Yu and Ji *BMC Medical Informatics and Decision Making* 2014, **14**(Suppl 1):S3
http://www.biomedcentral.com/1472-6947/14/S1/S3

BMC Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

# Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to iDASH healthcare privacy protection challenge

Fei Yu[1*], Zhanglong Ji[2]

*From* 3rd iDASH Privacy Workshop
San Diego, CA, USA. 24 March 2014

**Abstract**

In response to the growing interest in genome-wide association study (GWAS) data privacy, the Integrating Data for Analysis, Anonymization and SHaring (iDASH) center organized the *iDASH Healthcare Privacy Protection Challenge*, with the aim of investigating the effectiveness of applying privacy-preserving methodologies to human genetic data. This paper is based on a submission to the iDASH Healthcare Privacy Protection Challenge. We apply privacy-preserving methods that are adapted from Uhler et al. 2013 and Yu et al. 2014 to the challenge's data and analyze the data utility after the data are perturbed by the privacy-preserving methods. Major contributions of this paper include new interpretation of the $\chi^2$ statistic in a GWAS setting and new results about the Hamming distance score, a key component for one of the privacy-preserving methods.

## Introduction

Rapid developments in whole-genome sequencing technologies in recent years have made the collection of high quality genetic data faster and more economically feasible. Many types of genetic research can benefit from having a large amount of genetic data. For example, in genome-wide association studies (GWAS), which are a type of genetic research that examine a large number of single-nucleotide polymorphisms (SNPs) to identify genetic factors associated with a phenotype, which is typically a common disease, increasing the number of DNA samples available for analysis allows researchers to make more accurate statistical inference and improve the overall quality of the analysis.

Encouraging data sharing among researchers is the first step towards taking advantage of the benefits brought about by the rapid growth in genetic data collection. However, being able to share genetic data without compromising the study participants' privacy remains one of the biggest challenges in genetic research. While it is clear that individual level genetic data deserve a high level of protection, for many years it was widely considered safe to release to the public aggregate genetic data pooled from thousands of individuals without compromising genetic study participants' privacy. However, Homer et al. [1] in 2008 demonstrated that one can use publicly available aggregate genetic data, such as SNP data from the International HapMap Project http://hapmap.ncbi.nlm.nih.gov/, to infer whether an individual has participated in a study. Cautious about the potential breach of genetic study participants' privacy, the National Institute of Health (NIH) quickly responded to the Homer et al. [1] attack by mandating an elaborate approval process that every researcher has to go through in order to gain access to aggregate genetic data. This NIH policy remains in effect today.

Homer et al. [1]'s attack and NIH's subsequent reaction spurred research interest in privacy-preserving methodologies for GWAS data. A recent concept of *differential privacy* (e.g. [2]), introduced by the cryptographic community,

* Correspondence: feiy@stat.cmu.edu
[1] Department of Statistics, Carnegie Mellon University, 5000 Forbes Ave, PA 15213, Pittsburgh, PA, USA
Full list of author information is available at the end of the article

has shown great promise as a basis for privacy-preserving methodologies, as it provides a rigorous definition of privacy with meaningful privacy guarantees in the presence of arbitrary external information. We have seen privacy-preserving methods based on differential privacy applied to real human GWAS data in recent studies (e.g., [3-5]).

The iDASH Healthcare Privacy Protection Challenge, organized by *Integrating Data for Analysis, Anonymization and SHaring* (iDASH), aims to investigate the effectiveness of applying privacy-preserving methodologies to human genetic data [6]. This paper is based on a submission to the iDASH Healthcare Privacy Protection Challenge using privacy-preserving methods adapted from [3] and [5].

A major contribution of this paper is a new interpretation of the $\chi^2$ statistic in a GWAS setting and new results about the Hamming distance score, which plays an important role in the differentially private mechanisms proposed by [4] and [5]. In particular, we present a graphical interpretation of the allelic test $\chi^2$ statistic that will help us conceptualize the Hamming distance score. We also device an efficient algorithm for finding the Hamming distance score and prove that the *sensitivity* of the score function is 1; we hence address concerns raised in [5] about speed and sensitivity of alternative methods for finding the Hamming distance score.

We start by introducing background information on the iDASH Healthcare Privacy Protection Challenge. We briefly describe the characteristics of the data and define the allelic test $\chi^2$ statistic, which is used for evaluating the performance of submissions in the challenge. Then we summarize differentially private mechanisms applied to the challenge's data, which include a mechanism based on the Laplace mechanism and $\chi^2$ statistic, a mechanism based on the exponential mechanism and $\chi^2$ statistic, and a mechanism based on the exponential mechanism and Hamming distance score. We present a graphical interpretation of the allelic test $\chi^2$ statistic and an efficient algorithm for finding the Hamming distance score. We prove that our algorithm finds the shortest Hamming distance and therefore the Hamming distance score has sensitivity 1. We incorporate our improvements into the differentially private mechanisms and apply them to the challenge's data. We compare the performance of the mechanisms using risk-utility plots.

## Background information on iDASH challenge

The challenge has two tasks, both of which are concerned with the dissemination of aggregate GWAS data: (1) limiting the re-identification risks when releasing all aggregate data in a GWAS dataset, and (2) being compliant with differential privacy (Definition 2) when releasing the most significant SNPs. This paper focuses on the second task of releasing the most significant SNPs differentially privately.

The data used for the second task consist of 201 participants from the Personal Genome Project (http://www.personalgenomes.org/) and 174 participants from HapMap. Individuals from PGP are treated as cases and those from HapMap are treated as controls in the challenge. 106,129 SNPs are typed in all participants. [6] has more details on how the data are processed.

A subset containing 5,000 SNPs is selected by organizers of the challenge to form a representative sample of the entire set of SNPs. This paper uses the subset of SNPs to evaluate the performance of the privacy-preserving methods, as is recommended by organizers of the challenge.

In GWAS with $R$ cases and $S$ controls, we usually summarize the data for a single SNP using a 2 × 3 genotype contingency table shown in Table 1 or a 2 × 2 allelic contingency table shown in Table 2. In this challenge, we are given genotypes of individuals in the case group and allele frequencies of individuals in the control group. Therefore, the row pertaining to the case group in the allelic table can be easily derived from the genotypes of the case group. Furthermore, we will assume that the Hardy-Weinberg equilibrium holds so that we can derive the row pertaining to the control group in the genotype table from the allele frequencies of the control group.

In this challenge, the statistical significance of a SNP's association with the phenotype is assessed by the the allelic test statistic (Definition 1). For the rest of the paper, we will simply refer to the allelic test statistic as $\chi^2$ statistic. Assuming that the control group's data are public, we will use the differentially private mechanisms discussed in the next section to release the top $K$ SNPs while preserving the privacy of the case group.

**Definition 1** *The allelic test is also known as the Cochran-Armitage trend test for the additive model. The allelic test statistic based on a genotype contingency table (Table* 1*) is equivalent to the* $\chi^2$*-statistic based on the corresponding allelic contingency table (Table* 2*). The allelic test statistic can be written as*

$$Y_A = \frac{2N[(2r_0 + r_1)S - (2s_0 + s_1)R]^2}{RS(2n_0 + n_1)(n_1 + 2n_2)}$$

## Differential privacy: definitions and methods

The concept of differential privacy, recently introduced by the cryptographic community (e.g., [2]), provides a

**Table 1 Genotype table**

| | # of minor alleles | | | Total | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | | |
| Case | $r_0$ | $r_1$ | $r_2$ | R | |
| Control | $s_0$ | $s_1$ | $s_2$ | S | |
| Total | $n_0$ | $n_1$ | $n_2$ | | N |

**Table 2 Allelic table**

| | Allele type | | Total |
|---|---|---|---|
| | Minor | Major | |
| Case | $r_1 + 2r_2$ | $2r_0 + r_1$ | 2R |
| Control | $s_1 + 2s_2$ | $2s_0 + s_1$ | 2S |
| Total | $n_1 + 2n_2$ | $2n_0 + n_1$ | 2N |

notion of privacy guarantees that protect GWAS databases against arbitrary external information.

**Definition 2** (differential privacy) *Let* $\mathcal{D} = \{(X_1, \ldots, X_n) : X_i \sim \mathcal{P}\}$ *denote the set of all databases consisting of n individuals sampled independently from the same population* $\mathcal{P}$. *For* $D, D' \in \mathcal{D}$, *write* $D \sim D'$ *if D and D' differ in one individual. A randomized mechanism* $\mathcal{K}$ *is* ∈-differentially private *if, for all* $D \sim D'$ *and for any measurable set* $S \subset \mathbb{R}$,

$$\frac{Pr(\mathcal{K}(D) \in S)}{Pr(\mathcal{K}(D') \in S)} \leq e^{\varepsilon}.$$

Two methods are often used as building blocks for constructing more complex differentially private algorithms. One of the methods, due to [2], is called the *Laplace mechanism* (Definition 4), and the other method, due to [7], is called the *exponential mechanism* (Definition 5). Both methods require knowledge of the *sensitivity* of the score function, where sensitivity is defined as the smallest upper bound of how much the function can vary when one record in the input database changes (see Definition 3).

**Definition 3** *The* sensitivity *of a function* $f : \mathcal{D} \times \mathbb{R}^d \to \mathbb{R}$ *is the smallest number S(f) such that*

$$\sup_{x \in \mathbb{R}^d} ||f(D, x) - f(D', x)|| \leq S(f),$$

*for all databases* $D, D' \in \mathcal{D}$ *such that* $D \sim D'$.

**Definition 4** (Laplace mechanism) *Releasing* $f(D)+b$, *where* $b \sim Laplace\left(0, \frac{S(f)}{\varepsilon}\right)$, *satisfies the definition of* ∈*-differential privacy.*

**Definition 5** (exponential mechanism) *Let* $q : \mathcal{D} \times \mathbb{R}^d \to \mathbb{R}$ *be a function that outputs the score of an event or a value given a database. Define the random variable* $\varepsilon_q^{\varepsilon}$

$$\Pr(\varepsilon_q^{\varepsilon}(D) = x) = \frac{\exp\left(\frac{\varepsilon q(D,x)}{2S(q)}\right)}{\int_{\mathbb{R}^d} \exp\left(\frac{\varepsilon q(D,s)}{2S(q)}\right) \mathrm{d}s}.$$

*Then releasing* $\varepsilon_q^{\varepsilon}$ *satisfies the definition of* ∈*-differential privacy.*

## Methods for releasing the *K* most relevant SNPs

**Algorithm 1** The *E*-differentially private mechanism for releasing the *K* most relevant SNPs using the *Laplace mechanism* [3,5,8].

**Input:** The score of all *M* candidate SNPs, the number of SNPs, *K*, that we want to release, the sensitivity, *s*, of the score function, and the privacy budget *E*.

**Output:** *K* SNPs.

1: Add independent Laplace noise with mean zero and scale $\frac{2Ks}{\varepsilon}$ to each of the *M* SNPs scores.

2: Choose the top *K* SNPs based on the perturbed scores.

**Algorithm 2** The ∈-differentially private mechanism for releasing the *K* most relevant SNPs using the *exponential mechanism* [4,5].

**Input:** The scores (e.g. $\chi^2$ statistic or Hamming distance) of all *M* candidate SNPs, the number of SNPs, *K*, that we want to release, the sensitivity, *s*, of the score function, and the privacy budget ∈.

**Output:** *K* SNPs.

1: Initialize $\{q_i\}_{i=1}^{M}$ score of $SNP_i$.

2: Set $w_i = \exp\left(\frac{\varepsilon q_i}{2Ks}\right)$. Define $\Pr(\mathcal{T}(D) = i) = w_i \Big/ \sum_{j=1}^{M} w_j$.

3: Sample $j \sim \mathcal{T}(D)$. Record $SNP_j$. Set $q_j = -\infty$.

4: Repeat Step 2 and 3 until *K* SNPs have been recorded.

Algorithm 1 and Algorithm 2 extend the Laplace mechanism and the exponential mechanism, respectively, to release more than a single SNP differentially privately. In this paper, we consider three mechanisms for releasing the top *K* SNPs: a mechanism that is based on Algorithm 1 and uses $\chi^2$ statistic as score function, a mechanism that is based on Algorithm 2 and uses $\chi^2$ statistic as score function, and a mechanism that is based on Algorithm 2 and uses the Hamming distance score ([4]) as score function. In loose terms, the Hamming distance score is the smallest number of changes made to a genotype table until the significance of the table changes, where a change, counted as 1-Hamming distance in the space of genotype tables, is defined as changing the genotype of one individual and significance refers to whether the *p*-value of the $\chi^2$ statistic of the table is smaller than a pre-specified threshold value or not. See [5] for more details on the three mechanisms and applications of them to a real human GWAS dataset.

For mechanisms that use the $\chi^2$ statistic as score, we need to know the sensitivity of the $\chi^2$ statistic. An upper bound for the sensitivity is shown in [5], but [5] requires that the margins of the genotype contingency tables to be positive. Indeed, such requirement can be satisfied in the challenge's setting when we assume that Hardy-Weinberg equilibrium holds: because a typical GWAS dataset consists of only common SNPs, whose minor allele frequencies are greater than 1%, the control group's three genotypes derived from the allele frequency at each SNPs

will be nonnegative, which ensures that the derived genotype tables have positive margins.

For the mechanism that uses the Hamming distance score as score, we already know that, by construction, the sensitivity of the score function is 1 if the Hamming distance is the shortest Hamming distance [4]. However, as is pointed out in [4] and [5], it is a computationally onerous task to actually calculate the shortest Hamming distance, which, in the most naïve setting, involves examining all possible sequential changes made to the original genotype table that alter the significance status of the table. To make the calculations more computationally feasible, [4] and [5] use approximations of the shortest Hamming distance in their implementations of the mechanism, noting the caveat that the sensitivity of the approximated Hamming distance score may no longer be 1.

In the next section, we propose a new method of finding the Hamming distance score that is much more computationally efficient than those in [4] and [5]. We also prove that our method indeed produces the shortest Hamming distance, and therefore the sensitivity of the resulting Hamming distance score function is 1.

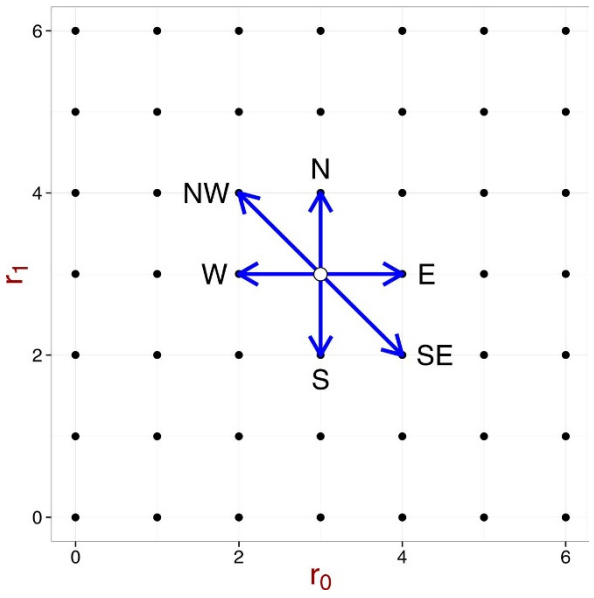


**Figure 1 Legal moves in the space of genotype tables with fixed *R*, *S*, $s_0$, $s_1$, and $s_2$.**

## Finding the Hamming distance score

Let's refer to the case group's data and the control group's data collectively as a database and call the data for an individual a record. We can think of the number of cases, $R$, and the number of controls, $S$ as fixed. Recall that we assume the control group's data are known to the public. Therefore, for a given genotype table, we assume that $s_0$, $s_1$, and $s_2$ are fixed. Then the $\chi^2$ statistic can be written as a function of $r_0$ and $r_1$. How the value of the $\chi^2$ statistic changes when we change one record in the database is illustrated in Figure 1. In Figure 1, each dot represent a value of the $\chi^2$ statistic given $r_0$ and $r_1$. When we change one record in the case group, there are 6 possible changes to the genotype table: $(r_0 \to r_0 + 1,\ r_1 \to r_1)$, $(r_0 \to r_0 + 1,\ r_1 \to r_1 - 1)$, $(r_0 \to r_0,\ r_1 \to r_1 - 1)$, $(r_0 \to r_0 - 1,\ r_1 \to r_1)$, $(r_0 \to r_0 - 1,\ r_1 \to r_1 + 1)$, and $(r_0 \to r_0,\ r_1 \to r_1 + 1)$; that is, $r_0$ and $r_1$ cannot increase or decrease by 1 simultaneously. The possible changes are shown as arrows in Figure 1. A change in the genotype table results in a change in the allelic table, and we get a new value for the $\chi^2$ statistic based on the new allelic table.

Let $p^*$ denote a pre-specified threshold $p$-value and let $c$ denote the $\chi^2$ statistic corresponding to $p^*$, the $p$-value of the $\chi^2$ distribution with 1 degree of freedom. Then for a given SNP in the pool of candidate SNPs, the genotype table of which we denote by $D$, the shortest Hamming distance is the smallest number of sequential changes made to $D$ such that the resulting genotype table, $D'$, satisfies $Y_A(D') \geq c$ if $Y_A(D) < c$ and $Y_A(D') < c$ if $Y_A(D) \geq c$; that is, if we call $c$ the significance threshold, then the goal is to make changes to the "insignificant" ("significant") table $D$ so that the $\chi^2$ statistic of $D'$ goes above (below) the significance threshold $c$, and $D'$ becomes a "significant" ("insignificant") table. The Hamming distance score is defined as $h$ = (shortest Hamming distance) − 1 if $Y_A(D) \geq c$ and $h$ = −(shortest Hamming distance) if $Y_A(D) < c$.

Let's consider the space of genotype tables, $\mathcal{B}_D$, defined by a genotype table $D$: for all $D' \in \mathcal{B}_D$, $D'$ shares the same values of $s_0$, $s_1$, $s_2$, $R$, $S$, and $N$ with $D$, but $D'$ does not necessarily have the same values of $r_0$, $r_1$, and $r_2$ as $D$. Let $n_{10} = 2s_0 + s_1$ denote the number of major alleles in the control group, and let $x = 2r_0 + r_1$ denote the number of major alleles in the control group, then we can write the $\chi^2$ statistic of a genotype table $D' \in \mathcal{B}_D$ as a function of $x$:

$$Y_A(D; \mathcal{B}_D) = Y(r_0, r_1 : D) = Y_A(x; D) = \frac{2N(xS - n_{10}R)^2}{RS(x + n_{10})(2N - x - n_{10})},$$

where $r_0$, $r_1$ and $x$ are derived from $D'$, and $n_{10}$, $R$, $S$ and $N$ are the same for $D$ and $D'$. For notational convenience, when $r_0$ and $r_1$ are also derived from $D$, we will simply write the $\chi^2$ statistic as $Y_A(D)$.

**Lemma 1** *$Y_A$ is an increasing function of × when $xS - n_{10}R > 0$, and it is a decreasing function of × when $xS - n_{10}R < 0$.*

*Proof.* [see Additional file 1].

To understand the implication of Lemma 1, let's consider Figure 2. In Figure 2, each dashed line has *slope* = −2, representing a value of $x$, which is defined as $x = 2r_0 + r_1$. Because we can consider each dot in Figure 2 to be

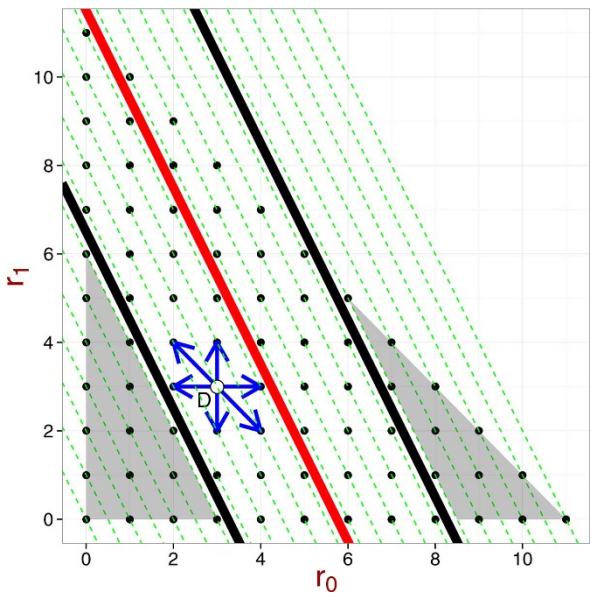


**Figure 2 An example of a genotype table, *D*, in the space of genotype tables with fixed *R*, *S*, $s_0$, $s_1$, and $s_2$.** Each dot represent a genotype table. Each dashed line has *slope* = fl2, representing the lines $x = 2r_0 + r_1$. The red line is $x = (2s_0 + s_1)R/S = 2r_0 + r_1$, and the two black lines correspond to values of $(2r_0 + r_1)$ such that $Y_A(r_0, r_1; \mathcal{B}_D) = c$, where $c$ is a pre-specified significance threshold value.

a unique genotype table in the space of genotype tables with fixed control data and a fixed number of cases, those tables that lie on the same dashed line will have the same value of $\chi^2$ statistic. Furthermore, because $0 \le r_0 + r_1 \le R$, $r_0 \ge 0$ and $r_1 \ge 0$, the space of genotype tables, represented as dots, fall within a triangle in Figure 2.

For the moment, let's treat $r_0$ and $r_1$ as continuous values. In Figure 2, the red solid line represents the line $2r_0 + r_1 = x = n_{10}R/S$ and the two solid black lines represent lines $2r_0 + r_1 = x$ such that $Y_A(D; \mathcal{B}_D) = c$. There are two black lines because by Lemma 1 $Y_A(x)$ is an increasing function when $x > n_{10}R/S$ and it is a decreasing function when $x < n_{10}R/S$; that is, there could be two values of $x$, say $x_1$ and $x_2$, such that $Y_A(x_1; \mathcal{B}_D) = Y_A(x_2; \mathcal{B}_D)$ and $x_1 < n_{10}R/S < x_2$. Because it is possible that

$$\begin{aligned} &\max_x Y_A(x; D) \\ &\le \max\{Y_A(0, \mathcal{B}_D), Y_A(2R, \mathcal{B}_D)\} \\ &= \max\left\{\frac{2NRn_{10}}{S(2N - n_{10})}, \frac{2NR(2S - n_{10})}{S(2R + n_{10})}\right\} < c, \end{aligned}$$

there could be genotype tables for which only one black line exists or no black line exists at all; in such cases, we will use the lines $0 = 2r_0 + r_1$ or $2R = 2r_0 + r_1$ wherever appropriate.

In Figure 2, the genotype table $D$ is insignificant and its $\chi^2$ statistic is below the threhold value. By Lemma 1, we know that the $\chi^2$ statistics of genotype tables, as represented by the dots on Figure 2, are greater than $c$ when they are in the shaded area, outside of the area between the two black lines and they are smaller than $Y_A(D^*)$ when they are inside the area between the two black lines. Therefore, finding the Hamming distance score for $D$ is to find the shortest Hamming distance from the genotype table $D$ to genotype tables in the shaded areas.

For genotype tables that are significant, they will fall into the shaded areas in Figure 2. Then finding the Hamming distance score for a significant genotype table is to find the shortest Hamming distance from the genotype table in one of the shaded areas to genotype tables in the non-shaded area.

**Proposition 2** *Given a significance threshold value c and an insignificant genotype table D (i.e., $Y_A(D) < c$), if there exists $D' \in \mathcal{B}_D$ such that $Y_A(D'; \mathcal{B}_D) \ge c$, then the shortest Hamming distance is* $\min\{H_1, H_2\}$, *where $H_1$ and $H_2$ are defined as follows:*

*(i) $H_1$ is the number of changes made to D in the following manner: (1) keep decreasing $r_0$ until the new genotype table, D', becomes significant (i.e., $Y_A(D'; \mathcal{D}) > c$); (2) when $r_0$ is minimized but the new table is still insignificant, keep decreasing $r_1$ until the new table becomes significant.*
*(ii) $H_2$ is the number of changes made to D in the following manner: (1) keep increasing $r_0$ until the new genotype table becomes significant; (2) if $r_0$ can no longer be increased without decreasing $r_1$ and the new table is still insignificant, increase $r_0$ and decrease $r_1$ in each change until the new table becomes significant.*

*If for all $D' \in \mathcal{B}_D$, $Y_A(D'; \mathcal{B}_D) < c$, then we define the shortest Hamming distance as* $\min\{H_1', H_2'\}$, *where $H_1'$ and $H_2'$ are defined as follows:*

*(i) When $r_0$ and $r_1$ are both minimized but the new table is still insignificant, set $H_1'$ to $1 + d_1$, where $d_1$ is smallest the number of changes needed to minimize $r_0$ and $r_1$.*
*(ii) When $r_0$ and $r_1$ are both maximized but the new table is still insignificant, set $H_2'$ to $1 + d_2$, where $d_2$ is smallest the number of changes needed to maximize $r_0$ and $r_1$.*

*Proof.* [see Additional file 1].

**Proposition 3** *Given a significance threshold value c and a significant genotype table D (that is, $Y_A(D) \geq c$), the shortest Hamming distance is* $\min\{H_1, H_2\}$, *where $H_1$ and $H_2$ are defined as follows:*

*(i) If $2r_0 + r_1 > (2s_0 + s_1)R/S$, set $H_1 = \infty$; otherwise, $H_1$ is the number of changes made to D in the following manner: keep decreasing $r_0$ until the new genotype table, D', becomes insignificant (i.e., $Y_A(D', D) < c$).*
*(ii) If $2r_0 + r_1 < (2s_0 + s_1)R/S$, set $H_2 = \infty$; otherwise, $H_2$ is the number of changes made to D in the following manner: keep decreasing $r_0$ until the new genotype table becomes insignificant.*

*Proof* The proof is similar to that of Proposition 2.

**Definition 6** (The Hamming distance score) *Given a threshold $\chi^2$ statistic value c and a genotye table D, the Hamming distance score of D is*

$$h \begin{cases} -d^-, & \text{if } Y_A(D) < c, \\ d^+ - 1 & \text{if } Y_A(D) \geq c, \end{cases}$$

*where $d^-$ is found using Proposition 2 and $d^+$ is found using Proposition 3.*

**Corollary 4** *The sensitivity of the Hamming distance score as defined in Definition 6 is 1.*

## Application to the challenge's data

In this section we apply all three differentially private mechanisms to the challenge's data and evaluate the performace of the mechanims by examining the data utility at several levels of privacy risk. Data utility is defined as follows: let $\mathcal{S}_0$ denote the set of top $K$ SNPs ranked according to their true $\chi^2$ statistics and let $\mathcal{S}$ be the set of top $K$ SNPs chosen after perturbation (either by Algorithm 1 or Algorithm 2). Then the data utility as a function of the privacy budget, $\in$, is

$$u(\varepsilon) = \frac{|\mathcal{S}_0 \cap \mathcal{S}|}{|\mathcal{S}_0|}.$$

In Figure 3 we compare the performace of the mechanisms given different privacy budget $E$ and different number of top SNPs to release, $K$. For the mechanism based

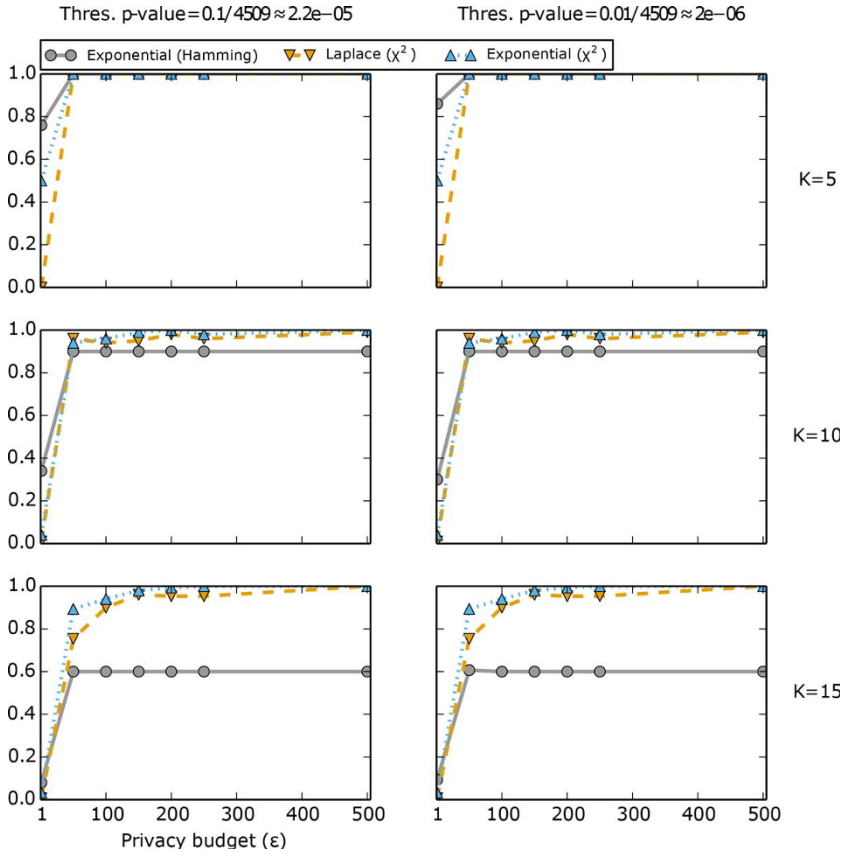


**Figure 3 Risk-utility plots**. Performance comparison of Algorithm 1 and Algorithm 2 with $\chi^2$ statistic or Hamming distance score as score function. Each row corresponds to a fixed $K$, the number of top SNPs to release. Each column corresponds to a fixed threshold $p$-value, which is relevant to the mechanism based on Hamming distance score only. The threshold $p$-values 0.1 and 0.01 divided by the number of SNPs.

on the Hamming distance score, we also consider different significance threshold values. We can see that the mechanism based on Algorithm 2 (a generalization of the exponential mechanism) and the Hamming distance score outperforms the other mechanisms when $\in$ is small ($\in$ = 1); on the other hand, unlike the other mechanisms, the data utilities of which continue to increase as $E$ increases, the data utility of the mechanism based on Algorithm 2 and the Hamming distance score may plateau before it reaches 1 even if we keep increasing $E$. This phenomenon is also observed in the analysis of a different GWAS dataset in [5]. The abnormality of the mechanism based on the Hamming distance score is due to the inconsistency in ranking: because the set of top $K$ SNPs based on the Hamming distance score is not always the same as the set of top $K$ SNPs based on the $\chi^2$ statistic, which is used to evaluate utility, therefore, as $E$ increases, the amount of noise decreases, and the set of $K$ SNPs resulting from the mechanism based on the Hamming distance score becomes more similar to the set of top $K$ SNPs based on the Hammming distance score, which may depart from the set of SNPS based on the $\chi^2$ statistic. [5] has a more detailed discussion of the characteristics of all three differentially private mechanisms.

It is also worth noting that even though the performance of the mechanism based on the Hamming distance score does not seem to be affected by the choice of threshold $p$-value, the analysis of the mechanism in [5] shows that whether the choice of threshold $p$-value has any effect on data utility also depends on the choice of $K$, the number of top SNPs to release. Therefore, the choice of threshold $p$-value should be justified before we use this mechanism.

## Conclusions

In our submission to the iDASH Healthcare Privacy Protection Challenge, we apply differentially-private methods proposed by [3] and [5] to the challenge's data. Our results show that the performance of the method based on Algorithm 2 and Hamming distance score is superior to that of other methods when the privacy budget, $\in$, is small. But we also point out problems with the Hamming distance score, such as the data utility plateauing at a level lower than other methods.

We devise an efficient algorithms for finding the Hamming distance score and prove that the sensitivity of the score function is 1. This addresses concerns raised in [5] regarding speed and sensitivity of alternative methods for finding the Hamming distance score. The graphical interpretation of the $\chi^2$ statistic that we present in the paper is instrumental in our discovery of the efficient algorithm for finding the Hamming distance score. We expect that the graphical interpretation can be extrapolated to other settings, such as the Pearson's $\chi^2$ statistic for 2 × 3 contingency tables and the setting in which data for the controls are not assumed to be public, and help with designing efficient algorithms for fining the Hamming distance score in those settings.

## Additional material

**Additional file 1: Proofs**

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

Fei Yu participated in the competition, performed the statistical analysis, and drafted the manuscript. Fei Yu and Zhanglong Ji both made significant contributions to devising the method for finding the Hamming distance score. Zhanglong Ji also participated in the revision of the manuscript. All authors read and approved the final manuscript.

#### Acknowledgements

This research was partially supported by NSF Awards EMSW21-RTG and BCS-0941518 to the Department of Statistics at Carnegie Mellon University. Publication of this article has been funded by iDASH(U54HL108460), NHGRI (K99HG008175), NLM(R00LM011392,R21LM012060), CTSA(UL1TR000100), and an NCBC-linked grant (R01HG007078).
This article has been published as part of *BMC Medical Informatics and Decision Making* Volume 14 Supplement 1, 2014: Critical Assessment of Data Privacy and Protection (CADPP). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcmedinformdecismak/supplements/14/S1.

#### Authors' details

[1] Department of Statistics, Carnegie Mellon University, 5000 Forbes Ave, PA 15213, Pittsburgh, PA, USA. [2]Department of Computer Science and Engineering, University of California, San Diego, CA 92092, La Jolla, CA, USA.

Published: 8 December 2014

#### References

1. Nils Homer, Szelinger Szabolcs, Redman Margot, Duggan David, Tembe Waibhav, Muehling Jill, Pearson VJohn, Stephan ADietrich, Nelson FStanley, Craig WDavid: **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays.** *PLoS Genetics* 2008, **4(8)**:e1000167.
2. Dwork Cynthia, McSherry Frank, Nissim Kobbi, Smith Adam: **Calibrating noise to sensitivity in private data analysis.** *Theory of Cryptography* 2006, 1-20.
3. Uhler Caroline, Slavkovic BAleksandra, Fienberg EStephen: **Privacy-preserving data sharing for genome-wide association studies.** *Journal of Privacy and Confidentiality* 2013, **5(1)**:137-166.
4. Johnson Aaron, Shmatikov Vitaly: **Privacy-preserving data exploration in genome-wide association studies.** *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2013, 1079-1087.
5. Yu Fei, Fienberg EStephen, Slavković BAleksandra, Uhler Caroline: **Scalable privacy-preserving data sharing methodology for genome-wide association studies.** *Journal of biomedical informatics* 2014, **50C**:133-141.
6. Jiang Xiaoqian, Zhao Yongan, Wang Xiaofeng, Malin Bradley, Wang Shuang, Ohno-Machado Lucila, Tang Haixu: **A community assessment of privacy preserving techniques on human genome data.** *BMC* 2014.
7. McSherry Frank, Talwar Kunal: **Mechanism Design via Differential Privacy.** *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)* 2007, 94-103.

8. Bhaskar Raghav, Laxman Srivatsan, Smith Adam, Thakurta Abhradeep: **Discovering frequent patterns in sensitive data.** *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10* New York, New York, USA, ACM Press; 2010, 503.

doi:10.1186/1472-6947-14-S1-S3
**Cite this article as:** Yu and Ji: Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to iDASH healthcare privacy protection challenge. *BMC Medical Informatics and Decision Making* 2014 **14**(Suppl 1):S3.