

# Differential Privacy for Clinical Trial Data: Preliminary Evaluations

Duy Vu  
Department of Statistics  
The Pennsylvania State University  
University Park, USA  
dqv100@stat.psu.edu

Aleksandra Slavković  
Department of Statistics  
The Pennsylvania State University  
University Park, USA  
sesa@stat.psu.edu

**Abstract**—The concept of differential privacy as a rigorous definition of privacy has emerged from the cryptographic community. However, further careful evaluation is needed before we can apply these theoretical results to privacy preservation in everyday data mining and statistical analysis. In this paper we demonstrate how to integrate a differential privacy framework with the classical statistical hypothesis testing in the domain of clinical trials where personal information is sensitive. We develop concrete methodology that researchers can use. We derive rules for the sample size adjustment whereby both statistical efficiency and differential privacy can be achieved for the specific tests for binomial random variables and in contingency tables.

**Index Terms**—Differential; Privacy; Clinical; Trials

## I. INTRODUCTION

Clinical trials test novel therapies and confirm findings from earlier studies [1]. Typically many confirmatory studies and careful meta-analyses are required to produce changes to medical practice or public policy. To improve the quality of research and of health care, new policies are aimed at promoting clinical trial transparency through clinical trial registries and public sharing of results, and in some cases sharing of data sets. ClinicalTrials.gov ([2], [3]) is currently the largest registry in the world; it warehouses more than 78,900 trials sponsored by government agencies and private industry. As the number and size of data registries grow, data mining tools enable rapid extraction, summary, and derivation of knowledge from the stored clinical information (e.g., [4], [5], [6]). However, results of statistical analyses and of data mining algorithms applied to such data registries may violate privacy [7].

Privacy-preserving data mining (e.g., [8]) and statistical disclosure limitation (e.g., [9]) both focus on obtaining valid statistical results while minimizing the loss of privacy for the individuals and organizations whose data are stored in statistical databases. Proposed schemes in both communities typically lack formal privacy guarantees and are only designed to deal with specific kinds of attacks. The emerging *differential privacy* framework claims precise guarantees on privacy in the presence of arbitrary side information [10], [11]. Some very recent developments aim to establish connections between differential privacy and traditional statistical inference; e.g., see [11] for private parametric estimation, [12] for link to robust statistics, and [13] for approximation of smooth densities.

In this paper we build on the theory of private parametric estimation proposed by [11], and focus on statistical elements relevant to the design and analysis of clinical trial data. We evaluate how private and non-private estimators compare for parametric exponential families in terms of their bias and asymptotic efficiency. Related to the additive perturbation techniques of [14] and [15], we present preliminary work aimed at developing concrete methodology that data analysts can use. More specifically, we propose *approximate* and *exact* sample size adjustment factors that are needed for sample size calculation and power analysis in classical hypothesis testing.

When data are already available, as in data mining applications, and multiple sources are combined, procedures proposed in this paper provide the lower bounds on the realistic sample size needed to achieve the same statistical power as the true sample size would if there was no infusion of noise for privacy protection. On the other hand, many funding agencies and ethics boards frequently request that a power analysis be completed before a study is conducted, or before a study's results are published. We propose sample size adjustment factors that will allow researchers to a priori determine the sample size needed to produce the desired power and “private” estimates for a set of specific statistical hypotheses.

We begin Section II with definitions of differential privacy, asymptotic efficiency, confidence level and power of a hypothesis test, and Pearson  $\chi^2$  test of independence. In Section III, we propose modifications to private parametric estimation described in [11] in order to avoid the biases that may come from estimates of smaller-size subsets of data. In Section IV, we present a preliminary investigation into calculating sample size under the differential privacy framework. To demonstrate the applicability of the proposed methodology, we outline our results for two types of basic hypothesis tests: (1) the test for a single proportion and (2) Pearson's  $\chi^2$  test of independence. For more details and other tests including the Bayesian setting, see the extended version of this paper [21]. The simulations in Section V demonstrate the interactive effects of sample size and privacy level  $\epsilon$ , and compare the asymptotic efficiency of the Maximum Likelihood Estimator (MLE) and a corresponding differentially private (DP) estimator. In the final section, we summarize our findings and point to ongoing related research.

## II. DEFINITIONS

**Differential Privacy.** Following [11], [10], and [16], we say that two data sets  $\mathbf{x} = (x_1, x_2, \dots, x_N)$  and  $\mathbf{x}' = (x'_1, x'_2, \dots, x'_N)$  are neighbors if and only if they are different at only one sample, i.e. we can rearrange  $\mathbf{x} = (x_1, x_2, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_N)$  and  $\mathbf{x}' = (x_1, x_2, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_N)$  for some  $i \in \{1 \dots N\}$ .

A statistic  $T(\cdot)$  is  $\epsilon$ -differentially private if for all neighboring data sets  $\mathbf{x}, \mathbf{x}'$ , and for all measurable subsets  $A$ :

$$\frac{P(T(\mathbf{x}) \in A)}{P(T(\mathbf{x}') \in A)} \leq e^\epsilon \quad (1)$$

where the parameter  $\epsilon > 0$  is a measure of the information leakage.

**Asymptotic Efficiency.** This study will focus on random variable  $X$  from the exponential family:

$$f(x, \theta) = \exp\left(\sum_i \theta_i T_i(x) - K(\theta)\right) \quad (2)$$

where  $T_i(x)$ s are sufficient statistics,  $\theta_i$ 's are canonical parameters, and  $K(\theta)$  is the normalizing constant. The Maximum Likelihood (ML) estimate for  $\theta_i$  is obtained by maximizing the likelihood function  $L(\mathbf{x}, \theta) = \prod_{k=1}^N f(x_k, \theta)$ , with  $N$  being the finite sample size. A classical result in statistics [17] states that an ML estimator has the following asymptotic property:

**Theorem (Cram er):** Let  $x_1, x_2, \dots, x_N$  be independently identically distributed with density  $f(x|\theta), \theta \in \Theta$  and let  $\theta_0$  denote the true value of  $\theta$ . Let the ML estimator of  $\theta_0$  be  $T_N(\mathbf{x})$ . Under appropriate regularity conditions

$$\sqrt{N}(T_N(\mathbf{x}) - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0)) \quad (3)$$

where  $I(\theta_0)$  is Fisher information.

We say that an arbitrary estimator  $T_N^\epsilon(\mathbf{x})$  is asymptotically efficient if it is also true that

$$\sqrt{N}(T_N^\epsilon(\mathbf{x}) - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0)). \quad (4)$$

**Hypothesis Testing.** One of the basic inferential problems in statistics is hypothesis testing [18]. Many research questions in the sciences are typically formulated in ways to test if we have sufficient evidence to reject some default theory or a *null hypothesis*; e.g., the new cancer treatment A and the state-of-the-art treatment B are equivalent against the *alternative hypothesis* that their performance is different. It is likely that many statistical tests are available for the same hypothesis testing problem. To evaluate the performance of these different tests, statisticians use the following criteria:

- 1) The confidence level of a hypothesis test is defined as  $1 - \alpha$  where  $\alpha$ , the type I error, is the probability of rejecting the null hypothesis when it is true.
- 2) The power of a hypothesis test calculated as  $1 - \beta$  is the probability of rejecting the null hypothesis when it is false where  $\beta$ , the type II error, is the probability of not rejecting the null hypothesis when it is false.

**Pearson's  $\chi^2$  test of independence.** In clinical trials, researchers are often interested in assessing disease-drug associations [5] where data can be presented in  $I \times J$  contingency

tables [19]. The independence can be tested with Pearson's chi-square test of independence via the  $\chi^2$  test statistic:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i},$$

where  $O_i$  is the observed frequency in a cell  $i$ , and  $E_i$  the expected frequency under the null model of independence.

**Distance Measures.** To analyze the simulation results in Section V, we use the standardized mean squared error (MSE). The mean square error of an estimator  $T(x)$  is defined as

$$MSE_{T(x)}(\theta) = E_\theta [(T(x) - \theta)^2]. \quad (5)$$

Note that if  $T(x)$  is unbiased,  $MSE_{T(x)}(\theta) = Var[T(x)]$ . For graphing purposes we standardize  $MSE_{T(x)}(\theta)$  by the variance  $Var[T_{ML}(x)]$  of the corresponding maximum likelihood estimator  $T_{ML}(x)$ . If  $T_{ML}(x)$  is asymptotically unbiased, its standardized MSE will be close to 1. We expect that the standardized MSEs of the  $\epsilon$ -differential privacy (DP) estimators are also close to 1. This indicates that the DP estimators achieve the same asymptotic efficiency as the corresponding maximum likelihood ones.

## III. DIFFERENTIAL PRIVACY FRAMEWORK

In parametric inference, sufficient statistics or estimates are needed for hypothesis testing or for building valid statistical models. The differential privacy mechanism basically manipulates these statistics by adding Laplace noise to them, thereby creating *perturbed* sufficient statistics or estimates. The precise noise added is described in *Algorithm 1*.

In the framework proposed by [11], the private and efficient estimators are constructed using the idea of "sample and aggregated" [20] method, whereby the data set is divided into  $k$  groups. In contrast to the original algorithm, in this paper we do not divide data into  $k$  groups in order to avoid biases coming from the estimates of small subsets of the data. We also modify the Laplace noise formula given in [11] by removing one of the parameters such that given a fixed sample size  $N$ , we only need to adjust the differential privacy level  $\epsilon$  to allow for trade-off between the privacy and the asymptotic properties of the returned estimators. Our general modifications of the algorithm from [11] are given in *Algorithm 1*.

**Lemma III.1.** *Algorithm 1 satisfies  $\epsilon$ -differential privacy.*

**Lemma III.2.** *Under the regularity conditions of normal asymptotic distributions of ML estimators [17], if the diameter of the parameter space  $\Lambda$  is bounded and  $\epsilon$  is fixed, the estimators  $T_i^\epsilon(x)$  is asymptotically unbiased, normal, and efficient, that is  $\sqrt{n}T_i^\epsilon(x) \xrightarrow{D} N(\theta_i, I_f^{-1}(\theta_i))$ .*

The proofs for  $\epsilon$ -differential privacy and the asymptotic properties here follow the proofs in [11] and will be presented in our extended version of this workshop paper; see [21].

---

**Algorithm 1** An  $\epsilon$ -Differential Privacy Algorithm

---

**Input:** A data set  $\mathbf{x} = (x_1, \dots, x_N) \in D^N, \epsilon > 0$ .

**Parameters:**  $\Lambda$  is the range of  $T_i(x)$  or diameter of the parameter space, and  $\epsilon$  is the level of privacy to be achieved.

**Output:** A set of estimators for the parameters  $\theta_1, \dots, \theta_m$ ; or sufficient statistics.

- 1) Obtain the sufficient statistics  $T_1(x), \dots, T_m(x)$  for  $\theta_1, \dots, \theta_m$ . A bias correction can be applied here.
  - 2) For each estimator  $T_i(x)$  draw a random observation  $L_i$  from a Laplace distribution with mean 0 and standard deviation  $\sqrt{2}\Lambda/(N\epsilon)$ .
  - 3)  $T_i^\epsilon(x) = T_i(x) + L_i$ .
  - 4) Return  $T_i^\epsilon(x)$  or MLE computed using  $T_i^\epsilon(x)$ .
- 

#### IV. SAMPLE SIZE DETERMINATION UNDER DIFFERENTIAL PRIVACY FRAMEWORK

In this section, we present our preliminary investigation into sample size estimation under the above described differential privacy framework. We denote with  $N$  the original or *true* sample size calculated without accounting for privacy, and with  $N'$  the differentially private or *realistic* sample size. All formulae correspond to a specified null hypothesis ( $H_0$ ) and one test statistic, and depend on the chosen confidence level  $1 - \alpha$  and the power  $1 - \beta$ . For practical implications, we consider two situations.

First, if the data are already available as in data mining applications, these procedures will provide the lower bounds of the *realistic* sample size needed to achieve the same statistical power as the *true*  $N$  would. Simulations in Section V show that when the true data size is large enough the difference between  $N$  and  $N'$  is not significant.

Second, in clinical and social sciences data collection and conducting experiments are costly, and the researchers are required to a priori determine the finite sample size needed to detect important differences. As discussed in previous sections, the proposed privacy estimators theoretically achieve asymptotic efficiency when the sample sizes go to infinity. Therefore, we need to establish a trade-off between the statistical efficiency and the privacy by increasing the required sample sizes to control for the effect of Laplace noise.

##### A. Classical Hypothesis Testing with Single Proportion

The U.S. FDA requires that efficacy be proven prior to giving its approval of a drug. Efficacy means that the tested dose of the drug is effective in treating the condition. Confidence intervals are often used to reflect the level of precision, but their width is a function of sample size. The researchers seek a certain precision, e.g.,  $\alpha = 0.05$ , and need to decide on the sample size of their study a priori. The simplest example occurs when the outcome is binary, e.g., the success or failure of a drug dosage in treating a heart condition.

Let  $x_1, x_2, \dots, x_N \sim \text{Bernoulli}(p)$ , and the sufficient statistic  $\hat{p} = \frac{1}{N} \sum_{i=1}^N x_i$  be the estimator of interest. We are interested

in testing the following hypothesis [18]:

$$H_0 : p = p_0 \text{ versus } H_a : p = p_0 + \delta, \quad (6)$$

where  $\hat{p}$ , for example, is the sample proportion of people who had a successful drug treatment. After adding Laplace noise  $L(\frac{\sqrt{2}}{\epsilon N})$  the sampling distribution of  $\hat{p}$  under the null hypothesis can be approximated by  $N(p_0, \frac{\sigma^2}{N}) + L(\frac{\sqrt{2}}{\epsilon N})$  and under the alternative hypothesis by  $N(p_0 + \delta, \frac{\sigma^2}{N}) + L(\frac{\sqrt{2}}{\epsilon N})$ . Here we define  $\sigma^2 = \bar{p}(1 - \bar{p})$  where  $\bar{p} = p_0 + \frac{\delta}{2}$  as in [18]. If we approximate  $L(\frac{\sqrt{2}}{\epsilon N})$  by a normal distribution  $N(0, \frac{2}{\epsilon^2 N^2})$ , the sampling distributions under privacy framework become  $N(p_0, \frac{\sigma^2}{N} + \frac{2}{\epsilon^2 N^2})$  and  $N(p_0 + \delta, \frac{\sigma^2}{N} + \frac{2}{\epsilon^2 N^2})$ .

To achieve the confidence level  $1 - \alpha$  and the power  $1 - \beta$ , the sample size  $N$  is calculated by solving:

$$p_0 + z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{N} + \frac{2}{\epsilon^2 N^2}} = p_0 + \delta - z_{1-\beta} \sqrt{\frac{\sigma^2}{N} + \frac{2}{\epsilon^2 N^2}} \quad (7)$$

The solution gives the following expression for the privacy-preserving sample size  $N'$

$$N' = N \left( \frac{1}{2} + \frac{1}{2} \sqrt{1 + \frac{8\delta^2}{\epsilon^2 (z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^4}} \right), \quad (8)$$

where  $N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{\delta^2}$  is the required classical sample size when no Laplace noise is added to the estimator  $\hat{p}$ . Here we are interested in the *approximate sample size correction factor* under the differential privacy framework:

$$K = \frac{1}{2} + \frac{1}{2} \sqrt{1 + \frac{8\delta^2}{\epsilon^2 (z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^4}}. \quad (9)$$

However, the above factor may not be safe because the Laplace distribution has a fatter tail than the normal distribution.

To compute the *exact sample size correction factor* under the differential privacy framework, we need to use a numerical method. Based on [22], the perturbed sampling distributions of  $\hat{p}$  under the null and alternative hypotheses are Normal Laplace distributions:  $X_1 \sim NL(p_0, \frac{\sigma^2}{N'}, \epsilon N', \epsilon N', 1)$ , and  $X_2 \sim NL(p_0 + \delta, \frac{\sigma^2}{N'}, \epsilon N', \epsilon N', 1)$ . Using a unique root-finding method for the equation  $F_{X_1}^{-1}(1 - \alpha/2) = F_{X_2}^{-1}(1 - \beta)$  with respect to the variable  $N'$ , we can find the exact *realistic* sample size  $N'$  and the exact sample size correction factor  $K$  by dividing the realistic  $N'$  by the *true* sample size  $N$ .

##### B. Pearson's $\chi^2$ Test of Independence

A comparative treatment efficacy (CTE) clinical trial consists of comparing two treatment groups with respect to the outcome, e.g., a disease-drug association study. In this paper, we consider the categorical outcomes and sample size estimation for a  $\chi^2$  test of independence in  $2 \times 2$  tables, where cell counts follow a *Multi*( $N, p$ ) distribution.

The proposed algorithm assumes that based on previous studies we can obtain prior estimates of cell proportions in the contingency tables  $\pi_i^0$ s. Given these prior estimates, the true sample size  $N$ , and a required level of differential privacy  $\epsilon$ ,

we can run a simulation to draw a graph of  $MSE(\hat{p}_i^\epsilon)$ s versus  $\epsilon$  as illustrated in the Experiments section on a  $\chi^2$  test of independence (Figure 3). In the simulation, we draw  $M$  samples of  $N$  data points from a *Multinomial* with parameters  $\pi_i^0$ s, add Laplace noise  $L(\frac{\sqrt{2}\Lambda}{N\epsilon})$  to the observed proportions  $\hat{p}_{i,t}$  in each sample  $t$  which results in  $\epsilon$ -differential proportion estimates  $\hat{p}_{i,t}^\epsilon$ , and then approximate  $MSE(\hat{p}_i^\epsilon)$ s by:

$$MSE(\hat{p}_i^\epsilon) = \frac{1}{M} \sum_{t=1}^M (\hat{p}_{i,t}^\epsilon - \pi_i^0)^2. \quad (10)$$

Note that we can increase  $M$  arbitrarily for better approximation of  $MSE(\hat{p}_i^\epsilon)$ s. From this plot, we select the worst  $MSE(\hat{p}_k^\epsilon)$  and since it is standardized, the y-coordinate value of this  $MSE(\hat{p}_k^\epsilon)$  curve gives the ratio:

$$c \approx \frac{MSE_N(\hat{p}_k^\epsilon)}{MSE_N(\hat{p}_k)} = \frac{MSE_N(\hat{p}_k^\epsilon)}{Var_N(\hat{p}_k)}, \quad (11)$$

where subscript  $N$  denotes the true sample size without injection of privacy-preservation noise. To achieve the same statistical power in the “private” analysis, we need to increase the original sample size  $N$  to the statistically equivalent differentially private sample size  $N'$ . The sample size correction factor can be approximated by solving:

$$\begin{aligned} MSE_N(\hat{p}_k) &= MSE_{N'}(\hat{p}_k^\epsilon) \\ MSE_N(\hat{p}_k) &= MSE_{N'}(\hat{p}_k^\epsilon) + MSE_{N'}(Lap) \\ MSE_N(\hat{p}_k) &= MSE_{N'}(\hat{p}_k^\epsilon) \sqrt{\frac{N}{N'}} + MSE_N(Lap) \frac{N}{N'} \end{aligned}$$

From (11), we have

$$MSE_N(\hat{p}_k^\epsilon) = MSE_N(\hat{p}_k) + MSE_N(Lap) \approx c MSE_N(\hat{p}_k) \quad (12)$$

then  $MSE_N(Lap) \approx (c - 1)MSE_N(\hat{p}_k)$ . Replacing  $MSE_N(Lap)$  and cancelling out  $MSE_N(\hat{p}_k)$  on both sides, we obtain the following equation:

$$\sqrt{\frac{N}{N'}} + (c - 1) \frac{N}{N'} = 1. \quad (13)$$

Set  $N' = KN$ , and solve the above equation to obtain the sample size correction factor  $K$ :

$$K = (\sqrt{1 + 4(c - 1)} + 1)^2 / 4 \quad (14)$$

## V. EXPERIMENTS

For each model described in Section IV, the simulations include two types of analysis. We demonstrate the interactive effects of the sample size and the privacy level  $\epsilon$  and compare the asymptotic efficiency of the maximum likelihood estimator and the differentially private one. The role of  $\epsilon$  is similar to that of confidence level  $1 - \alpha$  or the power  $1 - \beta$  in statistical analysis. Given the required level of privacy which is realized by the value of  $\epsilon$ , by controlling the sample size  $N$  we can achieve the asymptotic efficiency. We also show how the sample size correction factor changes with changing levels of privacy.

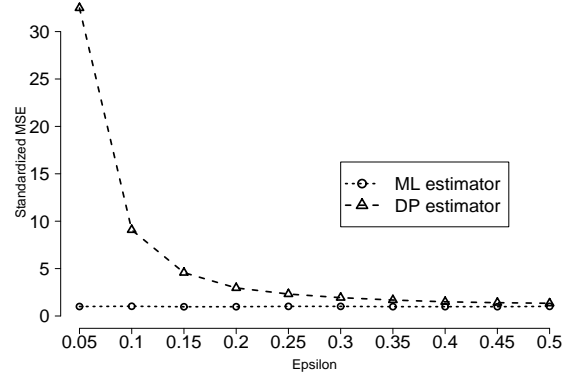


Fig. 1. ML versus DP estimator in a binomial model with  $p = .5$ , controlling for  $\epsilon$  with  $N = 100$  and simulation size  $M = 10000$ .

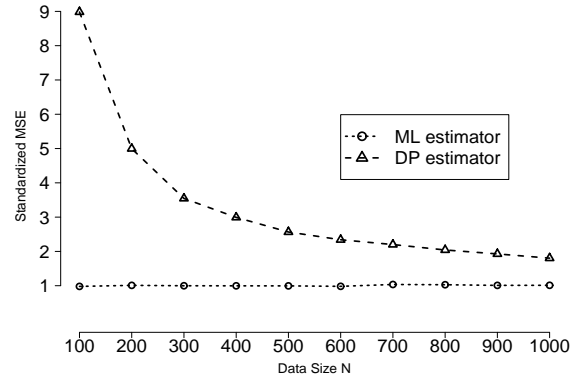


Fig. 2. ML versus DP estimator in a binomial model with  $p = .5$ , controlling for  $N$ , at privacy level  $\epsilon = .1$  and  $M = 10000$ .

### A. Classical Hypothesis Testing with Single Proportion

Figure 1 illustrates the trade-off between differential privacy and asymptotic efficiency for the binomial parameter  $p = 0.5$ , the sample size  $N = 100$  and the simulation size  $M = 10000$ . Recall that an unbiased ML or efficient DP estimator should have a standardized MSE close to 1. As expected, when we increase the value of the privacy level  $\epsilon$  allowing for more information leakage, we achieve better asymptotic efficiency.

Figure 2 illustrates the role of increasing sample size  $N$  in achieving asymptotic efficiency when the privacy parameter is fixed at  $\epsilon = 0.1$ . For larger  $N$ , there is a faster convergence of ML and DP estimators. The interesting question here is how to control for the sample size  $N$  to achieve both the desired asymptotic efficiency and privacy guarantees. As described in Section IV, in standard statistical practice, the required asymptotic efficiency could be realized through the specified confidence levels and statistical power. Under a differentially private framework, as proposed in Section IV, we need to make an adjustment to the sample size  $N$  using the correction factor  $K$ .

Tables I and II compare the proposed exact and approximate sample size factors,  $K$ , under the binomial hypothesis testing with two different power levels  $1 - \beta = .6$  and  $.9$ . Since

TABLE I

THE EXACT AND NORMAL-APPROXIMATED SAMPLE SIZE CORRECTION FACTOR  $K$  WITH  $\alpha = .05$ ,  $\beta = .4$ ,  $p_0 = .25$ ,  $\delta = .1$ . TRUE SAMPLE SIZE IS  $N = 103$  AND THE DP SAMPLE SIZE  $N' = KN$ .

$\epsilon$	0.1	0.2	0.3	0.4	0.5
Exact K	3.65	2.12	1.64	1.42	1.29
Normal-Approximated K	3.58	2.10	1.63	1.41	1.29

TABLE II

THE EXACT AND NORMAL-APPROXIMATED SAMPLE SIZE CORRECTION FACTOR  $K$  WHEN  $\alpha = .05$ ,  $\beta = .1$ ,  $p_0 = .25$ ,  $\delta = .1$ . TRUE SAMPLE SIZE  $N = 221$  AND THE DP SAMPLE SIZE  $N' = KN$ .

$\epsilon$	0.1	0.2	0.3	0.4	0.5
Exact K	2.62	1.64	1.35	1.22	1.15
Normal-Approximate K	2.64	1.65	1.35	1.22	1.15

the Laplace distribution has fatter tails than the corresponding normal distribution, the normal-approximated formula (9) underestimates the true  $K$  in comparison to the exact value. However, from the tables we can see that the differences are small. The other observation is that for the fixed  $\epsilon$  and for the higher required statistical power, the sample size correction factor  $K$  decreases. This phenomenon can be explained by the two following reasons. First, the classical sample sizes  $N$  increases when the required power increases. Second, when the sample size increases, the asymptotic efficiency of the  $\epsilon$ -differential estimators and their corresponding ML estimators are closer.

### B. Pearson's $\chi^2$ test of independence

For this paper we only consider simulations for  $2 \times 2$  contingency tables where the Laplace noise is added to the cell proportions under the saturated model, that is to the observed cell proportions. In the first set of simulations the samples are drawn from a ground truth  $Mult(.333, .167, .25, .25)$ . Similar to the binomial case, we compare the usual ML estimators of the cell probabilities under the saturated model with their  $\epsilon$ -differential private (DP) estimators. Figure 3 demonstrates the trade-off between the privacy and the asymptotic efficiency of the estimators for a fixed-size data set  $N = 100$  by varying the privacy level  $\epsilon$ . Figure 4 illustrates the role of the data size  $N$  with fixed  $\epsilon = 0.1$ . As expected, we observe that in order to have strong privacy we lose in utility (asymptotic efficiency), and the small cell probabilities have significantly larger  $MSEs$  than the larger ones; for example, see  $\epsilon = 0.1$ ,  $p = 0.167$  v.s.  $p = 0.333$ .

How big does the sample size need to be in order for the ML estimators to simultaneously achieve a certain level of privacy and the asymptotic efficiency under the saturated model? We assume that we know the prior estimates of the cell probabilities (e.g., from pilot or earlier studies) in order to estimate the value  $c$  in equation (11). Figure 5 demonstrates the range of the sample size correction factor  $K$  derived in (14) with the prior estimates of the cell probabilities as  $(.333, .167, .25, .25)$  when we vary the privacy level  $\epsilon$  and the original sample size  $N$ . For example, for  $\epsilon = 0.2$ , to achieve

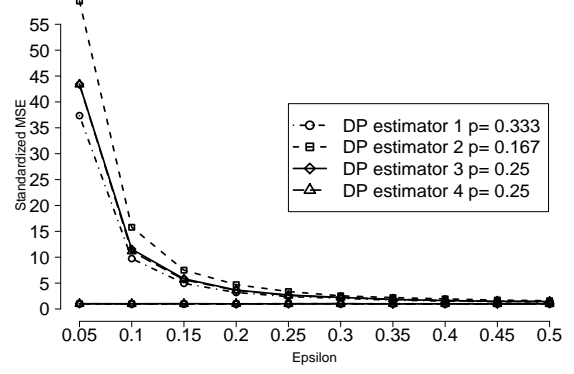


Fig. 3. ML versus DP estimators for  $Mult(.333, .167, .25, .25)$ , controlling for privacy level  $\epsilon$ , with  $N = 100$  and  $M = 10000$ . Overlapping horizontal lines at the bottom correspond to ML estimators.

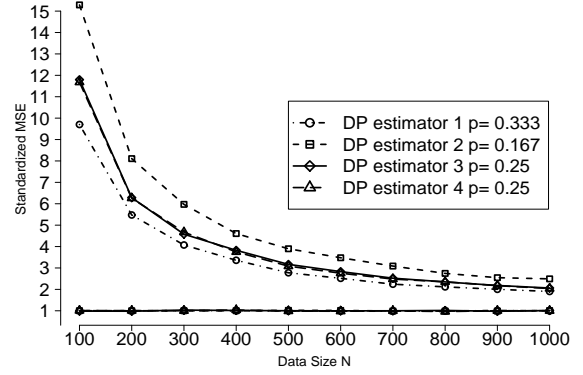


Fig. 4. ML estimators v.s. DP estimators in a  $Mult(.333, .167, .25, .25)$ , controlling for the sample size  $N$ . The privacy level  $\epsilon = .1$  and the simulation size  $M = 10000$ . The overlapped horizontal lines correspond to the ML estimators.

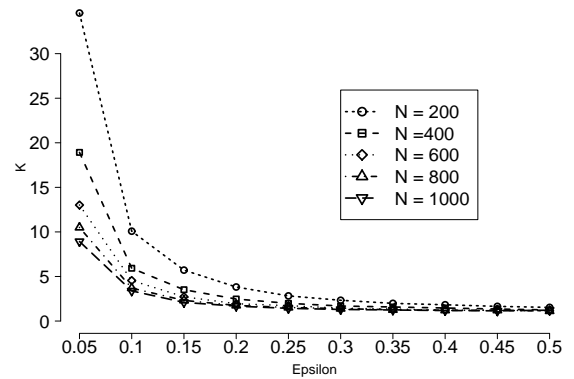


Fig. 5. The sample size correction factor  $K$  for a  $2 \times 2$  table, controlling for the sample size  $N$  and the privacy level  $\epsilon$ . Prior estimates of cell probabilities  $(.333, .167, .25, .25)$ .

the same efficiency with  $N = 200$  and no noise, the researcher must collect a sample that is about four times as large as the original.

For the next set of simulations assume again that we have

TABLE III

THE STATISTICAL POWER OF INDEPENDENCE  $\chi^2$  TEST WITH  $\alpha = .05$ , THE CLASSICAL SAMPLE SIZE  $N = 200$ , THE DP SAMPLE SIZE  $N' = KN$ , THE SIMULATION SIZE  $M = 10000$ . PRIOR ESTIMATES OF CELL PROBABILITIES ARE (.333, .167, .25, .25) WHEN THE GROUND-TRUTH ONES ARE (2/9, 5/18, 13/36, 5/36).

$\epsilon$	0.1	0.2	0.3	0.4	0.5
No Noise	0.9835	0.9811	0.9835	0.9824	0.9812
DP	1.0000	0.9999	0.9996	0.9984	0.9958

TABLE IV

THE STATISTICAL POWER OF INDEPENDENCE  $\chi^2$  TEST WITH  $\alpha = .05$ , THE CLASSICAL SAMPLE SIZE  $N = 200$ , THE DP SAMPLE SIZE  $N' = KN$ , THE SIMULATION SIZE  $M = 10000$ . PRIOR ESTIMATES OF CELL PROBABILITIES ARE (.333, .167, .25, .25) WHEN THE GROUND-TRUTH ONES ARE (.25, .25, .167, .333).

$\epsilon$	0.1	0.2	0.3	0.4	0.5
No Noise	0.6727	0.6723	0.6746	0.6776	0.6675
DP	0.9999	0.9865	0.9263	0.8743	0.8242

the prior cell estimates as (.333, .167, .25, .25) in order to get the  $c$  value. Then, Tables III and IV compare statistical power  $1 - \beta$  for the  $\chi^2$  test of independence, with confidence level  $1 - \alpha = .95$ , the original sample size  $N = 200$ , and the corrected sample size  $N' = KN$  with Laplace noise.

For Table III, the noise is added to the samples based on the ground truth values from  $Mult(2/9, 5/18, 13/36, 5/36)$  which are close to the prior estimates. However, for Table IV, the samples are based on the ground truth values drawn from  $Mult(.25, .25, .167, .333)$  which are further from our prior estimates. Then the  $\chi^2$  statistic is calculated based on the DP estimates. The simulated corrected sample sizes  $K$  based on the equation (14) work well under both situations, with good and bad priors. The  $\epsilon$ -differential estimates achieve the desired privacy and have better power because the corrected sample sizes are larger. However, we also observe that  $K$  is overestimated, that is conservative, when the prior is not good. This happens because in the approximation we do not take into account for the relations between the cell probabilities.

## VI. CONCLUSION

In this paper, we demonstrated how the revised  $\epsilon$ -differential private framework can be applied to clinical data mining via two representative examples: tests for binomial proportions and contingency tables. It is possible to integrate the  $\epsilon$ -differential private framework with the classical statistical hypothesis testing and modeling. The sample size plays a role in balancing the statistical efficiency (type I and type II errors,  $\alpha$  and  $\beta$ ) and the privacy level  $\epsilon$ . We derived the rules for the sample size adjustment whereby both statistical efficiency and differential privacy can be achieved. These rules can be used for planning new clinical experiments and for measuring the loss of statistical efficiency (the reduced true sample size) when the proposed privacy framework is deployed for mining of current clinical data.

In clinical and social sciences, log-linear models and logistic regression models are popular. We are currently working on applying the differential privacy framework to these models,

and deriving similar sample sizes adjustment methods. Some preliminary results of our work on Bayesian hypothesis testing are also discussed in the extended paper [21].

## ACKNOWLEDGMENT

The authors would like to thank Adam Smith and Ivan Simeonov for invaluable discussion. The research reported here was supported in part by ONR-MURI program, Award Number N00014-08-1-1015, and by NSF Grant SES-0532407 to the Department of Statistics, Pennsylvania State University.

## REFERENCES

- [1] S. Piantadosi, *Clinical Trials: A Methodologic Perspective Second Edition* (Wiley Series in Probability and Statistics). Wiley-Interscience, August 2005.
- [2] T. Tse, R. Williams, and D. Zarin, "Reporting Basic Results in ClinicalTrials.gov," *Chest*, vol. 136, no. 1, p. 295, 2009.
- [3] Food and D. Administration, "Amendments act of 2007. public law no. 110-85." [Online]. Available: <http://prsinfo.clinicaltrials.gov/fdaa.html>
- [4] X. Cao, K. Maloney, and V. Brusica, "Data mining of cancer vaccine trials: a bird's-eye view," *Immunome Research*, vol. 4, no. 1, p. 7, 2008.
- [5] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman, "Automated acquisition of disease-drug knowledge from biomedical and clinical documents: An initial study," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 87 – 98, 2008.
- [6] C. Bethel, L. Hall, and D. Goldgof, "Mining for Implications in Medical Data," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, 2006.
- [7] M. Kantarcoglu, J. Jin, and C. Clifton, "When do data mining results violate privacy?" in *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM Press, 2004, pp. 599–604.
- [8] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *SIGMOD Record*, vol. 33, p. 2004, 2004.
- [9] F. C. on Statistical Methodology, "Statistical policy working group 22 - report on statistical disclosure limitation methodology," 2005.
- [10] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *In Proceedings of the 3rd Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.
- [11] A. Smith, "Efficient, differentially private point estimators," *Preprint arXiv:0809.4794v1*, 2008.
- [12] C. Dwork and J. Lei, "Differential privacy and robust statistics," in *STOC*, 2009, pp. 371–380.
- [13] L. Wasserman and S. Zhou, "A statistical framework for differential privacy," *Preprint arXiv.org:0811.2501*, 2008.
- [14] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy, accuracy, and consistency too: a holistic solution to contingency table release," in *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. ACM New York, NY, USA, 2007, pp. 273–282.
- [15] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," *Lecture Notes in Computer Science*, vol. 3876, p. 265, 2006.
- [16] C. Dwork and A. Smith, "Differential privacy for statistics: What we know and what we want to learn," 2008.
- [17] T. S. Ferguson, *A Course in Large Sample Theory*. Chapman & Hall, 1996.
- [18] G. van Belle, *Statistical Rules of Thumb*, 2nd ed. John Wiley & Sons, 2008.
- [19] Y. Bishop, S. Fienberg, and P. Holland, *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge MA, 1975.
- [20] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 2007, pp. 75–84.
- [21] D. Vu, A. Slavković, and A. Smith, "Differential privacy and statistical hypothesis testing," *Manuscript in preparation available at <http://www.stat.psu.edu/~sesa/privacy.html>*, 2009.
- [22] F. Wu, "Applications of the normal laplace and generalized normal laplace distributions," Master's thesis, University of Victoria, 2008.