setting $N = 1,000$ and the values of $A_i$, $B_i$, $C_i$, and $D_i$ in the contingency table were randomly computed for each SNP $i$ as follows: $A_i = \text{Binomial}(2N, 1/3)$, $B_i = \text{Binomial}(2N - A_i, 1/3)$, $C_i = \text{Binomial}(2N - A_i - B_i, 2/5)$, and $D_i = 2N - A_i - B_i - C_i$. The results over 10 runs are shown in Fig. S1.

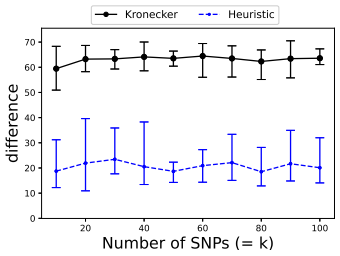


Fig. S1. Comparison of the accuracy of $\chi^2$-statistics between the existing Kronecker product-based method (black, solid) and our heuristic method (blue, dashed) when the privacy level for the entire datset is fixed. The $x$-axis represents the number of SNPs; that is, the number of $k$. The $y$-axis represents the average difference between the original and differentially private statistics. The error bar represents the range of all results.

In Fig. S1, the difference when using our heuristic method is almost less than half of that when using the existing method. This result indicates that our method can provide high accuracy in important data analysis. Considering the results in our main paper, it is expected that larger values of $\epsilon_i$ or smaller values of $a_i$ will have even increased impact.

## REFERENCES


[1] Acharya, J., Sun, Z., Zhang, H.: Hadamard response: Estimating distributions privately, efficiently, and with little communication. In: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. vol. 89, pp. 1120–1129. PMLR (2019)

[2] Cummings, R., Desai, D.: The role of differential privacy in GDPR compliance (2018)

[3] Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. S. Halevi and T. Rabin, (eds) Theory of Cryptography **3876**, 265–284 (2006)

[4] Erlingsson, U., Pihur, V., Korolova, A.: RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. pp. 1054–1067 (2014)

[5] Fienberg, S.E., Slavkovic, A., Uhler, C.: Privacy preserving GWAS data sharing. In: IEEE 11th International Conference on Data Mining Workshops. pp. 628–635 (2011)

[6] Gaboardi, M., Rogers, R.: Local private hypothesis testing: Chi-square tests. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. vol. 80, pp. 1626–1635 (2018)

[7] Holohan, N., Leith, D.J., Mason, O.: Optimal differentially private mechanisms for randomised response. IEEE Transactions on Information Forensics and Security **12**(11), 2726–2735 (2017)

[8] Jacobs, K.B., Yeager, M., Wacholder, S., Craig, D., Kraft, P., Hunter, D.J., Paschal, J., Manolio, T.A., Tucker, M., Hoover, R.N., Thomas, G.D., Chanock, S.J., Chatterjee, N.: A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nat. Genet. **41**(11), 1253–1257 (2009)

[9] Kairouz, P., Bonawitz, K., Ramage, D.: Discrete distribution estimation under local privacy. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. pp. 2436–2444 (2016)

[10] Kulkarni, T.: Answering range queries under local differential privacy. In: Proceedings of the 2019 International Conference on Management of Data. p. 1832–1834. SIGMOD '19, Association for Computing Machinery, New York, NY, USA (2019)

[11] Liu, C., Chen, S., Zhou, S., Guan, J., Ma, Y.: A general framework for privacy-preserving of data publication based on randomized response techniques. Information Systems **96**, 101648 (2021)

[12] Ren, X., Yu, C.M., Yu, W., Yang, S., Yang, X., McCann, J.A., Yu, P.S.: LoPub : High-dimensional crowdsourced data publication with local differential privacy. IEEE Transactions on Information Forensics and Security **13**(9), 2151–2166 (2018)

[13] Sankararaman, S., Obozinski, G., Jordan, M.I., Halperin, E.: Genomic privacy and limits of individual detection in a pool. Nat. Genet. **41**(9), 965–967 (2009)

[14] Wan, Z., Hazel, J.W., Clayton, E.W., Vorobeychik, Y., Kantarcioglu, M., Malin, B.A.: Sociotechnical safeguards for genomic data privacy. Nat. Rev. Genet. **23**(7), 429–445 (2022)

[15] Wang, T., Blocki, J., Li, N., Jha, S.: Locally differentially private protocols for frequency estimation. In: Proceedings of the 26th USENIX Conference on Security Symposium. p. 729–745 (2017)

[16] Wang, Y., Wu, X., Hu, D.: Using randomized response for differential privacy preserving data collection. In: Palpanas, T., Stefanidis, K. (eds.) Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops 2016, Bordeaux, France, March 15, 2016. vol. 1558 (2016)

[17] Warner, S.L.: Randomized response: A survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. **60**(309), 63–66 (1965)

[18] Xiong, X., Liu, S., Li, D., Cai, Z., Niu, X.: A comprehensive survey on local differential privacy. Security and Communication Networks p. 8829523 (2020)

[19] Yamamoto, A., Shibuya, T.: Privacy-preserving genomic statistical analysis under local differential privacy. In: Atluri, V., Ferrara, A.L. (eds.) Data and Applications Security and Privacy XXXVII. pp. 40–48. Springer Nature Switzerland, Cham (2023)