Privacy-preserving chi-squared test of independence for small samples

doi: 10.1186/s13040-021-00238-x

Sei and Ohsuga *BioData Mining* (2021) 14:6
https://doi.org/10.1186/s13040-021-00238-x

BioData Mining

RESEARCH Open Access

Privacy-preserving chi-squared test of independence for small samples

Yuichi Sei* and Akihiko Ohsuga

*Correspondence: seiuny@uec.ac.jp
The University of Electro-Communications, Tokyo, Japan

Abstract

Background: The importance of privacy protection in analyses of personal data, such as genome-wide association studies (GWAS), has grown in recent years. GWAS focuses on identifying single-nucleotide polymorphisms (SNPs) associated with certain diseases such as cancer and diabetes, and the chi-squared (χ^2) hypothesis test of independence can be utilized for this identification. However, recent studies have shown that publishing the results of χ^2 tests of SNPs or personal data could lead to privacy violations. Several studies have proposed anonymization methods for χ^2 testing with ϵ-differential privacy, which is the cryptographic community's de facto privacy metric. However, existing methods can only be applied to 2×2 or 2×3 contingency tables, otherwise their accuracy is low for small numbers of samples. It is difficult to collect numerous high-sensitive samples in many cases such as COVID-19 analysis in its early propagation stage.

Results: We propose a novel anonymization method (RandChiDist), which anonymizes χ^2 testing for small samples. We prove that RandChiDist satisfies differential privacy. We also experimentally evaluate its analysis using synthetic datasets and real two genomic datasets. RandChiDist achieved the least number of Type II errors among existing and baseline methods that can control the ratio of Type I errors.

Conclusions: We propose a new differentially private method, named RandChiDist, for anonymizing χ^2 values for an $I \times J$ contingency table with a small number of samples. The experimental results show that RandChiDist outperforms existing methods for small numbers of samples.

Keywords: Differentical privacy, Chi-squared testing, Privacy-preserving data mining

Introduction

Examining genes involves comparing several groups of genes [1, 2], with three or more groups possibly involved in several instances. Generally, statistical analyses such as the chi-squared (χ^2) test of independence are used to determine whether single-nucleotide polymorphisms (SNPs) can be considered significantly different. The findings from such analyses are frequently shared between researchers and government agencies to facilitate new discoveries.

A genome can contain sensitive information about an individual such as genetic disease factors and disease risk. Each person's genome is 99.9% identical, with the remaining 0.1% difference producing peoples' various characteristics. The variation among individuals at a single position in a genome is known as a SNP. A genome-wide association study (GWAS) is a method of analyzing the statistical relationship between SNPs and diseases by finding SNPs that are related to a specific disease. To accomplish this, χ^2 testing has been used. Homer et al. [3] reported that an attacker may be able to statistically determine whether someone is a member of a group with a specific disease if the attacker is familiar with the potential victim's SNPs and the aggregate allele frequencies within that specific disease group.

The underlying assumption that the attacker is familiar with the potential victim's SNPs, which can be obtained from a very small blood sample, is realistic because of the increasing availability in cost-effective genotyping services [4, 5]. Furthermore, Wang et al. [6] suggested that the allele frequency of the group SNP values can be determined from standard statistical data such as p-values or χ^2 values. Consequently, an anonymization procedure should always be applied to χ^2 values when publishing SNP datasets [6–8].

Data sharing in genemic research is very important [9]. To avoid such leakage of private information, we should execute a privacy protection mechanism on GWAS results. Existing studies add a relatively large amount of noise to GWAS results to protect privacy. However, our aim is to reduce the amount of noise while maintaining the same level of privacy protection. In other words, we can achieve the same level of privacy protection as existing studies with privacy-preserving χ^2 testing and increase the usefulness of GWAS results.

The recent GWAS analysis methods are not limited to only the chi-squared test. For example, mixed linear model based methods have been used. However, the chi-squared test is still an important analysis method.

Although other methods for GWAS exist, a lot of recent research papers employ the chi-squared test for GWAS, such as [10–13], which were published in 2019 or 2020. Furthermore, the chi-squared test is used in numerous papers on GWAS to analyze COVID-19 [14–18]. Thus, because the chi-squared test has been adopted in many cases, it is worth studying.

Other tests, such as Kruskal-Wallis test and Wilcoxon test, are also employed for GWAS [19, 20]. Couch et al. [21] proposed differentially private methods for these tests. Dealing with other tests in our research remains an issue to be addressed in future work.

The most influential privacy metric within the privacy community is ϵ-differential privacy [22], which has been intensively investigated [23–26]. Several researchers, such as Fienberg et al. [27], Uhlerop et al. [28], and Yu et al. [7], have suggested approaches to facilitate sharing of χ^2 values while conforming with ϵ-differential privacy parameters. However, these proposed methods are currently only applicable to 2×2 or 2×3 contingency tables. In other words, it is currently not possible to analyze contingency tables larger than 2×3. However, the requirement to analyze SNPs based on an $I \times J$ contingency table is crucial. For example, previous studies have evaluated higher degrees of freedom within a contingency table [8, 29]. However, these methods have relatively poor accuracy, particularly in cases with small sample populations. This condition applies to many situations where the sample sizes being considered can range from dozens to several hundred samples [30–32].

Although we live in an era of big data where datasets with a large number of samples are becoming available in many domains, obtaining sensitive information is still difficult due to privacy regulations such as General Data Protection Regulations (GDPR). Sensitive patient biomedical data cannot be shared without permission [33]. Moreover, there are a lot of rare diseases, and obtaining such information of the patients is very difficult [34, 35]. Further, it is difficult to collect a large number of samples when there is a need for rapid analysis for a new disease such as COVID-19. Someone might provide his or her sensitive information without any privatization schemes; however, more people would provide their sensitive information by conducting privatization schemes [36, 37]. Moreover, many studies [38–40] have considered contingency tables larger than 2×3. Therefore, private χ^2 testing for large contingency tables with small samples is an important problem.

In this paper, we propose a new method, named RandChiDist, for anonymizing χ^2 values for an $I \times J$ contingency table with a small number of samples, and we experimentally evaluate this method using real datasets. RandChiDist adds the minimized Laplace noise to the true χ^2 value based on the contingency table and controls the ratio of Type I errors (i.e., false positives). The evaluation uses the synthetic and real datasets, including two genomic datasets. The evaluation shows that RandChiDist can control the ratio of Type I errors strictly and can reduce Type II errors (i.e., false negatives) more than existing methods that can control the ratio of Type I errors. Several methods reduce Type II errors more than RandChiDist; however, the methods cannot control the ratio of Type I errors.

Several approaches exist for non-private χ^2 testing, and RandChiDist can be used to calculate the global sensitivity of the χ^2 value of the simplest χ^2 testing and to add noise to the χ^2 value based on the global sensitivity. Thus, the added noise is minimized according to the Laplace mechanism theorem [22].

The motivation of this paper is summarized as follows. Chi-squared test can be employed for various data analyses, such as the identification of SNPs associated with certain diseases; however, publishing the chi-squared value can lead to privacy leakage. Thus, we propose a privacy-preserving chi-squared testing algorithm for a small number of samples due to the difficulty in collecting a large number of samples of a rare disease or new disease.

In our research, samples of less than about 1,000 in number are considered as a small sample size.

The rest of this paper is organized as follows: "Preliminaries" section introduces χ^2 hypothesis test and differential privacy. "Related work" section discusses related work. "Proposed method" section presents our proposed method and "Evaluation" section presents the results of our simulations. "Discussion" Section discusses the evaluation results, and the need for adaptation to a large contingency table and a small sample. "Conclusion" section concludes the paper.

Preliminaries

χ^2 hypothesis test of Independence

We consider a contingency table with I rows and J columns. Let $[i, j]$ denote the ith row and jth column's cell of the table. $O_{i,j}$ represents the value of cell $[i, j]$, and $E_{i,j}$ represents the expected value of cell $[i, j]$.

Let $m_i = \sum_j O_{i,j}$, $s_j = \sum_i O_{i,j}$, and $n = \sum_i m_i = \sum_j s_j$. Table 1 provides an example of a contingency table.

Table 1 An example of case-control analysis

	(Combination of allele types of SNP1 and SNP2)				
	(major,major)	**(major,minor)**	**(minor,major)**	**(minor,minor)**	**Total**
Case	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$	$O_{1,4}$	m_1
Control	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$	$O_{2,4}$	m_2
Total	s_1	s_2	s_3	s_4	n

The χ^2 value is calculated as

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} V_{i,j}, \quad \text{where } V_{i,j} = \frac{(E_{i,j} - O_{i,j})^2}{E_{i,j}}, \quad E_{i,j} = s_j \cdot \frac{m_i}{n}. \tag{1}$$

We determine the significance level α (i.e., the probability of a Type I error occurring) and the null hypothesis H_0 in advance. We then calculate χ^2 based on Eq. (1) and determine whether to reject H_0 using the χ^2 *distribution table*. Thus, χ^2_ν represents the probability density function of the χ^2 distribution with ν degrees of freedom. The χ^2 distribution table presents the percentage point $P(\chi^2_\nu > x) = \alpha$ for several combinations of ν and α.

Privacy model

In recent years, ϵ-differential privacy [22] has been considered the de facto standard for privacy metrics [33, 41–43].

The privacy parameter ϵ reflects the privacy level, with a large ϵ value indicating a low privacy. We consider neighboring databases to represent two databases differing by a maximum of one record. The ϵ-differential privacy is defined as follows:

Definition 1 (ϵ-differential privacy) *Let D and D' be neighboring databases. A randomized mechanism $\mathcal{M}$ satisfies the ϵ-differential privacy if, for any D and D' and any subset of outputs $Y \subset Range(\mathcal{M})$, it holds that*

$$P(\mathcal{M}(D) \in Y) \leq e^{\epsilon} P(\mathcal{M}(D') \in Y). \tag{2}$$

The Laplace mechanism, which adds noise generated using a Laplace distribution, can satisfy Theorem 1 [22]. To explain this mechanism, we first outline the concept of global sensitivity.

Definition 2 (Global sensitivity) *Let f be a function $f : \mathcal{D} \to \mathbb{R}^d$, where $\mathcal{D}$ is a collection of databases. When f satisfies for any neighoring databases D and D'*

$$\Delta f = \max_{D,D'} ||f(D) - f(D')||_1, \tag{3}$$

the global sensitivity of f is Δf.

Theorem 1 (Laplace Mechanism [22]) *A randomized mechanism $\mathcal{M}$ realizes ϵ-differential privacy if $\mathcal{M}$ outputs $f(D) + Lap(\Delta f/\epsilon)$, where $Lap(\nu)$ returns independent Laplace random variables with scale parameter ν.*

Table 2 A contingency table

	Condition 1	Condition 2
Group 1	25	30
Group 2	20	25

Related work

In χ^2 testing, a contingency table such as Table 2 is used. This contingency table can be represented as Table 3, and Tables 2 and 3 are equivalent. In research on privacy-preserving χ^2 testing, databases such as those shown in Table 3 are considered. For example, Tables 3 and 4 are neighboring databases because the tables contain the same data with exception of one record

Yu et al. [7] demonstrated that the global sensitivity of the χ^2 value of 2×3 contingency tables can be calculated as

$$\Delta_Y = \frac{n^2}{m_1 m_2}\left(1 - \frac{1}{\max\{m_1, m_2\} + 1}\right), \tag{4}$$

if m_1 and m_2 are known (i.e., published).

Fienberg et al. [27] and Uhlerop et al. [28] demonstrated that if $m_1 = m_2$, the global sensitivity of the χ^2 value can be calculated as

$$\Delta_F = \frac{4n}{n+2}. \tag{5}$$

The global sensitivities, Δ_F and Δ_Y, have been shown to be optimal values. However, they can only be applied to 2×2 or 2×3 contingency tables.

Kakizaki et al. [44, 45] proposed a unit circle mechanism that can achieve a high degree of accuracy. However, they assumed only 2×2 contingency tables. Additionally, they did not publish the differentially private χ^2 value used in their method; however, they did publish the differentially private result of the χ^2 testing based on the given significance level, α. Therefore, if a data holder wants to publish the private χ^2 testing results of several α values (e.g., $\alpha = 0.05, 0.01, 0.005$, and 0.001), the data holder must independently execute the privacy mechanisms multiple times (e.g., three times). Following the composition theorem [46], if a privacy mechanism outputs K times based on ϵ-differential privacy, the resulting privacy level thus becomes $K\epsilon$ (i.e., the privacy level decreases). Moreover, Banerjee et al. [47] state that publishing P-value could be important for data analysis.

The aforementioned studies all assumed that m_i $(i = 1, \ldots, I)$ is not sensitive information. We can share each value of m_i without privatization schemes.

Gaboardi et al. [8] proposed several methods for arbitrary contingency tables. First, they show a straightforward method that does not add Laplace noise to the χ^2 value, but rather adds it to each cell of the contingency table with a global sensitivity of 2.

Table 3 A raw database

Pseudo ID	Group number & Condition number
1	1&2
2	1&1
3	2&2
...	...
100	2&1

Table 4 Database that contains the same data as Table 3 except with 3's data

Pseudo ID	Group number & Condition number
1	1&2
2	1&1
3′	2&1
...	...
100	2&1

In this paper, we name this method as RandCell. RandCell is also known as SNPpval, which was proposed by Jonson and Shmatikov [48]. The χ^2 value of the contingency table to which RandCell adds Laplace noise tends to be large, meaning that RandCell yields many false positives. Therefore, Gaboardi et al. proposed several other methods known as PrivIndep, MCIndep with Laplace mechanism, and MCIndep with Gaussian mechanism. They showed that MCIndep with Laplace mechanism had the best performance of their proposed methods. Hence, we describe MCIndep with Laplace mechanism in detail in this paper and refer to MCIndep with Laplace mechanism as MCIndep for simplicity.

MCIndep generates many contingency tables randomly based on m_i and s_j of the contingency table with added Laplace noise and compares their χ^2 values. The original contingency table can be considered to reject H_0 if the χ^2 value of the contingency table to which RandCell adds Laplace noise is greater than the top $\alpha \times 100\%$ of the generated contingency tables' χ^2 value. Other methods for (ϵ, δ)-differential privacy are proposed [49], which relaxes the ϵ-differential privacy as their privacy metric. We focus on ϵ-differential privacy in this paper, and applying our method to (ϵ, δ)-differential privacy is an issue to be addressed in future work.

Sei et al. [50] proposed several theorems for differentially private χ^2 testing, but there were no detailed proofs for the theorems and the equations provided in their study. Moreover, there were no experiments that evaluated the performance of χ^2 testing.

More recently, Gaboardi et al. [29] proposed χ^2 test algorithms (LocalNoiseIND, LocalExpIND, and LocalBitFlipIND) for privacy-preserving χ^2 testing of independence based on local differential privacy. LocalNoiseIND is also known as zCDP general chi-squared test, which was proposed by Kifer and Rogers [51]. In their paper, they showed that LocalExpIND had the best performance of the three methods for most parameter settings. These methods can be applied to arbitrary contingency tables, and address a local model of privacy and assume there is no trusted entity. In this paper, we assume that a trusted entity has all the raw data.

Canonne et al. [52] calculated the sample complexity bounds of an ϵ-differentially private test for distinguishing between two distributions. They also applied differentially private change-point detection. Their method is for a parametric setting that requires that the two distributions are perfectly known. In contrast, our method can be used for a nonparametric setting.

Csail et al. [53] proposed an algorithm for testing the closeness of two distributions in a private manner. Their algorithm can also test the independence of two random variables. However, execution for privacy-preserving χ^2 testing was not described.

Liu et al. [54] showed how ϵ influences the accuracy of differentially private hypothesis testing. They proposed a method to determine an appropriate value for ϵ that can be

useful for determining the ϵ value for our proposed algorithm; however, determining ϵ is outside the scope of our paper.

Couch et al. [21] proposed a differentially private hypothesis testing method for the Kruskal-Wallis test, Mann-Whitney test, Wilcoxon test, and one-sample t-test. This hypothesis testing method is not for nominal scale data, which are suitable for χ^2 testing, but rather for ordinal or interval scale data.

The methods for arbitrary $I \times J$ contingency tables have relatively poor accuracy, particularly in cases with small-sample populations. We show the comparison between existing methods and the proposed method in "Evaluation" section.

Adversarial model

The adversarial model is described as follows. The server has a database, and it wants to share the result of the chi-squared test with data analysts who are potential attackers. The attacker is considered to be a semi-honest entity, that is, the attacker follows the protocol between the server. However, the attacker might attempt to extract individual information from the result of the chi-squared test.

Proposed method

Overview

We propose RandChiDist, which adds Laplace noise to the χ^2 value obtained from a target contingency table. Calculating the Laplace noise to be added requires the global sensitivity of the $I \times J$ contingency table's χ^2 value. The method for calculating global sensitivity is described in 1.

Typically, the χ^2 distribution table is used to determine whether to reject H_0. However, RandChiDist adds noise to the χ^2 value, thus we need a modified χ^2 distribution table. The method for calculating this is described in 1. RandChiDist uses this table to determine whether to reject H_0. We consider bounding the Type I error to be at most α to be a hard constraint.

Our main symbols are summarized in Table 5.

Global sensitivity of χ^2 value

As was assumed in other studies, we assume that m_i $(i = 1, \ldots, I)$ is also provided to a data analyzer. We consider contingency tables D_1 and D_2, which are generated from neighboring databases. Because the neighboring databases differ by one record, their contingency tables differ by a maximum of two cells. The value of cell $[a, k]$ in table D_2 is greater than that of cell $[a, k]$ in D_1 by 1, and the value of cell $[a, l]$ in table D_2 is less than that of cell $[a, l]$ (s.t. $l \neq k$) in table D_1 by 1.

Table 5 Symbols

n	Number of samples
I	Number of rows of a table
J	Number of columns of a table
m_i	Total value of ith row
s_j	Total value of jth column
$O_{i,j}(D)$	Observed value of cell $[i, j]$ in database D
$E_{i,j}(D)$	Expected value of cell $[i, j]$ in database D

Because the values of $m_i (i = 1, \ldots, I)$ are released to the public, the collection of databases in Definition 2.2 only include databases that satisfy the released values of m_i, and the neighboring databases are elements of the collection. Therefore, the global sensitivity is calculated based on the neighboring databases that satisfy the released values of m_i.

Thus, we calculate the possible maximum value of the difference of χ^2 values between tables D_1 and D_2.

RandChiDist satisfies differential privacy by adding Laplace noise with global sensitivity because of Theorem 1. We thus propose RandChiDist, which adds Laplace noise with global sensitivity,

$$\Delta_R = \begin{cases} \frac{(m_\alpha + m_\beta)n}{m_\alpha(1+m_\beta)} & J \geq 3 \\ \frac{n^2}{m_\alpha(n-m_\alpha+1)} & J = 2, \end{cases} \tag{6}$$

where

$$\begin{aligned} \alpha &= \arg\min_i m_i \quad \text{and} \\ \beta &= \arg\min_{i \neq \alpha} m_i, \end{aligned} \tag{7}$$

to the calculated χ^2 value from (1). Here, we have the following theorem:

Theorem 2 *RandChiDist satisfies ϵ-differential privacy.*

Proof We prove that Δ_R is the global sensitivity of χ^2 of the $I \times J$ contingency table. We can then uphold Theorem 2 because RandChiDist adds $Lap(\Delta_R/\epsilon)$ to the original value based on the Laplace mechanism theorem (Theorem 1).

Let $O_{i,j}(D)$ denote the observed value of cell $[i, j]$ in database D and let $\chi^2(D)$ denote the χ^2 value of database D. Without a loss of generality, we consider neighboring databases D_1 and D_2, which satisfy the following equations:

$$\begin{cases} O_{a,k}(D_2) = O_{a,k}(D_1) + 1 \\ O_{a,l}(D_2) = O_{a,l}(D_1) - 1, \end{cases} \tag{8}$$

where k and l are arbitrary natural numbers satisfying $k, l \in \{1, \ldots, J\}$ and $k \neq l$.

From Proposition 2, when J is greater than or equal to 3 and we are given the value a, neighboring databases that satisfy the following constraints maximize the difference between the χ^2 values of tables D_1 and D_2 (see Fig. 1a).

$$\begin{aligned} &O_{a,k}(D_1) + 1 = O_{a,k}(D_2) = m_a \\ &O_{a,l}(D_1) - 1 = O_{a,l}(D_2) = 0 \\ &\quad O_{b,l}(D_1) = O_{b,l}(D_2) = m_b \text{ where } b \neq a \\ &\quad O_{i,k}(D_1) = 0 \text{ where } i \neq a \\ &\quad O_{i,l}(D_1) = 0 \text{ where } i \neq a, b \\ &\quad O_{i,j}(D_1) = O_{i,j}(D_2) = \text{arbitrary values that} \\ &\qquad\qquad \text{satisfy the constraint } \sum_j O_{i,j} = m_i \\ &\qquad\qquad \text{where } [i,j] \neq [a,k], [a,l], \text{ and } [b,l]. \end{aligned} \tag{9}$$

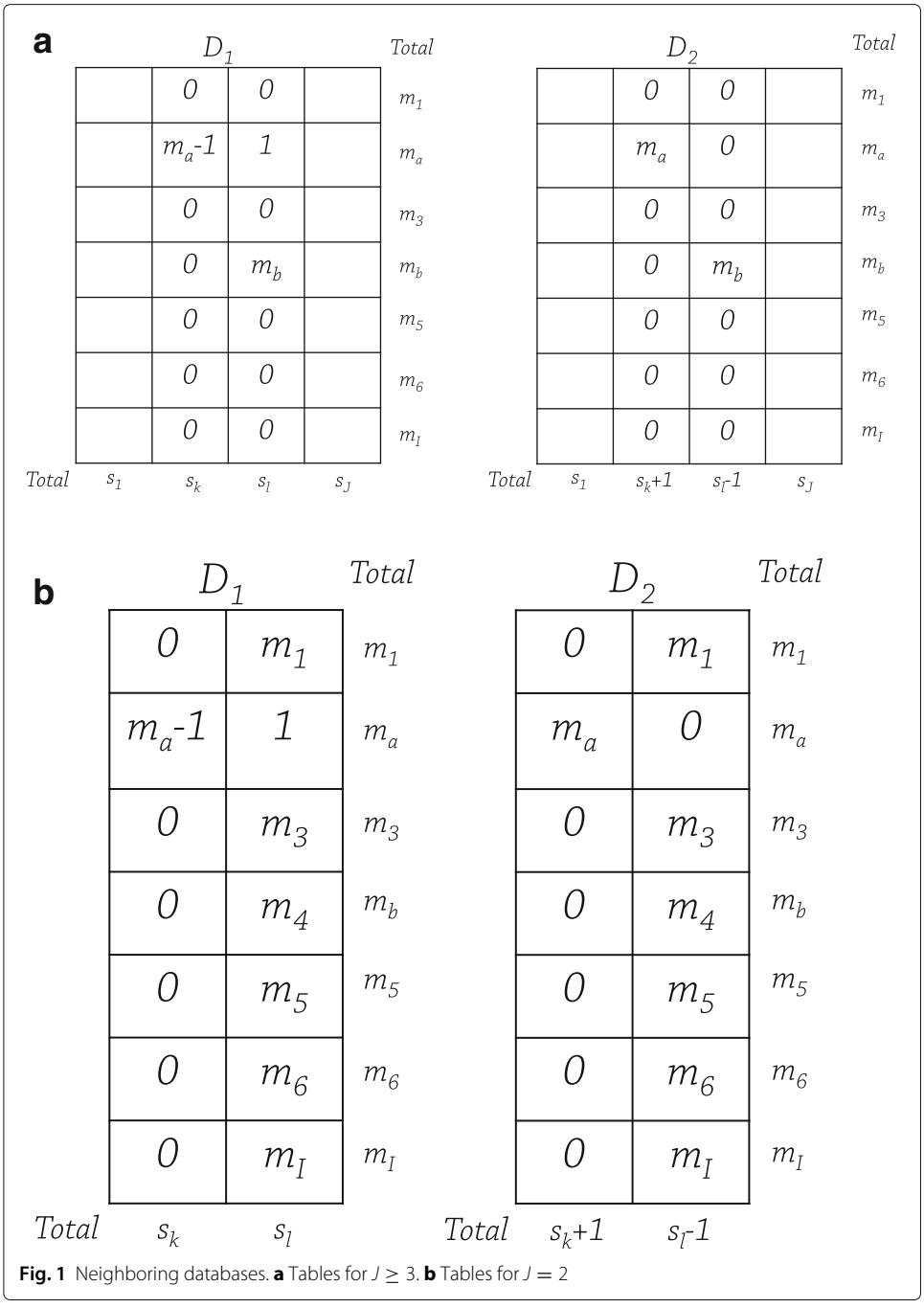

Fig. 1 Neighboring databases. **a** Tables for $J \geq 3$. **b** Tables for $J = 2$

From constraint (9), we understand that the sum of the kth column of D_2 (i.e., s_k of D_2) is equal to m_a.

Let $V_{i,j}(D)$ denote $V_{i,j}$ in Eq. (1) for database D.

The symbol b is an arbitrary integer from 1 to I but not a. The symbol l is an arbitrary integer from 1 to J but not k.

The difference between the χ^2 values of tables D_1 and D_2 that satisfies the constraint (9) is thus calculated by

$$\sum_i (V_{i,k}(D_2) + V_{i,l}(D_2) - V_{i,k}(D_1) - V_{i,l}(D_1)) = V_{a,k}(D_2) + \sum_{i\neq a} V_{i,k}(D_2) + V_{b,l}(D_2)$$
$$+ \sum_{i\neq b} V_{i,l}(D_2) - V_{a,k}(D_1) - \sum_{i\neq a} V_{i,k}(D_1) - V_{a,l}(D_1) - V_{b,l}(D_1)$$
$$- \sum_{i\neq a,b} V_{i,l}(D_1) = \frac{(m_a\frac{m_a}{n} - m_a)^2}{m_a\frac{m_a}{n}} + \sum_{i\neq a}\frac{(m_a\frac{m_i}{n})^2}{m_a\frac{m_i}{n}} + \frac{(m_b\frac{m_b}{n} - m_b)^2}{m_b\frac{m_b}{n}} + \sum_{i\neq b}\frac{(m_b\frac{m_i}{n})^2}{m_b\frac{m_i}{n}}$$
$$- \frac{((m_a-1)\frac{m_a}{n} - (m_a-1))^2}{(m_a-1)\frac{m_a}{n}} - \sum_{i\neq a}\frac{((m_a-1)\frac{m_i}{n})^2}{(m_a-1)\frac{m_i}{n}} - \frac{((m_b+1)\frac{m_a}{n} - 1)^2}{(m_b+1)\frac{m_a}{n}}$$
$$- \frac{((m_b+1)\frac{m_b}{n} - m_b)^2}{(m_b+1)\frac{m_b}{n}} - \sum_{i\neq a,b}\frac{((m_b+1)\frac{m_i}{n})^2}{(m_b+1)\frac{m_i}{n}} = \frac{(m_a+m_b)n}{m_a(1+m_b)} \tag{10}$$

Therefore, given a, when the value of J is greater than or equal to 3, global sensitivity is represented by Eq. (10). Moreover, from Proposition 1, global sensitivity is represented by Eq. (6) when the value of J is greater than or equal to 3 and a is not given.

When $J = 2$ and a is given, neighboring databases that satisfy the following constraints will maximize the difference between the χ^2 values of contingency tables D_1 and D_2 from Proposition 3 (see Fig. 1b).

$$\begin{aligned} O_{a,k}(D_1) + 1 &= O_{a,k}(D_2) = m_a \\ O_{a,l}(D_1) - 1 &= O_{a,l}(D_2) = 0 \\ O_{i,k}(D_1) &= O_{i,k}(D_2) = 0 \text{ for all } i \text{ except for } i = a \\ O_{i,l}(D_1) &= O_{i,l}(D_2) = m_i \text{ for all } i \text{ except for } i = a \end{aligned} \tag{11}$$

The difference between the χ^2 values of tables D_1 and D_2 that satisfy the constraint (11) can be calculated as

$$\sum_i (V_{i,k}(D_2) + V_{i,l}(D_2) - V_{i,k}(D_1) - V_{i,l}(D_1)) = V_{a,k}(D_2) + \sum_{i\neq a} V_{i,k}(D_2) + V_{a,l}(D_2)$$
$$+ \sum_{i\neq a} V_{i,l}(D_2) - V_{a,k}(D_1) - \sum_{i\neq a} V_{i,k}(D_1) - V_{a,l}(D_1) - \sum_{i\neq a} V_{i,l}(D_1) = \frac{(m_a\frac{m_a}{n} - m_a)^2}{m_a\frac{m_a}{n}}$$
$$+ \sum_{i\neq a}\frac{(m_a\frac{m_i}{n})^2}{m_a\frac{m_i}{n}} + \frac{((n-m_a)\frac{m_a}{n})^2}{(n-m_a)\frac{m_a}{n}} + \sum_{i\neq a}\frac{((n-m_a)\frac{m_i}{n} - m_i)^2}{(n-m_a)\frac{m_i}{n}}$$
$$- \frac{((m_a-1)\frac{m_a}{n} - (m_a-1))^2}{(m_a-1)\frac{m_a}{n}} - \sum_{i\neq a}\frac{((m_a-1)\frac{m_i}{n})^2}{(m_a-1)\frac{m_i}{n}} - \frac{((n-m_a+1)\frac{m_a}{n} - 1)^2}{(n-m_a+1)\frac{m_a}{n}}$$
$$- \sum_{i\neq a}\frac{((n-m_a+1)\frac{m_i}{n} - m_i)^2}{(n-m_a+1)\frac{m_i}{n}} = \frac{n^2}{m_a(n-m_a+1)}. \tag{12}$$

Because $n^2/(m_a(n-m_a+1))$ decreases when m_a decreases, the global sensitivity can be represented by Eq. (6) when J is equal to 2 and a is not given. □

When we use a 2×3 contingency table, Δ_R is identical to Δ_Y, and when we use a 2×3 contingency table with $m_1 = m_2$, Δ_R is identical Δ_F.

Propositions 1 and 2 used in the proof of Theorem 2 are described below.

Proposition 1 *Δ_R in Eq. (6) is maximized when the minima (7) are satisfied.*

Proof By differentiating Eq. (6) with respect to m_a, we obtain

$$-\frac{m_b n}{m_a^2(1+m_b)}. \tag{13}$$

By differentiating Eq. (6) with respect to m_b, we obtain

$$-\frac{(m_a-1)n}{m_a(1+m_b)^2}. \tag{14}$$

From Expressions (13) and (14), m_a and m_b should thus be minimized to maximize Eq. (6).

Let *min* denote the minimum value in m_i $(i = 1, \ldots, I)$ and let *min*+x denote the second most minimum value in m_i $(i = 1, \ldots, I)$, where $x \geq 0$. If m_a is *min* and m_b is $min + x$, Eq. (6) can then be expressed as

$$\frac{n(2min+x)}{min(1+min+x)}. \tag{15}$$

If m_a is $min + x$ and m_b is *min*, Eq. (6) can then be expressed as

$$\frac{n(2min+x)}{(1+min)(min+x)}. \tag{16}$$

Because Expression (15) is always greater than or equal to Expression (16), we find that the value of Δ_R in Eq. (6) is maximized when (7) is satisfied. □

Proposition 2 *When J is greater than or equal to 3 and a is given, neighboring databases that satisfy the constraints (9) maximize the difference between the χ^2 values of tables D_1 and D_2.*

Proof There are many neighboring databases that satisfy Eq. (8); however, we prove that neighboring databases that satisfy the constraints (9) have the greatest difference, $\delta(D_1, D_2)$, between $\chi^2(D_1)$ and $\chi^2(D_2)$ when $J \geq 3$. We assume that m_i is a fixed value for all values of i.

Thus, we write $O_{i,j}(D_1)$ as $O_{i,j}$ for any i and j in the following manner.

Following Lemma 1, $O_{a,k}$ should be maximized to maximize $\delta(D_1, D_2)$. As a result, the value of $O_{a,k}$ becomes $m_a - 1$ because of the constraints (17).

From Eq. (8) we have the following constraints:

$$O_{a,k}(D_1) \leq m_a - 1 \quad \text{and} \quad 1 \leq O_{a,l}. \tag{17}$$

Following Lemma 2, $O_{i,k}$ should be zero to maximize $\delta(D_1, D_2)$ for all values of i except $i = a$.

Following Lemma 3, $O_{a,l}$ should be minimized to maximize $\delta(D_1, D_2)$. As a result, the value of $O_{a,l}$ becomes 1 because of the constraints (17).

Following Lemma 4, $O_{\mu,l}$ $\mu \neq a$ should be m_μ and $O_{i,l}$ for all i, except for $i = a$, and $i = \mu$ should be zero to maximize $\delta(D_1, D_2)$.

As a result, we can maximize $\delta(D_1, D_2)$ when tables D_1 and D_2 satisfy the constraints (9) by replacing μ in Lemma 4 with b. □

Lemma 1 *To maximize $\delta(D_1, D_2)$, $O_{a,k}$ should be maximized (and correspondingly, $O_{a,r}$ for all r, except for $r = k, l$, should be adjusted to satisfy m_a).*

Proof We have

$$\begin{aligned}
&\delta(D_1, D_2) = \chi^2(D_2) - \chi^2(D_1) \\
&= V_{a,k}(D_2) - V_{a,k}(D_1) + \sum_{i \neq a}(V_{i,k}(D_2) - V_{i,k}(D_1)) \\
&+ V_{a,l}(D_2) - V_{a,l}(D_1) + \sum_{i \neq a}(V_{i,l}(D_2) - V_{i,l}(D_1)) \\
&= -2 + \frac{m_a}{n} + \frac{n(-O_{a,k}^2 + s_k + 2s_k O_{a,k})}{m_a s_k(1+s_k)} \\
&+ \sum_{i \neq a} \frac{m_i^2 s_k(1+s_k) - n^2 O_{i,k}^2}{m_i n s_k(1+s_k)} \\
&+ 2 - \frac{m_a}{n} + \frac{n(O_{a,l}^2 + s_l - 2s_l O_{a,l})}{m_a(s_l - 1)s_l} \\
&+ \sum_{i \neq a} \frac{n^2 O_{i,l}^2 - m_i^2(s_l - 1)s_l}{m_i n(s_l - 1)s_l}.
\end{aligned} \tag{18}$$

By differentiating Eq. (18) with respect to $O_{a,k}$, we obtain

$$\frac{n(O_{a,k} - s_k)^2(1+2s_k)}{m_a s_k^2(1+s_k)^2} + \sum_{i \neq a} \frac{nO_{i,k}^2(1+2s_k)}{m_i s_k^2(1+s_k)^2}, \tag{19}$$

because we have

$$\frac{\partial s_k}{\partial O_{a,k}} = 1. \tag{20}$$

Because Expression (19) is always ≥ 0, Eq. (18) increases as $O_{a,k}$ increases.

Thus, we can say that $O_{a,k}$ should be increased to maximize $\delta(D_1, D_2)$. As a result, we have $O_{a,k} = m_a - 1$. □

Lemma 2 *To maximize $\delta(D_1, D_2)$, $O_{i,k}$ should be minimized (and correspondingly $O_{i,r}$ for all r, except for $r = k, l$, should be adjusted to satisfy m_i) for all values of i except for $i = a$.*

Proof We focus on $\mu \in \{1, \ldots, I\}$ such that $\mu \neq a$. By differentiating Eq. (18) with respect to $O_{\mu,k}$, we obtain

$$\begin{aligned}
&\frac{n(O_{a,k} - s_k)(O_{a,k} + s_k + 2s_k O_{a,k})}{m_a s_k^2(1+s_k)^2} \\
&+ \frac{nO_{\mu,k}(O_{\mu,k} + 2s_k O_{\mu,k} - 2s_k(1+s_k)}{m_\mu s_k^2(1+s_k)^2} \\
&+ \sum_{i \neq a,\mu} \frac{nO_{i,k}^2(1+2s_k)}{m_i s_k^2(1+s_k)^2},
\end{aligned} \tag{21}$$

because we have

$$\frac{\partial s_k}{\partial O_{\mu,k}} = 1. \tag{22}$$

Let $\Theta = \sum_{i \neq a,\mu} O_{i,k}^2 / m_i$. By solving equation (21)$= 0$ for Θ, we obtain

$$\begin{aligned} &(m_\mu (s_k - O_{a,k})(O_{a,k} + s_k + 2 s_k O_{a,k}) \\ &+ m_a O_{u,k}(2 s_k - O_{u,k} + 2 s_k (s_k - O_{u,k})) \\ &/ (m_a m_\mu (1 + 2 s_k)). \end{aligned} \tag{23}$$

Expression (23) is always greater than zero. When $\Theta = 0$ in Expression (21), the value of Expression (21) is less than 0.

Thus, when Θ is less than Expression (23), Expression (21) is less than zero. Similarly, when Θ is greater than Expression (23), Expression (21) is greater than zero. That is, to maximize Eq. (18), the value of $O_{\mu,k}$ should be either minimized or maximized. From this observation, to maximize Eq. (18), we can say that $O_{i,k}$ should be either minimized (i.e., zero) or maximized (i.e., m_i) for all i except for $i = a$,.

From Lemma 1, we have $O_{a,k} = m_a - 1$. Therefore, when $O_{i,k} = 0$ for all i except $i = a$, we have $s_k = m_a - 1$. In this case, $\delta(D_1, D_2)$ is

$$\begin{aligned} &-2 + \frac{m_a}{n} + \frac{n}{m_a} + \frac{1}{n} \sum_{i \neq a} m_i \\ &+ V_{a,l}(D_2) - V_{a,l}(D_1) + \sum_{i \neq a} (V_{i,l}(D_2) - V_{i,l}(D_1)). \end{aligned} \tag{24}$$

In contrast, when $O_{i,k} = m_i$ for all i except $i = a$, $s_k = \sum_i m_i - 1 = n - 1$. In this case, $\delta(D_1, D_2)$ is

$$-2 + \frac{m_a}{n} + \frac{n}{m_a} + \sum_{i \neq a} \frac{m_i}{n} - \frac{(n - m_a)^2}{m_a (n-1) n} + \sum_{i \neq a} \frac{m_i}{n - n^2}. \tag{25}$$

By subtracting Expression (25) from Expression (24), we obtain

$$\frac{(n - m_a)^2}{m_a (n-1)} + \sum_{i \neq a} \frac{m_i}{n-1}. \tag{26}$$

Because Expression (26) is always greater than zero, $O_{i,k}$ for all i except $i = a$ should be zero. □

Lemma 3 *To maximize $\delta(D_1, D_2)$, $O_{a,l}$ should be minimized (and correspondingly $O_{a,r}$ for all r except $r = k, l$ should be adjusted to satisfy m_a).*

Proof By differentiating Eq. (18) with respect to $O_{a,l}$, we obtain

$$\frac{n(O_{a,l} - s_l)^2 (1 - 2 s_l)}{m_a (s_l - 1)^2 s_l^2} + \sum_{i \neq a} \frac{n O_{i,l}^2 (1 - 2 s_l)}{m_i (s_l - 1)^2 s_l^2}, \tag{27}$$

because we have

$$\frac{\partial s_l}{\partial O_{a,l}} = 1. \tag{28}$$

Because (27) is always less than zero, Eq. (18) increases as $O_{a,l}$ decreases. □

Lemma 4 *To maximize $\delta(D_1, D_2)$, $O_{\mu,l}$ ($\mu \neq a$) should be maximized (and correspondingly, $O_{\mu,r}$ for all r except $r = k, l$ should be adjusted to satisfy m_μ). Additionally, $O_{i,l}$ should be minimized (and correspondingly, $O_{i,r}$ for all r except $r = k, l$ should be adjusted to satisfy m_i) for all i except $i = a$ and $i = \mu$.*

Proof By differentiating Eq. (18) with respect to $O_{\mu,l}$ ($\mu \neq a$), we obtain

$$\begin{aligned}&\frac{n(s_l - O_{a,l})(2s_l O_{a,l} - O_{a,l} - s_l)}{m_a(s_l-1)^2 s_l^2}\\&+\frac{nO_{\mu,l}(O_{\mu,l} - 2s_l - 2s_l O_{\mu,l} + 2s_l^2)}{m_\mu(s_l-1)^2 s_l^2}\\&+\sum_{i\neq a,\mu}\frac{nO_{i,l}^2(1-2s_l)}{m_i(s_l-1)^2 s_l^2},\end{aligned} \tag{29}$$

because we have

$$\frac{\partial s_l}{\partial O_{\mu,l}} = 1. \tag{30}$$

Let $\Theta = \sum_{i\neq a,\mu} O_{i,l}^2/m_i$. We have $O_{a,l} = 1$ from Lemma 3. By solving Expression (29)=0 for Θ, we obtain

$$\frac{m_\mu(s_l-1)^2 + m_a O_{u,l}(2s_l(s_l - O_{\mu,l} - 1) + O_{\mu,l})}{m_a m_\mu(2s_l - 1)}. \tag{31}$$

Expression (31) is always greater than zero. When $\Theta = 0$ and $O_{a,l} = 1$ in Expression (29), (29) can be expressed as

$$\frac{n(m_\mu(s_l-1)^2 + m_a O_{\mu,l}(2s_l(s_l - O_{\mu,l} - 1) + O_{\mu,l}))}{m_a m_\mu(s_l-1)^2 s_l^2} \geq 0. \tag{32}$$

Therefore, when Θ is less than or equal to Expression (31), Expression (29) is greater than zero, and when Θ is greater than Expression (31), Expression (29) is less than zero. Thus, to maximize Eq. (18), the value of $O_{\mu,l}$ should be either minimized (i.e., zero) or maximized (i.e., m_μ).

Thus, to maximize $\delta(D_1, D_2)$, the value of $O_{\mu,l}$ should be either minimized or maximized. Let us have $x = \sum_{i\neq\mu} O_{i,l}$. When $O_{\mu,l}$ is maximized (i.e., $O_{\mu,l} = m_\mu$), we $\delta(D_1, D_2)$ is

$$\begin{aligned}&2 - \frac{m_a + m_\mu}{n} + \frac{m_\mu n}{m_\mu + x - 1} - \frac{n + m_a m_\mu n}{m_a(m_\mu + x)}\\&+\sum_{i\neq a,\mu}\left(\frac{O_{i,l}^2 n}{m_i(m_\mu + x - 1)(m_\mu + x)} - \frac{m_i}{n}\right).\end{aligned} \tag{33}$$

In contrast, when $O_{\mu,l}$ is minimized (i.e., $O_{\mu,l} = 0$), $\delta(D_1, D_2)$ is

$$\begin{aligned}&2 - \frac{m_a + m_\mu}{n} - \frac{n}{m_a x}\\&+\sum_{i\neq a,\mu}\frac{O_{i,l}^2 n^2 - m_i^2(x-1)x}{m_i n(x-1)x}.\end{aligned} \tag{34}$$

By subtracting Expression (34) from Expression (33), we obtain

$$\begin{aligned}&\frac{m_\mu n(-1 + m_\mu + x + m_a x)}{m_a x(-1 + m_\mu + x)(m_\mu + x)}\\&-\sum_{i\neq a,\mu}\frac{n m_\mu O_{i,l}^2(-1 + m_\mu + 2x)}{m_i(x-1)x(m_\mu + x - 1)(m_\mu + x)}.\end{aligned} \tag{35}$$

When $I = 2$, the second term of Expression (35) is zero. Therefore, Expression (35) is always greater than zero and Lemma 4 holds when $I = 2$.

We then consider the situation where $I \geq 3$. We assume that $O_{i,l}$ is zero for all values of i except $i = a$ and $i = \mu$. In this case the second term of Expression (35) is zero and the first term of Expression (35) is greater than zero; therefore, we can say that Expression (35) is always greater than zero. Thus, $O_{\mu,l}$ should be maximized to m_μ when $O_{i,l}$ is zero for all values of i except $i = a$ and $i = \mu$.

Next, we focus on ν such that $\nu \in \{1, \ldots, I\}$ and $\nu \neq a, \mu$. We demonstrate that Expression (35) is always ≤ 0 when $O_{\nu,l}$ is maximized to m_ν. Additionally, the second term of Expression (35) is minimized when $I = 3$. In this case, we obtain

$$(35) \leq -\frac{(m_a - 1)m_\mu n}{m_a(1 + m_\nu)(1 + m_\nu + m_\mu)} < 0, \tag{36}$$

because $x = m_\nu + 1$.

Therefore, each $O_{i,l}$ for all i except $i \neq a, \mu$ should be minimized to zero.

From this observation, Lemma 4 also holds when $I \geq 3$. □

Proposition 3 *When J equals 2 and a is given, neighboring databases that satisfy the constraints (11) maximize the difference between the* χ^2 *values of tables* D_1 *and* D_2.

The proof can be conducted in a similar manner as Lemma 2.

Differentially private hypothesis testing

We can now calculate the anonymized χ^2 value from an original table,

$$\chi^{2*} = \chi^2 + Lap(\Delta_R/\epsilon), \tag{37}$$

where χ^{2*} is the anonymized χ^2 value.

From the definitions of the Laplace distribution and χ^2 distribution, the probability density function of a χ^2 value possessing ν degrees of freedom with the addition of Laplace noise and global sensitivity Δ can be expressed as

$$g_{\nu,\Delta,\epsilon}(x) = \int_{\mu=-\infty}^{\infty} \mathcal{L}_{\mu,\beta}(x)\mathcal{Z}_\nu(\mu)d\mu, \tag{38}$$

where

$$\beta = \Delta/\epsilon, \tag{39}$$

$$\mathcal{L}_{\mu,\beta}(x) = \begin{cases} \frac{\exp\left(-\frac{x-\mu}{\beta}\right)}{2\beta} & x \geq \mu \\ \frac{\exp\left(-\frac{\mu-x}{\beta}\right)}{2\beta} & otherwise, \end{cases} \tag{40}$$

and

$$\mathcal{Z}_\nu(u) = \begin{cases} \frac{2^{-\nu/2}\exp(-u/2)u^{-1+\nu/2}}{\Gamma(\nu/2)} & x > 0 \\ 0 & otherwise, \end{cases} \tag{41}$$

where $\Gamma(\nu/2)$ represents the $\nu/2$ gamma function, that is,

$$\Gamma(\nu/2) = \int_0^\infty x^{\nu/2-1}e^{-x}dx. \tag{42}$$

When we set the significance level to α, our proposed RandChiDist rejects H_0 if the χ^2 value calculated using Eq. (1), with the addition of Laplace noise and the scale Δ_R/ϵ,

is greater than or equal to α, as calculated by solving the following equation with regard to α;

$$\int_{x=t}^{\infty} g_{\nu,\Delta,\epsilon}(x) = \alpha. \tag{43}$$

Lastly, we compare the ${\chi^2}^*$ value calculated using Eq. (37) to the t value calculated using Eq. (43). When ${\chi^2}^*$ is greater than or equal to t, RandChiDist outputs "reject the null hypothesis H_0," and otherwise outputs "fail to reject the null hypothesis H_0."

Algorithm 1 shows the overall RandChiDist algorithm.

Algorithm 1 Algorithm of RandChiDist

Input: Significance level α, Privacy parameter ϵ, Original cross table T

Output: Result of the null hypothesis test

1: Calculate original χ^2 value from (1)
2: Calculate global sensitivity ΔR from (6)
3: ${\chi^2}^* \Leftarrow \chi^2 + Lap(\Delta R/\epsilon)$
4: Calculate value of t from (43)
5: **if** ${\chi^2}^* \geq t$ **then**
6: Return "reject the null hypothesis H_0"
7: **else**
8: Return "fail to reject the null hypothesis H_0"
9: **end if**

If we want an anonymized version of the p value, RandChiDist calculates and outputs

$$\int_{x={\chi^2}^*}^{\infty} g_{\nu,\Delta,\epsilon}(x). \tag{44}$$

The data analysis can thus conduct a χ^2 hypothesis test using an arbitrary α by comparing Expression (44) and α.

Complexity analysis

Calculating original χ^2 yields a computational complexity of $O(I \times J)$. Calculating global sensitivity ΔR requires finding the largest value and the second largest value of m_i $(i = 1, \ldots, I)$; therefore, the computational complexity is $O(I)$. Calculating (43) and (44) requires the calculation of an integration. For example, Monte Carlo integration can be adopted to calculate an integration. The computational complexity of Monte Carlo integration is not influenced by the cross table. There are numerous Monte Carlo integration methods that can be calculated extremely fast [55].

Therefore, the computational complexity of the proposed algorithm is $O(I \times J + M)$, where M denotes the computational complexity of calculating an integration.

Evaluation

We compared RandChiDist, RandCell, MCIndep, and LocalExpIND as described in "Related work" section.

LocalExpIND was proposed especially for local privacy; therefore, LocalExpIND can be used for more scenarios than RandChiDist. Thus, the local model of privacy is another avenue for future exploration.

Moreover, to clarify the contribution of calculating the private χ^2 distribution table's value (proposed in "Differentially private hypothesis testing" section), we also compared a method that uses the global sensitivity Δ_R calculated using Eq. (6) that does not use the private χ^2 distribution table's value calculated using Eq. (24). We refer to this method as RandChi, which is also proposed in this paper.

The source code for the RandChi and RandChiDist methods can be obtained from https://uecdisk.cc.uec.ac.jp/index.php/s/pic3T9GEp03qy6y.

We should use Bonferroni's corrected threshold when conducting multiple χ^2 testing [56]. In this paper, we conducted many χ^2 tests; however, we consider each to be independent. Thus, Bonferroni's corrected threshold was not used in this paper to compare the performance among our proposed methods and methods from existing studies for independent χ^2 testing. This paper shows the average results of each independent χ^2 test. Additionally, previous studies of privacy-preserving χ^2 testing, such as [7, 8, 27–29, 44, 45], did not use the Bonferroni's corrected threshold.

We varied the values of n from 100 to 900, α from 0.005 to 0.05, and ϵ from 0.01 to 10. We set the parameters of MCIndep the same as in [8].

Significance results

We first evaluated the significance to confirm that RandChiDist guarantees a significance of at least $1-\alpha$. We randomly generated 2×2 contingency tables based on a multinomial distribution with probabilities of (0.25, 0.25, 0.25, 0.25) 1,000 times. Each time, we evaluated whether each method correctly output "fail to reject the null hypothesis H_0." Figure 2 shows the results with an ϵ value of 0.1. The significance of each method should be approximately $1-\alpha$.

The significance levels of RandChiDist, MCIndep, and LocalExpIND were controlled around $1-\alpha$ for any n, ϵ, and α values. In contrast, RandCell and RandChi had significance values much less than $1-\alpha$ when ϵ was less than 1.

We conducted the same experiments for randomly generated 4×4 contingency tables based on a multinomial distribution with probabilities of $1/16,\ldots,1/16$. Figure 3 shows the results with $\epsilon = 0.1$. As with the 2×2 contingency tables, the significance values of RandChiDist, MCIndep, and LocalExpIND were approximately $1-\alpha$. In contrast, RandCell and RandChi significance values were less than $1-\alpha$, especially when ϵ was small.

RandCell adds a Laplace noise to each cell. The probability that at least one Laplace noise becomes very large increases when the number of cells is large. Therefore, RandCell has many false positives (i.e., significance results are small) when contingency tables are

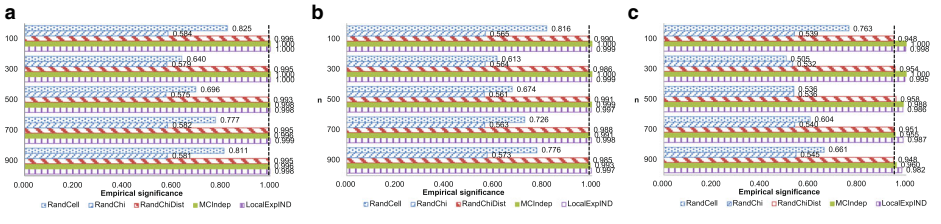

Fig. 2 Significance results based on 2×2 contingency tables. The dashed lines represent $1-\alpha$. **a** Results with $\epsilon = 0.1, \alpha = 0.005$. **b** Results with $\epsilon = 0.1, \alpha = 0.01$. **c** Results with $\epsilon = 0.1, \alpha = 0.05$

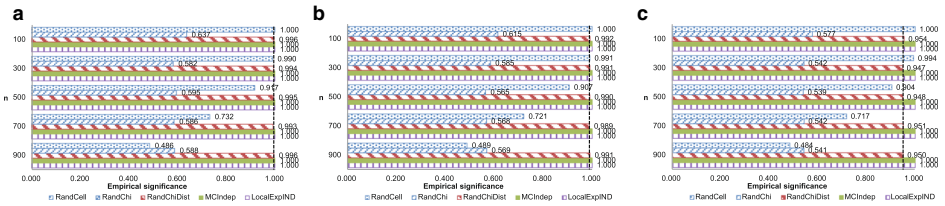

Fig. 3 Significance results based on 4 × 4 contingency tables. The dashed lines represent $1-\alpha$. **a** Results with $\epsilon = 0.1, \alpha = 0.005$. **b** Results with $\epsilon = 0.1, \alpha = 0.01$. **c** Results with $\epsilon = 0.1, \alpha = 0.05$

large. On the other hand, if the Laplace noise is a large negative value, the cell value with the noise could be less than five (or negative). In this case, RandCell fails to reject the null hypothesis based on the rule of thumb. Therefore, RandCell's results of 4 ×4 tables are smaller than those of 2 ×2 tables only when n is large.

RandChi's significance results do not vary greatly by the table size or n. This is because the global sensitivity calculated from Eq. 7 also does not vary greatly by the table size or n.

Power results

We then evaluated each method's power. The values of parameters α, ϵ, and n were identical to those in the significance experiments; however, we randomly generated 2×2 contingency tables based on a multinomial distribution with probabilities of $(0.25+0.01, 0.25-0.01, 0.25-0.01, 0.25+0.01)$ and $(0.25+0.15, 0.25-0.15, 0.25-0.15, 0.25+0.15)$. We also used another probability set $(0.3+0.15, 0.3-0.15, 0.2-0.15, 0.2+0.15)$ to determine whether RandChiDist can be applied to unbalanced tables. Moreover, we randomly generated 3×4 contingency tables based on a multinomial distribution with probabilities of $(1/12+0.07, 1/12-0.07, 1/12, 1/12, 1/12-0.07, 1/12+0.07, 1/12, \ldots, 1/12)$. Each time, we evaluated whether each method correctly output "reject the null hypothesis H_0." Figures 4, 5, 6, and 7 show the results for $\epsilon = 0.1$.

In the experiment on a multinomial distribution with probabilities of $(0.25+0.01, 0.25-0.01, 0.25-0.01, 0.25+0.01)$, the empirical power of Non-private, which does not consider privacy at all, is very low, which is approximately from 0 to 0.2. Hence, all privacy-preserving algorithms that can control Type I errors do not realize high empirical power, although RandChiDist, which we proposed, is just slightly better than the other algorithms.

In the experiments on other multinomial distributions, MCIndep has relatively low empirical power. MCIndep generated many contingency tables from its algorithm based

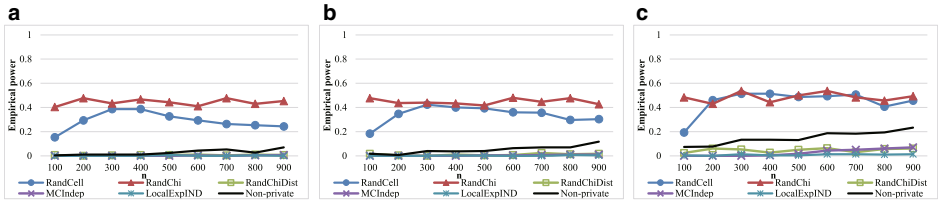

Fig. 4 Empirical power results with 2 × 2 Contingency tables generated with probabilities of $(0.25+0.01, 0.25-0.01, 0.25-0.01, 0.25+0.01)$. **a** Results with $\epsilon = 0.1, \alpha = 0.005$. **b** Results with $\epsilon = 0.1$, $\alpha = 0.01$. **c** Results with $\epsilon = 0.1, \alpha = 0.05$

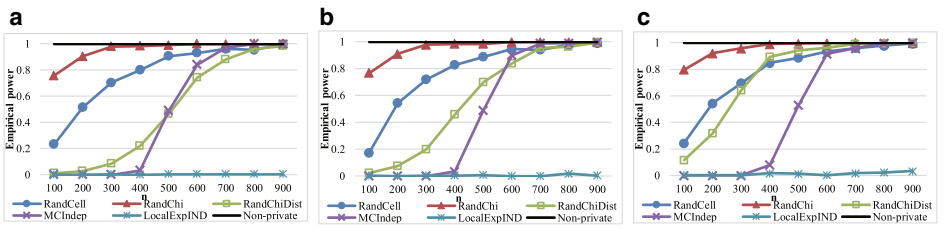

Fig. 5 Empirical power results with 2 × 2 Contingency tables generated with probabilities of $(0.25 + 0.15, 0.25 - 0.15, 0.25 - 0.15, 0.25 + 0.15)$. **a** Results with $\epsilon = 0.1, \alpha = 0.005$. **b** Results with $\epsilon = 0.1$, $\alpha = 0.01$. **c** Results with $\epsilon = 0.1, \alpha = 0.05$

on the original contingency table. MCIndep quickly outputs "fail to reject H_0" when at least one cell in the generated contingency tables has a value of less than five. Therefore, even if all the target contingency table's cells have values greater than five, MCIndep is likely to output "fail to reject H_0" if several values are close to 5 (for example a value of 10).

In contrast, RandCell and RandChi both achieved high empirical power at the expense of empirical significance. The empirical power of MCIndep is high when there are many samples and the data are uniformly distributed. RandChiDist achieved higher empirical power with fewer samples than MCIndep while also achieving empirical significance.

In hypothesis testing that includes χ^2 testing, we should avoid Type I errors (i.e., false positives). In general, we adjust the Type I error probability by the value of α (e.g., 0.05). Even if the empirical power is high, the algorithm is of no use if the empirical significance is less than 1-α. The empirical power of RandCell and RandChi is greater than that of RandChiDist; however, RandCell and RandChi have empirical significance values much than 1-α. That is, they cannot control Type I errors (false positives) in many cases. Therefore, we can conclude that RandChiDist outperforms RandCell and RandChi. Among RandChiDist, MCIndep, and LocalExpIND, which can control Type I errors, RandChiDist has the highest power.

Results of real datasets

We used two genomic datasets[1]. The first dataset is the Human Genome Diversity Project genotype dataset (HGDP) used by Conrad [57], which consists of 2,834 SNPs and has 1,244 records after the records in which unknown values are eliminated. The other is the International Haplotype Map Project genotype dataset (HapMap) used by [58], which consists of 1,853 SNPs and has 420 complete records.

We randomly generated contingency tables for linkage disequilibrium analysis for each dataset and set the numbers of columns and rows to four. Following the "rule of thumb," if any values of the created contingency table are less than five, we re-created another contingency table and then conducted normal χ^2 testing on the original contingency tables. We then carried out the privacy-preserving methods. We generated contingency tables and conducted χ^2 testing 100 times, and then calculated the mean results of false positive and false negative rates.

The results for the HGDP genotype and HapMap genotype datasets are shown in Figs. 8 and 9, respectively. RandChiDist outperformed MCIndep and LocalExpIND for most of the parameter settings used in this paper.

[1] https://web.stanford.edu/group/rosenberglab/hgdpsnpDownload.html (accessed May 26, 2017)

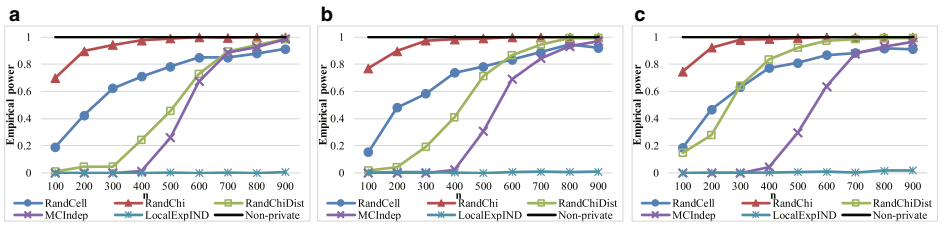

Fig. 6 Empirical power results with 2 × 2 Contingency tables generated with probabilities of $(0.3 + 0.15, 0.3 - 0.15, 0.2 - 0.15, 0.2 + 0.15)$. **a** Results with $\epsilon = 0.1, \alpha = 0.005$. **b** Results with $\epsilon = 0.1$, $\alpha = 0.01$. **c** Results with $\epsilon = 0.1, \alpha = 0.05$

Discussion

According to the evaluation results, RandCell and RandChi could not control the ratio of Type I errors—that is, they caused a lot of false positives. On the contrary, RandChiDist, MCIndep, and LocalExpIND could control the ratio of Type I errors. RandChiDist achieved the least number of Type II errors among RandChiDist, MCIndep, and LocalExpIND. When testing a hypothesis, data analyzers determine the significance level α (i.e., the ratio of Type I errors) ahead of time. That is, they reject a true null hypothesis with a probability no greater than α. A high false positive rate means that a true null hypothesis is rejected with a probability greater than α, which leads to the false interpretation of datasets. Therefore, if we want to avoid such false interpretations, RandChiDist is the preferred method.

There are several approaches for non-private χ^2 testing. The simplest approach is shown in "χ^2 hypothesis test of Independence" section. RandChi and RandChiDist calculate the global sensitivity of the χ^2 value of the simplest chi-squared testing and adds noise based on the global sensitivity to the χ^2 value. Thus, the added noise is minimized following the Laplace mechanism theorem (Theorem 1).

In contrast, RandCell calculates the global sensitivity of each value of each cell and adds noise to each value. The summation of added noises thus become very large. MCIndep takes another approach for calculating non-private χ^2 testing, as shown in "Related work" section. MCIndep first estimates the parameters of the underlying multinomial distribution generating the samples. By the estimated the multinomial distribution, MCIndep generates more than $1/\alpha$ contingency tables. When the number of samples is small, the estimated parameters of the underlying multinomial distribution have low accuracy. Because of this low accuracy estimation, MCIndep could have low performance when the number of samples is small. LocalExpIND assumes that each piece data is anonymized for

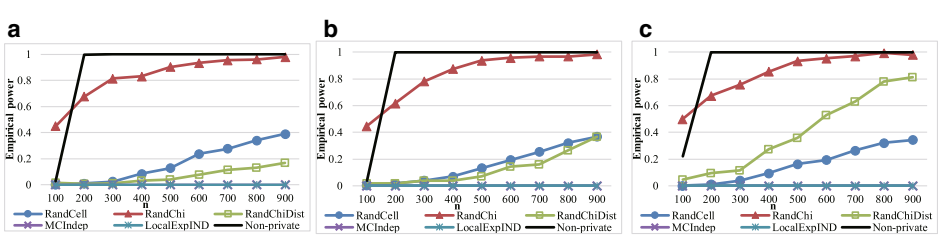

Fig. 7 Empirical power results with 3 × 4 Contingency tables generated with probabilities of $(1/12 + 0.07, 1/12 - 0.07, 1/12, 1/12, 1/12 - 0.07, 1/12 + 0.07, 1/12, \ldots, 1/12)$. **a** Results with $\epsilon = 0.1$, $\alpha = 0.005$. **b** Results with $\epsilon = 0.1, \alpha = 0.01$. **c** $\epsilon = 0.1, \alpha = 0.05$

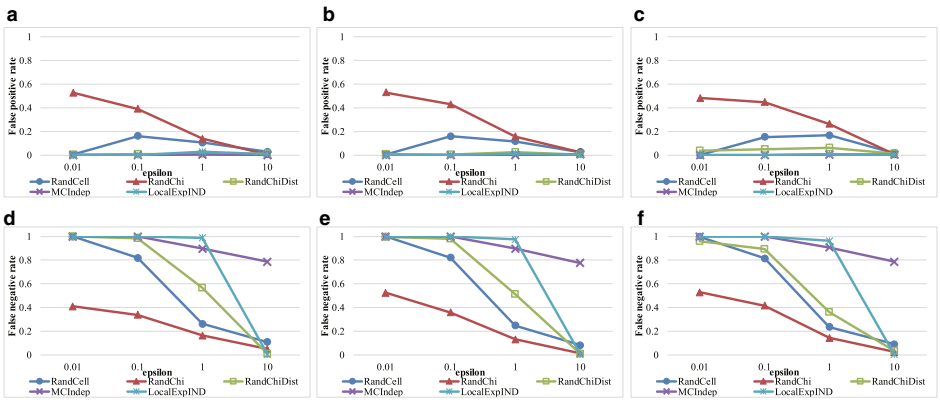

Fig. 8 Results of the HGDP genotype datasets. **a** Results with False Positive Rate ($\alpha = 0.005$). **b** Results with False Positive Rate ($\alpha = 0.01$). **c** False Positive Rate ($\alpha = 0.05$). **d** False Negative Rate ($\alpha = 0.005$). **e** False Negative Rate ($\alpha = 0.01$). **f** False Negative Rate ($\alpha = 0.05$)

each person and that there is no trusted entity. Because noise is added to each data point, the amount of the total noise becomes large.

Sharpe claimed that if we can avoid χ^2 hypothesis testing for contingency tables larger than 2×2, doing so is desirable [59]. However, he showed an understanding that in some cases we could not avoid this and also reported that approximately 30% of χ^2 tests are conducted for contingency tables larger than 2×2. This is based on his survey of journals published by the American Psychological Association for 2012, 2013, and early 2014. χ^2 hypothesis testing has been widely used for GWAS as well as many other personal databases [38–40]. Moreover, some studies [38–40] have considered contingency tables larger than 2×3. Therefore, we consider the application of ϵ-differential privacy to χ^2 hypothesis testing for contingency tables larger than 2×3 to be an important issue.

Our proposed method can be used not only GWAS but also other private data analysis for small samples. For example, the characteristics of COVID-19 patients (n=403) (the number of died patients was 100 and the number of recovered patients was 303) were analyzed by χ^2 test with α being 0.05 [60]. Poyiadi et al. analyzed the COVID-19 with acute

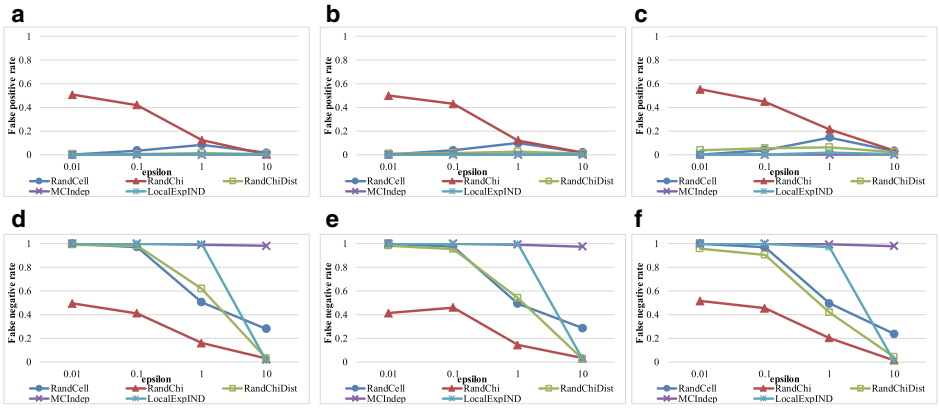

Fig. 9 Results of the HapMap genotype datasets. **a** Results with False Positive Rate ($\alpha = 0.005$). **b** False Positive Rate ($\alpha = 0.01$). **c** False Positive Rate ($\alpha = 0.05$). **d** False Negative Rate ($\alpha = 0.005$). **e** False Negative Rate ($\alpha = 0.01$). **f** False Negative Rate ($\alpha = 0.05$)

pulmonary embolism and the COVID-19 without acute pulmonary embolism [61]. The number of patients was n=328. They conducted χ^2 test with α being 0.05. The influence on sexual activity for COVID-19 was analyzed by Jacob et al. [62]. The number of samples was 868. As these studies show, there is a high need for testing with a small sample size. In particular, it is difficult to collect a large number of samples when there is a need for rapid analysis for a new disease such as COVID-19.

We assume that the data holder publishes m_i as well as the differentially private chi-square value. In general, the information of m_i and a sample size is necessary to interpret a chi-square value accurately [63]. For example, even if in the case of trivial differences between two datasets, a very small chi-square value is obtained when every m_i is very large [64]. Therefore, m_i is very useful information for data analysts.

Publishing m_i also provides several other types of information. For example, we know that $O_{i,j}$ for all j are less than or equal to m_i. However, we cannot know each value of $O_{i,j}$, and we cannot know which value is greater ($O_{i,j}$ or $O_{i,j'}$) for any j or j', even if we know m_i and the (differentially private) chi-square value. Our proposed algorithm can protect chi-square values based on differential privacy, and we can ensure that it is impossible to reconstruct the original cross table. To the best of our knowledge, no researchers have claimed that publishing m_i could cause privacy issues.

Conclusion

χ^2 testing is widely used in GWAS and other types of data analysis. We proposed the RandChiDist method, which anonymizes the χ^2 value of contingency tables. If we have a lot of samples for data analysis, it is easy to conduct statistical analysis precisely. However, obtaining highly sensitive data is quite difficult due to privacy reasons.Existing methods on privacy-preserving χ^2 testing such as MCIndep are a better choice when the number of samples n is large; however, we demonstrated that RandChiDist outperforms existing methods when n is small.

Future work will include evaluating other relevant datasets. We also plan to apply our method to other hypothesis testing methods such as Student's t-test and Fisher's exact test.

Authors' contributions
YS contributed to the study design, the conception, data acquisition, analysis, interpretation, writing, drafting, revision, and creation of software. AO supervised the study and contributed to the study design, the conception, analysis, interpretation, writing. All authors read and approved the final manuscript.

Funding
This work was supported by JSPS KAKENHI Grant Numbers JP17H04705, JP18H03229, JP18H03340, JP18K19835, JP19K12107, JP19H04113. This work was supported by JST, PRESTO Grant Number JPMJPR1934.

Availability of data and materials
All data generated or analysed during this study are included in this published article.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 24 August 2020 Accepted: 5 January 2021
Published online: 22 January 2021

References

1. Wu X, Dong H, Luo L, Zhu Y, Peng G, Reveille JD, Xiong M. A Novel Statistic for Genome-Wide Interaction Analysis. PLoS Genet. 2010;6(9):1001131.

2. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet. 2003;4(9):701–9.
3. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW, Egeland T, Dalen I, Mostad P, Hu Y, Fung W, Balding D, Clayton T, Whitaker J, Sparkes R, Gill P, Cowell R, Lauritzen S, Mortera J, Pearson J, Huentelman M, Halperin R, Tembe W, Melquist S, Bill M, Gill P, Curran J, Clayton T, Pinchin R, Jobling M, Gill P, Ladd C, Lee H, Yang N, Bieber F, Goodwin W, Linacre A, Vanezis P, Coble M, Just R, O'Callaghan J, Letmanyi I, Peterson C, Parsons T, Coble M, Just R, Irwin J, O'Callaghan J, Saunier J, Coble M, Vallone P, Just R, Coble M, Butler J, Parsons T, Kidd K, Pakstis A, Speed W, Grigorenko E, Kajuna S, Kennedy G, Matsuzaki H, Dong S, Liu W, Huang J, Macgregor S, Zhao Z, Henders A, Nicholas M, Montgomery G, Chakraborty R, Meagher T, Smouse P, Weir B, Triggs C, Starling L, Stowell L, Walsh K. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genet. 2008;4(8):1000167.
4. Dorfman R, Mamzer-Bruneel M-F, Vogt G, Hervé C, Izatt L, Jacobs C, Donaldson A, Brady A, Cuthbert A, Harrison R. Falling prices and unfair competition in consumer genomics. Nat Biotechnol. 2013;31(9):785–6.
5. Savage N. Privacy: The myth of anonymity. Nature. 2016;537(7619):70–72.
6. Wang R, Li YF, Wang X, Tang H, Zhou X. Learning your identity and disease from research papers: information leaks in genome wide association study. In: Proc. ACM CCS. New York City: Association for Computing Machinery; 2009. p. 534–44.
7. Yu F, Fienberg SE, Slavković AB, Uhler C. Scalable privacy-preserving data sharing methodology for genome-wide association studies,. J Biomed Informa. 2014;50:133–41.
8. Gaboardi M, woo Lim H, Rogers R, Vadhan S. Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing. In: Proc. ICML. Cambridge: Journal of Machine Learning Research, Inc.; 2016.
9. Pereira S, Gibbs R, McGuire A. Open Access Data Sharing in Genomic Research. Genes. 2014;5(3):739–47. https://doi.org/10.3390/genes5030739.
10. Terao C, Momozawa Y, Ishigaki K, Kawakami E, Akiyama M, Loh P-R, Genovese G, Sugishita H, Ohta T, Hirata M, Perry JRB, Matsuda K, Murakami Y, Kubo M, Kamatani Y. GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. Nat Commun. 2019;10(1):. https://doi.org/10.1038/s41467-019-12705-5.
11. Schmidt-Kastner R, Guloksuz S, Kietzmann T, van Os J, Rutten BPF. Analysis of GWAS-Derived Schizophrenia Genes for Links to Ischemia-Hypoxia Response of the Brain. Front Psychiatry. 2020;11:. https://doi.org/10.3389/fpsyt.2020.00393.
12. Lee K-Y, Leung K-S, Ma SL, So HC, Huang D, Tang NL-S, Wong M-H. Genome-Wide Search for SNP Interactions in GWAS Data: Algorithm, Feasibility, Replication Using Schizophrenia Datasets. Front Genet. 2020;11:. https://doi.org/10.3389/fgene.2020.01003.
13. Yuan J, Xing H, Lamy AL, Lencz T, Pe'er I. Leveraging correlations between variants in polygenic risk scores to detect heterogeneity in GWAS cohorts. PLOS Genet. 2020;16(9):. https://doi.org/10.1371/journal.pgen.1009015.
14. Armstrong J, Rudkin JK, Allen N, Crook DW, Wilson DJ, Wyllie DH, O'Connell AM. Dynamic linkage of COVID-19 test results between Public Health England's Second Generation Surveillance System and UK Biobank. Microb Genom. 2020;6(7):. https://doi.org/10.1099/mgen.0.000397.
15. Shelton JF, Shastri AJ, Ye C, Weldon CH, Filshtein-Somnez T, Coker D, Symons A, Esparza-Gordillo J, Team C, Aslibekyan S, Auton A. Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19 susceptibility and severity. medRxiv. 20202020–090420188318. https://doi.org/10.1101/2020.09.04.20188318.
16. Asselta R, Paraboschi EM, Mantovani A, Duga S. ACE2 and TMPRSS2 Variants and Expression as Candidates to Sex and Country Differences in COVID-19 Severity in Italy. SSRN Electron J. 2020. https://doi.org/10.2139/ssrn.3559608.
17. Galmés S, Serra F, Palou A. Current State of Evidence: Influence of Nutritional and Nutrigenetic Factors on Immunity in the COVID-19 Pandemic Framework. Nutrients. 2020;12(9):2738. https://doi.org/10.3390/nu12092738.
18. Das R, Ghate SD. Investigating the likely association between genetic ancestry and COVID-19 manifestations. medRxiv. 2020;20054627. https://doi.org/10.1101/2020.04.05.20054627.
19. Ren W-L, Wen Y-J, Dunwell JM, Zhang Y-M. pKWmEB: integration of Kruskal–Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. Heredity. 2018;120(3):. https://doi.org/10.1038/s41437-017-0007-4.
20. Casto AM, Feldman MW. Genome-Wide Association Study SNPs in the Human Genome Diversity Project Populations: Does Selection Affect Unlinked SNPs with Shared Trait Associations? PLoS Genet. 2011;7(1):. https://doi.org/10.1371/journal.pgen.1001266.
21. Couch S, Kazan Z, Shi K, Bray A, Groce A. Differentially private nonparametric hypothesis testing. In: Proc. ACM CCS. New York City: Association for Computing Machinery; 2019. p. 737–51.
22. Dwork C, McSherry F, Nissim K, Smith A. Calibrating Noise to Sensitivity in Private Data Analysis. In: Proc. Theory of Cryptography (TCC). Berlin: Springer; 2006. p. 265–84.
23. Ren H, Li H, Liang X, He S, Dai Y, Zhao L. Privacy-Enhanced and Multifunctional Health Data Aggregation under Differential Privacy Guarantees. Sensors. 2016;16(9):1463. https://doi.org/10.3390/s16091463.
24. Sei Y, Ohsuga A. Differential Private Data Collection and Analysis Based on Randomized Multiple Dummies for Untrusted Mobile Crowdsensing. IEEE Trans Inf Forensic Secur. 2017;12(4):926–39.
25. Liu Y, Wang H, Peng M, Guan J, Xu J, Wang Y. DeePGA: A Privacy-Preserving Data Aggregation Game in Crowdsensing via Deep Reinforcement Learning. IEEE Internet Things J. 2020. https://doi.org/10.1109/jiot.2019.2957400.
26. Ukil A, Jara AJ, Marin L. Data-Driven Automated Cardiac Health Management with Robust Edge Analytics and De-Risking. Sensors. 2019;19(12):2733–1273318. https://doi.org/10.3390/s19122733.
27. Fienberg SE, Slavkovic A, Uhler C. Privacy Preserving GWAS Data Sharing. In: Proc. IEEE International Conference on Data Mining Workshops. New York City: Institute of Electrical and Electronics Engineers; 2011. p. 628–35.
28. Uhlerop C, Slavković A, Fienberg SE, Uhler C, Slavković A, Fienberg SE. Privacy-Preserving Data Sharing for Genome-Wide Association Studies. J Privacy Confidentiality. 2013;5(1):137–66.

29. Gaboardi M, Rogers R. Local Private Hypothesis Testing: Chi-Square Tests. In: Proc. ICML. Cambridge: Journal of Machine Learning Research, Inc.; 2018. p. 1626–35.
30. Kohutek ZA, Wu AJ, Zhang Z, Foster A, Din SU, Yorke ED, Downey R, Rosenzweig KE, Weber WA, Rimner A. FDG-PET maximum standardized uptake value is prognostic for recurrence and survival after stereotactic body radiotherapy for non-small cell lung cancer. Lung Cancer. 2015;89(2):115–20.
31. and others, Shi SQ, White MJ, Borsetti HM, Pendergast JS, Hida A, Ciarleglio CM, De Verteuil PA, Cadar AG, Cala C, McMahon D. Molecular analyses of circadian gene variants reveal sex-dependent links between depression and clocks. Transl Psychiatry. 2017;6(3):748.
32. Möckel M, Schindler R, Knorr L, Müller C, Heller Jr G, Störk TV, Frei U. Prognostic value of cardiac troponin T and I elevations in renal disease patients without acute coronary syndromes: a 9-month outcome analysis. Nephrol Dial Transplant Off Publ Eur Dial Transplant Assoc Eur Ren Assoc. 1999;14(6):1489–95.
33. Kim JW, Jang B, Yoo H. Privacy-preserving aggregation of personal health data streams. PLoS ONE. 2018;13(11): 0207639. https://doi.org/10.1371/journal.pone.0207639.
34. Schieppati A, Henter JI, Daina E, Aperia A. Why rare diseases are an important medical and social issue. Lancet. 2008;371(9629):2039–41. https://doi.org/10.1016/S0140-6736(08)60872-7.
35. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Le Cam Y, Rath A. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28(2):165–73. https://doi.org/10.1038/s41431-019-0508-0.
36. Capponi A, Fiandrino C, Kantarci B, Foschini L, Kliazovich D, Bouvry P. A Survey on Mobile Crowdsensing Systems: Challenges, Solutions, and Opportunities. IEEE Commun Surv Tutorials. 2019;21(3):2419–65. https://doi.org/10.1109/COMST.2019.2914030.
37. Gao H, Xu H, Zhang L, Zhou X. A Differential Game Model for Data Utility and Privacy-Preserving in Mobile Crowdsensing. IEEE Access. 2019;7:128526–33. https://doi.org/10.1109/ACCESS.2019.2940096.
38. Bosu A, Carver JC, Bird C, Orbeck J, Chockley C. Process Aspects and Social Dynamics of Contemporary Code Review: Insights from Open Source Development and Industrial Practice at Microsoft. IEEE Trans Softw Eng. 2017;43(1):56–75.
39. Pantforder D, Vogel-Heuser B, Grams D, Schweizer K. Supporting Operators in Process Control Tasks–Benefits of Interactive 3-D Visualization. IEEE Trans Human-Machine Syst. 2016;46(6):895–907.
40. Mukherjee P, Jansen BJ. Information Sharing by Viewers Via Second Screens for In-Real-Life Events. ACM Trans Web. 2017;11(1):1–24.
41. Ren X, Yu CM, Yu W, Yang S, Yang X, McCann JA, Yu PS. LoPub : High-dimensional crowdsourced data publication with local differential privacy. IEEE Trans Inf Forensics Secur. 2018;13(9):2151–66. https://doi.org/10.1109/TIFS.2018.2812146. arXiv:1612.04350v2.
42. Torra V. Random dictatorship for privacy-preserving social choice. Int J Inf Secur. 20191–9. https://doi.org/10.1007/s10207-019-00474-7.
43. Grining K, Klonowski M, Syga P. On practical privacy-preserving fault-tolerant data aggregation. Int J Inf Secur. 2019;18(3):285–304. https://doi.org/10.1007/s10207-018-0413-5.
44. Kakizaki K, Fukuchi K, Sakuma J. Differential Privacy Based on Geometrical Interpretation of Chi-squared Testing. In: Computer Security Symposium. Tokyo: Information Processing Society of Japan; 2016. p. 1199–206.
45. Kakizaki K, Fukuchi K, Sakuma J. Differentially private chi-squared test by unit circle mechanism. In: Proc. ICML. Cambridge: Journal of Machine Learning Research, Inc.; 2017. p. 1761–70.
46. McSherry F, Talwar K. Mechanism Design via Differential Privacy. In: Proc. IEEE FOCS. New York City: Institute of Electrical and Electronics Engineers; 2007. p. 94–103.
47. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127.
48. Johnson A, Shmatikov V. Privacy-preserving data exploration in genome-wide association studies. In: Proc. ACM KDD. New York City: Association for Computing Machinery; 2013. p. 1079–87.
49. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our data, ourselves: privacy via distributed noise generation. In: Proc. Eurocrypt, vol. 4004. Berlin: Springer; 2006. p. 486–503.
50. Sei Y, Ohsuga A. Privacy-Preserving Chi-Squared Testing for Genome SNP Databases. In: Proc. 39th International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC); 2017. https://doi.org/10.1109/EMBC.2017.8037705.
51. Kifer D, Rogers R. A New Class of Private Chi-Square Tests. In: Proc. International Conference on Artificial Intelligence and Statistics (AISTATS). Cambridge: Journal of Machine Learning Research, Inc.; 2017. p. 991–1000.
52. Canonne CL, Kamath G, McMillan A, Smith A, Ullman J. The structure of optimal private tests for simple hypotheses. In: Proc. ACM STOC. New York City: Association for Computing Machinery; 2019. p. 310–21.
53. Csail MA, Diakonikolas I, Kane D, Rubinfeld R. Private Testing of Distributions via Sample Permutations. In: Proc. NeurIPS. La Jolla: Neural Information Processing Systems Foundation, Inc.; 2019. p. 10878–89.
54. Liu C, He X, Chanyaswad T, Wang S, Mittal P. Investigating Statistical Privacy Frameworks from the Perspective of Hypothesis Testing. In: Proc. PET. Warsaw: Sciendo; 2019. p. 233–54.
55. Atanassov E, Dimov IT. What Monte Carlo models can do and cannot do efficiently?,. Appl Math Model. 2008;32(8): 1477–500.
56. Cabin RJ, Mitchell RJ. To Bonferroni or Not to Bonferroni: When and How Are the Questions. Bull Ecol Soc Am. 2000;81(3):246–248.
57. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet. 2006;38(11):1251–60.
58. Pemberton TJ, Jakobsson M, Conrad DF, Coop G, Wall JD, Pritchard JK, Patel PI, Rosenberg NA. Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India. Ann Hum Genet. 2008;72(4):535–46.
59. Sharpe D. Your Chi-Square Test Is Statistically Significant: Now What? Pract Assess Res Eval. 2015;20(8):1–10.

60. Luo X, Xia H, Yang W, Wang B, Guo T, Xiong J, Jiang Z, Liu Y, Yan X, Zhou W, Ye L, Zhang B. Characteristics of patients with COVID-19 during epidemic ongoing outbreak in Wuhan, China. medRxiv. 20201–17. https://doi.org/10.1101/2020.03.19.20033175.
61. Poyiadi N, Cormier P, Patel PY, Hadied MO, Bhargava P, Khanna K, Nadig J, Keimig T, Spizarny D, Reeser N, Klochko C, Peterson EL, Song T. Acute Pulmonary Embolism and COVID-19. Radiology. 2020;201955:1–9. https://doi.org/10.1148/radiol.2020201955.
62. Jacob L, Smith L, Butler L, Barnett Y, Grabovac I, McDermott D, Armstrong N, Yakkundi A, Tully MA. COVID-19 Social Distancing and Sexual Activity in a Sample of the British Public. J Sex Med. 2020;17(7):1229–36. https://doi.org/10.1016/j.jsxm.2020.05.001.
63. Bearden WO, Sharma S, Teel JE. Sample Size Effects on Chi Square and Other Statistics Used in Evaluating Causal Models. J Mark Res. 1982;19(4):425–30. https://doi.org/10.1177/002224378201900404.
64. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures,. Psychol Bull. 1980;88(3):588–606. https://doi.org/10.1037/0033-2909.88.3.588.

Publisher's Note