Protecting count queries in study design

Staal A Vinterbo,[1] Anand D Sarwate,[2] Aziz A Boxwala[1]

[1]Division of Biomedical Informatics, Department of Medicine, University of California San Diego, La Jolla, California, USA
[2]Toyota Technological Institute at Chicago, Chicago, Illinois, USA

Correspondence to
Dr Staal A Vinterbo, Division of Biomedical Informatics, Department of Medicine, University of California San Diego, 9500 Gilman Drive #0728, La Jolla, CA 92093-0728, USA; sav@ucsd.edu

Received 5 July 2011
Accepted 15 March 2012
Published Online First 17 April 2012

ABSTRACT

Objective Today's clinical research institutions provide tools for researchers to query their data warehouses for counts of patients. To protect patient privacy, counts are perturbed before reporting; this compromises their utility for increased privacy. The goal of this study is to extend current query answer systems to guarantee a quantifiable level of privacy and allow users to tailor perturbations to maximize the usefulness according to their needs.

Methods A perturbation mechanism was designed in which users are given options with respect to scale and direction of the perturbation. The mechanism translates the true count, user preferences, and a privacy level within administrator-specified bounds into a probability distribution from which the perturbed count is drawn.

Results Users can significantly impact the scale and direction of the count perturbation and can receive more accurate final cohort estimates. Strong and semantically meaningful differential privacy is guaranteed, providing for a unified privacy accounting system that can support role-based trust levels. This study provides an open source web-enabled tool to investigate visually and numerically the interaction between system parameters, including required privacy level and user preference settings.

Conclusions Quantifying privacy allows system administrators to provide users with a privacy budget and to monitor its expenditure, enabling users to control the inevitable loss of utility. While current measures of privacy are conservative, this system can take advantage of future advances in privacy measurement. The system provides new ways of trading off privacy and utility that are not provided in current study design systems.

BACKGROUND

In 1991 the Institutes of Medicine stated that "perhaps the impediment to computer-based patient records (CPRs) that is of greatest concern is the issue of privacy".[1] Since then, privacy in healthcare practice and research has received growing attention from the public, legislators, and researchers. In December 2000 the Federal Register issued the Portability and Accountability Act (HIPAA) Privacy Rule (45 CFR Part 160 and Subparts A and E of Part 164), which contains criteria for deciding whether data are subject to dissemination restrictions. These criteria are based on whether the data are 'de-identified'—that is, whether a particular record or similar piece of information can be linked back to the identity of the person from whom it stems. Associating privacy with preventing data-to-identity linkage is reflected in both historic and recent definitions of privacy[2–5] as well as empirical risk quantification.[6 7] Indeed, Winkler states that "the highest standard for estimating (re-identification risk) is where record linkage is used to match information from public data with the masked microdata".[8] However, many approaches to de-identifying or 'sanitizing' data sets have been shown to be subject to attacks[9–11] which use public data to compromise privacy. In 2003 Dinur and Nissim[12] showed that a data curator can only provide useful answers to a limited number of questions about the data without divulging most of the sensitive information contained in it. In 2006 Dwork *et al* provided a definition of privacy ('differential privacy')[13] that quantifies risk to individuals of unwanted disclosures. This quantification allows systems to account for privacy loss over time and to track a 'privacy budget' to manage this loss.

Whereas the HIPAA de-identification standard and other anonymity-based definitions of privacy such as *k*-anonymity[2] are based on the properties of the data to be disseminated, differential privacy is based on the properties of the process used to disseminate the data. Process-based privacy agrees with commonsense notions of privacy. Suppose a data holder has two data sets which have been individually deemed safe to disseminate and they decide to release the first or the second based on the HIV status of a particular individual. While the data themselves are privacy-preserving, the dissemination process is not—the choice of disseminated data set reveals the HIV status of the individual. Current approaches that treat privacy as properties of the data are not able to address this type of privacy threat. Furthermore, they impose losses of utility that are deemed major barriers to the secondary use of clinical data in research.[14–17]

Cohort identification is a common task in the secondary use of clinical data. Researchers can issue count queries about how many patients exist who match a particular profile (eg, patients who are male, have secondary diabetes, and are the age of 30). Allowing unrestricted access to count queries endangers patient privacy because even a few queries can reveal sensitive information. For example, suppose the above query returns three. If this query is extended with an additional clause 'and HIV-positive' and still returns three, we can infer that any 30-year-old man with secondary diabetes in the database is HIV-positive. This information can easily be linked with auxiliary knowledge of a diabetes patient (eg, knowing that a colleague or neighbor seeks care for diabetes at your institution) to infer that they are HIV-positive, which constitutes a serious privacy breach.

To protect patient privacy some institutions only allow researchers to receive approximate counts and access is mediated by systems such as the i2b2[18] framework and the Stanford University Medical

School STRIDE[19] environment. In particular, i2b2 and STRIDE add Gaussian noise to the true count and then round to the nearest multiple of one and five, respectively. Importantly, these systems provide privacy through the process by which they answer queries. However, when they were developed there were no metrics to quantify protection from re-identification, so it is not surprising that there are no quantitative analyses of how the scheme affects privacy loss over time. Instead, they estimate how many times a single query has to be repeated in order to estimate the true count by averaging out the perturbation.[20]

Generally speaking, fixed magnitude perturbations affect small counts much more than large counts. Unfortunately, we cannot reduce the perturbation for small counts and provide the same privacy. However, there are situations where the direction of the perturbation matters for the user, and if we allow flexibility in this regard, we can mitigate the problem to some degree without compromising privacy. Consider the following two fictitious use cases.

Researcher A wants to conduct a clinical trial for a new treatment for cancers affecting the salivary glands. The trial must accrue at least 40 patients for the study to have sufficient power, and researcher A needs to determine the length of the trial to develop a budget. Furthermore, there would be a high cost to not enrolling a sufficient number of patients. If the actual count in the database is 38 patients diagnosed with primary carcinoma of the salivary glands per year and the query tool returns a perturbed count of 45, the researcher may budget for a too short trial run, resulting in an underpowered study at high cost. Researcher B designs a study that requires her to enrol consecutive patients admitted with a diagnosis of heart failure for the first 6 months of the year. All these patients would be offered physiological home monitoring on discharge. This kit would monitor various cardiovascular parameters and electronically transmit them to a server. The protocol budgets for 75 patients based on the use of the query tool. If the real number is 85, the proposed budget would be too small, resulting in a missed opportunity.

OBJECTIVE

Our objective is to demonstrate an extension to currently employed count query systems for study design that: (1) provides current functionality with stronger privacy guarantees, serving as an example of a service that can be included in a larger enterprise-wide system that manages privacy and accountability; (2) provides the option to incorporate user preferences with regard to individual query responses, thereby increasing utility to users without compromising privacy; (3) supports privacy budgeting to increase utility to users across multiple queries; and (4) is implementable and practical in use.

The main tools we employ to meet this objective are the recent statistically motivated privacy metric differential privacy[13] and the exponential mechanism of McSherry and Talwar.[21]

METHODS

A system for count perturbation

The following description is intended to give an overview of what our system does and to serve as a recipe for implementation. We present the mathematical properties in subsequent sections, as well as some potential enhancements to the system later in the Discussion.

In describing our system we will denote by n the total number of individual records in the database. The administrator sets parameters $r_{\min}$ and $r_{\max}$, which are upper and lower bounds on the possible answers to be returned by the query mechanism, and assigns each user a total privacy budget $\varepsilon_{\text{total}}$ that represents the total privacy risk they are allowed to incur prior to their access being reviewed. The administrator also assigns a per-query privacy level ε for the user.

The perturbation is parameterized by positive numbers α^+, β^+, α^-, and β^- which define the following function $U_c(r)$, giving the utility placed on receiving a result r if the real count is c:

$$U_c(r) = \begin{cases} -\beta^+(r-c)^{\alpha^+} & \text{if } r \geq c \\ -\beta^-(c-r)^{\alpha^-} & \text{otherwise} \end{cases}$$

Because α and β parameters are positive, the utility $U_c(r)$ is maximal at $r=c$. The utility is specified independently for $r \geq c$ and $r<c$ to reflect asymmetric preferences with respect to over- or underestimation.

The parameters α^+, α^-, β^+, and β^- are chosen using a fictitious value $\hat{c}$ for c. In practice, the parameter choice will be aided by a tool like in figure 4. The right side plot shows a utility that is linearly decreasing with absolute distance from the real count c, with a 3× steeper decrease on the right side of c, representing a bias towards underestimation. Other shapes can be seen in figure 1. The system implementor can offer users a variety of preset parameter settings from which they can choose.

When the user chooses parameters and issues a query to the database, the system first computes the true count c. User parameters and c are used to compute a parameter η for the per-query privacy level ε by:

$$\begin{aligned} \Delta^+ &= \max\left(\beta^+, \alpha^+\beta^+ r_{\max}^{\alpha^+-1}\right) \\ \Delta^- &= \max\left(\beta^-, \alpha^-\beta^-\left(n - r_{\min}\right)^{\alpha^- - 1}\right) \\ \Delta &= \max\left(\Delta^-, \Delta^+\right) \\ \eta &= \frac{\varepsilon}{2\Delta}. \end{aligned}$$

Using the computed η, the system then translates U_c into a probability distribution $P(r|c)$ by:

$$P(r|c) = \frac{\exp(\eta U_c(r))}{N} \text{where}$$

$$N = \sum_{r'=r_{\min}}^{r_{\max}} \exp(\eta U_c(r')). \quad (1)$$

Finally, the system response r is randomly chosen among integer values lying between $r_{\min}$ and $r_{\max}$ with a probability given by $P(r|c)$. The mean and variance of $P(r|c)$ given the system and user parameters are given by:

$$\mu = \sum_{r=r_{\min}}^{r_{\max}} r \cdot \frac{\exp(\eta U_c(r))}{\sum_{r'} \exp(\eta U_c(r'))}. \quad (2)$$

$$\sigma^2 = \sum_{r=r_{\min}}^{r_{\max}} (r-\mu)^2 \cdot \frac{\exp(\eta U_c(r))}{\sum_{r'} \exp(\eta U_c(r'))}. \quad (3)$$

Quantifying privacy

We now describe the mathematics behind our new system and how it compares with existing methods. We model records as

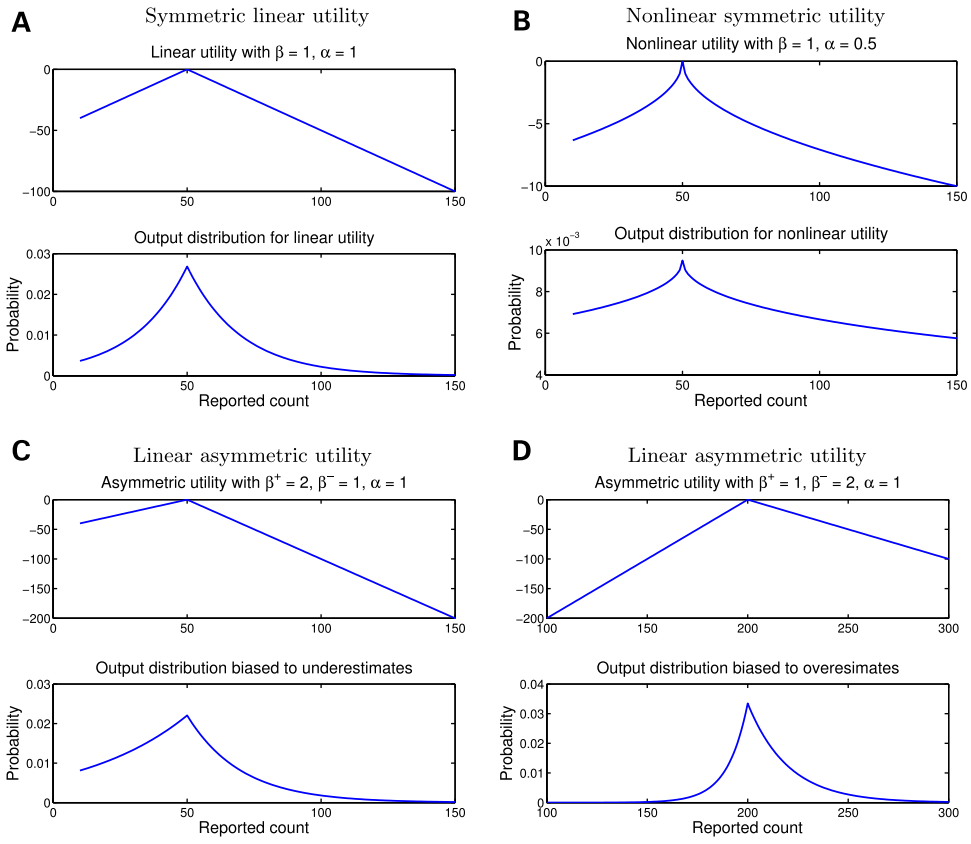

Figure 1 Utility shapes and corresponding noise distributions. (A–C) show the distributions centered at a true count of 50 and (D) at a true count of 200.

vectors in a feature space V and a database D as a collection of n such vectors. A query or predicate p is a function $p : V \rightarrow \{0,1\}$ taking as input a point in V, and producing 0 when the record does not match the predicate and 1 when it does. The number of records in D for which p evaluates to 1 is the (true) count for p:

$$c(p,D) = \sum_{x \in D} p(x).$$

Response mechanisms

A perturbed response mechanism $\mu(c(p,D), n, \theta)$ takes a predicate and database and releases a number chosen randomly according to a distribution that is a function of p, D, and a parameter vector θ. For our proposed system, if the true count is $c(p,D)=c$, then $\mu(c, n, \theta)$ generates the response from the distribution $P(r|c)$ in (1) parameterized by $\theta=(\varepsilon, r_{\min}, r_{\max}, \beta^+, \beta^-, \alpha^+, \alpha^-)$. The approaches used in i2b2[20] and STRIDE can be expressed as:

$$\mu(c,n,\theta) = [c + v]_k.$$

where v is drawn from a Gaussian distribution with mean 0 and SD θ and $[\cdot]_k$ stands for rounding to the nearest multiple of k. In i2b2, $k=1$, and in STRIDE, $k=5$. Both approaches report values produced by μ at or below a given threshold as 'at or below' the given threshold.

Differential privacy

A response mechanism satisfies ε-differential privacy[13] if, for any two databases, D and D' differing in a single point, any query p, and any response r, we have:

$$\frac{P(\mu(c(p,D),n,\theta) = r)}{P(\mu(c(p,D'),n,\theta) = r)} \leq e^{\varepsilon}. \quad (4)$$

That is, the probability that the mechanism returns a count r from running query p on D is very close to the probability it returns a count r from running p on the database D'. This closeness from changing a single individual's data is the source of the name 'differential' and also illustrates the strength of the measure—it guarantees privacy to any (and every) individual in the database regardless of any additional information available.

McSherry and Talwar's exponential mechanism[21] translates the utility $U_c(r)$ of getting r when the true count is c into the probability $P(r|c)$. By specifying a utility function that matches the user's own preferences, their results show that for any non-negative integers c' and c such that $|c-c'| \leq 1$ and for all r between $r_{\min}$ and $r_{\max}$:

$$\frac{P(r|c)}{P(r|c')} \leq e^{2\eta\Delta},$$

where $\Delta = \max_{r_{\min} \leq r \leq r_{\max}} \{|U_c(r) - U'_c(r)|\}$. Conversely, for a fixed ε, the translation must employ $\eta = \varepsilon(2\Delta)^{-1}$ in order to provide ε-differential privacy.

The sensitivity Δ of U_c is computed as follows. We start by noting that U_c is of the form $h=-\beta|r-c|^{\alpha}$ with different values for α and β depending on whether or not $r \geq c$. Consider the $r \geq c$ case where $\beta=\beta^+$ and $\alpha=\alpha^+$ and we denote the resulting h as h^+. We then have Δ^+ as the maximum change in h^+ over all possible values r and c, c' such that $|c-c'| \leq 1$. Formally, we can express this as:

$$\Delta^+ = \max_{r,c}\left(\frac{dh^+}{dc}\right), \text{where}$$

$$\frac{dh^+}{dc} = \alpha^+\beta^+\frac{|r-c|^{\alpha^+}}{r-c}. \quad (5)$$

We note that (5) is not defined for $r=c$ as this is where U_c discontinuously 'switches' from h^- to h^+. When $r=c$, we have $U_c(c)=0$ and consequently $|U_c(c+1)-U_c(c)|=|U_c(c+1)|=\beta^+$ (and $|U_c(c-1)-U_c(c)|=|U_c(c-1)|=\beta^-$). If $\alpha<1$, β^+ is always larger than (5) and, if $\alpha^+=1$, then (5) reduces to β^+. Finally, if $\alpha^+>1$, we have (5) is maximal for $c=0$ and $r=r_{\max}$. In summary, we have:

$$\Delta^+ = \max\left(\beta^+, \alpha^+\beta^+ r_{\max}^{\alpha^+-1}\right).$$

An analogous argument leads to:

$$\Delta^- = \max(\beta^-, \alpha^-\beta^-(|D|-r_{\min})^{a^--1}).$$

and the overall Δ for U_c can be written as:

$$\Delta = \max(\Delta^-, \Delta^+).$$

Adding Gaussian noise

For a fixed SD σ, the release value density at r corresponding to adding Gaussian noise is proportional to

$$\exp\left(-\frac{1}{(2\sigma^2)}(r-c)^2\right)$$

. The corresponding level of differential privacy is:

$$\log\left(\frac{\exp\left(-\frac{1}{(2\sigma^2)}(r-c)^2\right)}{\exp\left(-\frac{1}{(2\sigma^2)}(r-c+1)^2\right)}\right) \geq \frac{1}{(2\sigma^2)}\left((r_{\max}-r_{\min})+1\right) \quad (6)$$

For a mechanism adding Gaussian noise, the differential privacy ε is at least (6). For a SD of 1.33 and $r_{\min}=3$, values suggested in the analysis of Murphy and Chueh[20] and $r_{\max}=10^6$, which is smaller than the number of patient records in typical academic medical centers where such count query tools are deployed, we get the differential privacy afforded has an $\varepsilon>282661$ as opposed $\varepsilon=2.037$ in our system. Conversely, if we require $\varepsilon=2.037$, we need to require Gaussian noise with a SD exceeding 495. Rounding the reported values to the nearest multiple of k does not fundamentally change this behavior. Consequently, the approach described by Murphy and Chueh[20] does not guarantee practical differential privacy. A similar analysis can be carried out for STRIDE.

Tracking privacy expenditures and privacy budgeting

In order to protect against averaging query responses, we can only allow a finite number of queries, determined by the SD of the response distribution as well as how accurate an estimate of the count we want to tolerate. The allowed number of queries is an example of a 'privacy budget' and it appears implicitly in the i2b2 system, which disallows more than a fixed number of queries that return the same true count.[20] However, averaging the result of repeating the same query is not the only form of attack.[11 22]

Under differential privacy, the total decrease in privacy resulting from a user's queries is at most the sum of the privacy afforded by each query. In general, querying k times using ε-differential privacy gives a total loss of $k\varepsilon$-differential privacy and, if the ith query p_i issued by a user has differential privacy ε_i associated with it, i queries cost $\sum_i \varepsilon_i$. This cumulative loss represents the statistical risk of breach for any attack, not just the repeated query attack described above.

For a total privacy budget $\varepsilon_{\text{total}}$, a user could ask $\varepsilon_{\text{total}}/\varepsilon_i$ queries with cost ε_i each before exhausting their budget, after which there can be another administrative review of their project. A simple modification of the system could allow users to pick from a predefined set of ε levels per query. For preliminary queries where the expected true count is large, the user could use a small level of ε and incur more noise, thereby 'saving' the privacy budget for later narrower queries. Because of the properties of differential privacy, the total risk to any individual in the database is not affected by this flexibility as long as the total expenditure stays the same.

The privacy budget $\varepsilon_{\text{total}}$ can be chosen according to a person's role in the institution and how trusted they are. We can then associate the level of privacy protection with the level of trust; more trusted users can use a larger ε_i and obtain more accurate counts than less trusted users. We call such an arrangement a 'multi-trust level architecture'. Because such a system can monitor and track the total privacy expenditures of users, it can automatically flag users who expend their privacy budget, facilitating the auditing of query logs and simplifying administrative overheads for approving preliminary research.

Optimizing usefulness of individual queries

In this section we describe how statistical properties of the returned counts depend on the parameters. We also give some interpretation of how user parameters and privacy level interact with the utility shape and reported response mean and variance in (2) and (3). We can determine parameter settings that yield responses with a specified mean and variance (eg, those given by existing count query systems).

The utility function is determined by the corresponding α and β parameters. If $\alpha=1$, the utility decreases linearly with distance from the real count with slope β. If $\alpha<1$, we decrease the slope to an increasing degree the further away we get from the real count. The effect is to 'flatten' out the utility moving further away from the real count. If $\alpha>1$, we do the opposite. Setting $\beta^-<\beta^+$ gives a bias towards underestimation while $\beta^->\beta^+$ gives a bias toward overestimation. Figure 1 shows how utility preferences translate into the output distribution.

Given a true count of 50, figure 2 shows the mean and variance of the response distribution for fixed β as we vary ε. In all cases, the variance of the response decreases as ε increases. Small ε corresponds to more privacy and hence a larger perturbation.

An important point illustrated in figure 2 is that the variance of the error can be quite large for small ε. For symmetric utilities, there is no change in the response variance with changing β for fixed ε. For linear asymmetric utilities, the response variance is shown in figure 3A for a few different values of ε. As the privacy requirement becomes higher, ε is smaller and larger asymmetry in the utility results in higher variance. On the other hand, as shown in figures 2 and 3, reducing the degree of asymmetry reduces the variance.

Introducing non-linear utilities with $\alpha<1$ increases the variance, as shown in figure 3B. The same happens in general for $\alpha>1$. However, reductions in variance can be achieved by applying $\alpha=1+\delta$ for small positive δ on the side with smaller Δ as long as this

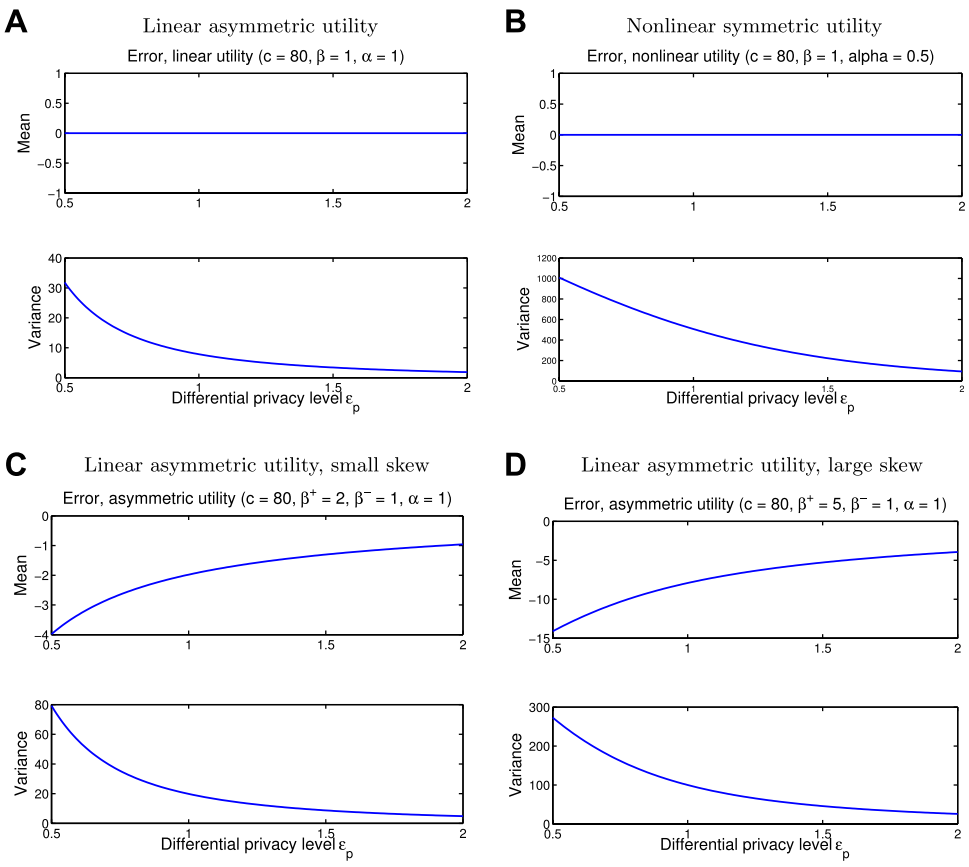

Figure 2 (A–D) Mean and variance of noise values for true count equal to 80 versus the differential privacy parameter ε_P. For symmetric noise distributions the mean of the noise is effectively 0, but in the asymmetric case the mean increases as the privacy level increases.

does not increase this Δ to become the larger of the two. As an example, consider the parameter settings in figure 4. Here $\Delta^-=1$ and $\Delta^+=3$. We can increase α^- from 1 to a little more than 1.128 and still have $\Delta = \max(\Delta^-, \Delta^+) = 3$, but decreasing the variance from 9.25 to 5.60. However, the mean also changes from 36.08 to 36.70, decreasing the underestimation bias.

The preceding description is meant to illustrate how to trade off different parameters in the system design. In order to facilitate the exploration of parameter settings, we constructed the graphical web-based tool shown in figure 4. This can be used to develop preset options for users to select when issuing queries to the database.

RESULTS

We created an open source tool that fully implements the privacy protection mechanism of the system and also includes a graphical interface for tuning system parameters and exploring user preferences. The implementation only employs about 240 lines of JavaScript and 107 lines of standard HTML, and is available at http://ptg.ucsd.edu/cq and by request from the authors.

In order to show that our proposed system provides similar utility to i2b2, we simulated four queries 1000 times each with true counts of 600, 430, 250, and 80. Figure 5 shows the histogram of returned values for these queries for i2b2 with SD 1.33 and our proposed system with $\varepsilon=2$ (yielding variance slightly larger than 1.33). These values represent returned counts from a sequence of queries designed to identify a cohort. While both methods returned similar value ranges, our method returned the true count more often (by a factor 1.61) while at the same time guaranteeing 2-differential privacy.

How changing the privacy expenditure per query can help is shown in table 1, where a sample run of queries was executed

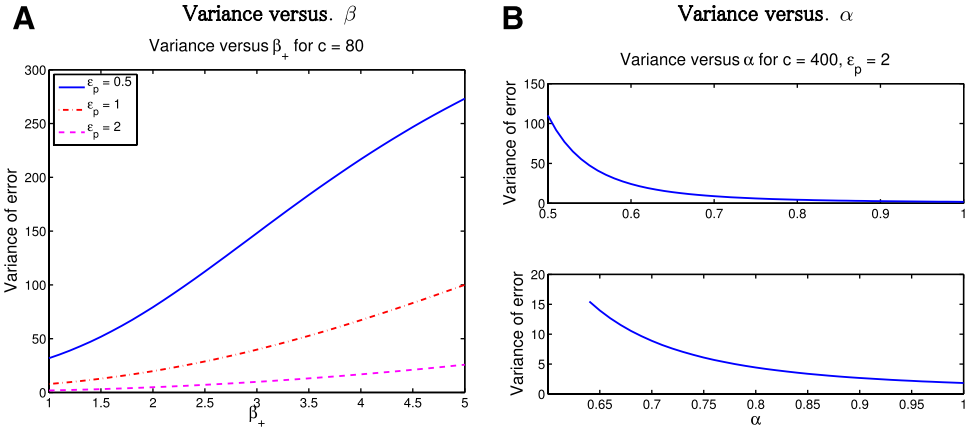

Figure 3 Variance of noisy counts as functions of β and α. (A) Variance of noise for true count equal to 80 versus parameter β^+ when $\beta^-=1$ in the linear asymmetric utility. For larger ε, the variance is smaller even with a larger degree of asymmetry. (B) Variance of noise for true count equal to 400 versus α in the non-linear symmetric utility with $\beta=1$. The upper plot shows a larger range of α values and the lower plot is zoomed in for larger α.

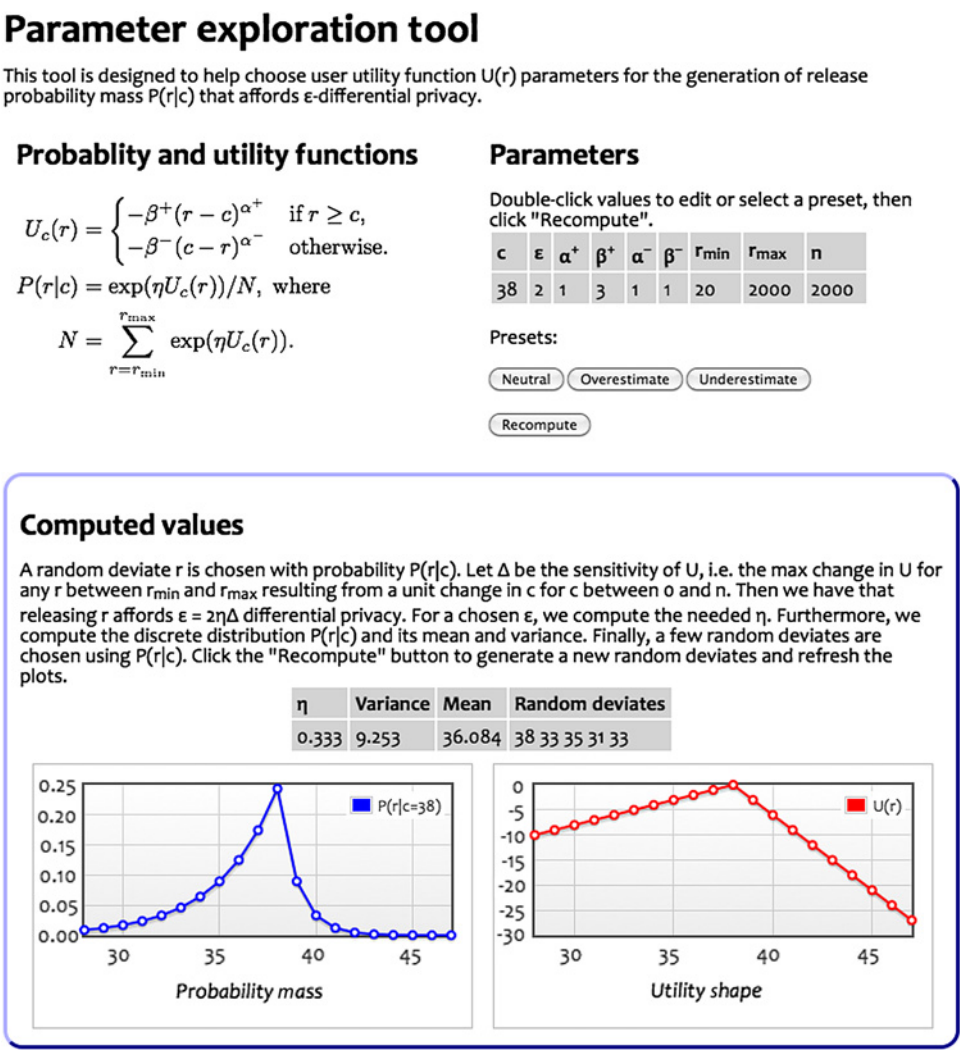

Figure 4 Screen capture of prototype parameter exploration tool. The parameters are set to the underestimation preset. The shape of the utility function, the resulting probability mass function together with its mean and variance are shown, as well as five random deviates, each of which represents a possible system response. As can be seen from these five random deviates, the biasing towards underestimation was successful.

using our tool. In the first run the privacy expense is ε_i=1 per query, which results in too accurate counts for broad queries (in the upper rows of the table) and less accurate counts for narrower queries with more clauses (in the lower rows). In the second run we can choose ε_i to vary per query, expending more of the privacy budget in the narrower queries. This results in a more accurate count. However, in both runs the total privacy risk is ε=5. This shows how quantifiable privacy can help give users some flexibility over a 'one size fits all' solution.

Finally, we applied our tool to the scenarios for researcher A and researcher B described earlier. The results obtained for researcher A are shown in figure 4. For researcher B, using overestimation (β^-=3) produces responses of 84, 86, 86, 86, and 88 from a distribution with a mean of 86.95 and variance of 9.84. As can be seen, while we are over- and underestimating, we are not forced to do so grossly.

DISCUSSION

In accordance with our objective, we have shown in detail how existing count query systems can be extended to provide strong privacy guarantees, allow the incorporation of user preferences with respect to individual query responses, and allow budgeting of privacy loss over time. In this study we address two fundamental questions that arise: (1) how much privacy does perturbing the count yield, and (2) how does the perturbation of counts affect the usefulness to the end user? To optimize the usefulness of the individual perturbed counts returned, we have proposed a new method for perturbing the true counts that is not based on noise addition but, instead, employs the exponential mechanism of McSherry and Talwar.[21] Our method provides a specified level of differential privacy[13] and explicitly translates user preferences into a distribution on approximate count values from which the returned count is drawn. Using differential privacy, we can quantify how well our mechanism protects information with regard to individuals, including their identity.

The strength of our differential privacy guarantees stems from very conservative assumptions. First, the privacy risk to each individual is assessed assuming that all other records of the individual have been revealed. Second, queries are treated independently and privacy risk increases additively per query. Ongoing theoretical work seeks to mitigate these assumptions either through additional modeling of the data[23] or through processing queries in a non-interactive or batch manner.[24] With an extension to multiple query types, our system resembles PINQ.[25] PINQ is a specific prototype query language which supports privacy-preserving queries on databases with the goal of making the response mechanism transparent to the end user. In contrast, our goal here is to let the user tune the response mechanism for his or her specific needs. Furthermore, rather than creating a system-dependent solution, we propose

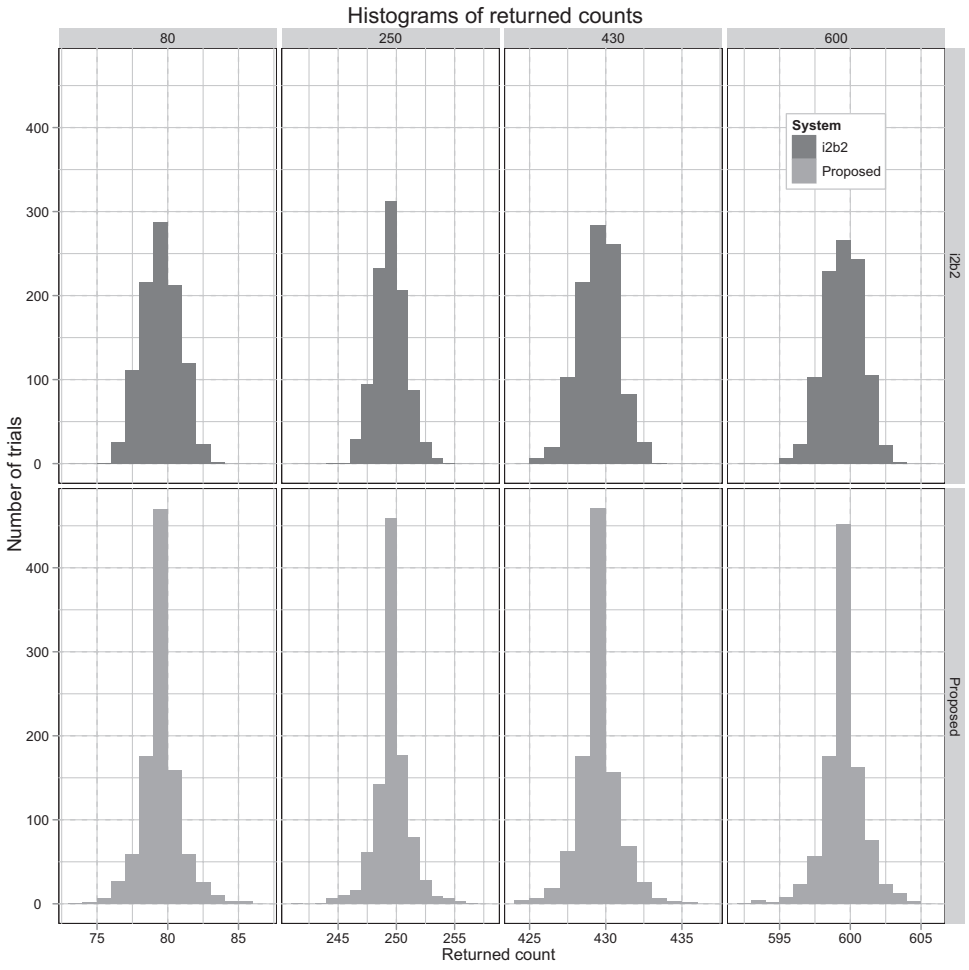

Figure 5 Comparison of count perturbations of the proposed system with those of i2b2.

modifying the response mechanism of existing study design tools to provide quantifiable privacy and maintain the privacy budget.

Our approach offers the use of a total 'privacy budget' $\varepsilon_{\text{total}}$ and a bound on the maximum privacy risk per query $\varepsilon_i^{\max}$ that limits the total number of queries. Without these limitations a user could approximately recover the entire database from perturbed queries.[12] Many differential privacy methods assume that access to the query tool is public[12 22 23] and hence the total privacy budget may be too small to make an effective study design aid. However, in institutional interactive query systems, administrators can control the environment by restricting access to trusted users, allowing larger budgets. If a user exhausts their privacy budget, it can be renewed after a review. A natural point at which renewal can happen in is when the IRB approves a study and allows the researcher access to the data. The quantity $\varepsilon_i^{\max}$ should be set such that the error per query is sufficiently small for effective cohort identification, and the total privacy budget should be set to the typical number of queries needed to identify an appropriate study cohort, which can be as high as 100 (Murphy SN, personal communication, 2011). In the end, determining these numbers is a policy question that an institution can address in consultation with statisticians. Regardless of regulations, institutions may continue to protect themselves by adopting stronger privacy oversight for patient data than required.

We envisage three ways of extending i2b2 and STRIDE by replacing their noise addition with our mechanism: (1) only offering symmetric linear utility and ε=2.037; (2) offering select simple preset parameter profiles developed by system designers and statisticians; and (3) offering a tool like ours for users to develop their own presets. All three approaches offer privacy budgeting and all three can co-exist in a single implementation.

The tool we have created demonstrates both the feasibility of such a system and the practicality of implementing it. This tool will be useful both for selecting profiles for a simplified user

Table 1 Two example runs of queries

Query	True count	Privacy cost ε_i	Perturbed count	Privacy cost ε_i	Perturbed count
Admitted with heart failure	6000	1	5996	0.5	5977
+ within last year	600	1	599	0.5	606
+ previous diagnosis of hypertension	430	1	433	1	436
+ male	250	1	251	1	246
+ under 65	80	1	70	2	79

Column 1 indicates clauses added to the query from the preceding row, column 2 is the true count of individuals satisfying the query in a hypothetical database, columns 3 and 4 are for constant ε_i per query, and columns 5 and 6 show the query results from changing ε_i per query.

interface for users who are not inclined to explore individual parameter settings and for privacy policy makers who want to explore the consequences of different privacy level settings. Furthermore, our approach is already partially implemented in our Clinical Data Warehouse for Research at the University of California San Diego with a full implementation planned.

Because privacy in our system is quantified using a common privacy 'currency', future systems can allow queries for statistics beyond simple counts provided these are answered by methods that guarantee differential privacy. Current such methods include learning models by empirical risk minimization including logistic regression and support vector machines,[26] and producing robust descriptive statistics and estimators.[27] This has the potential to enhance the ability of researchers to design studies while staying within their privacy budget, as well as providing the basis on which an enterprise-wide comprehensive privacy architecture can be built.

CONCLUSION

Counts supplied by a count query tool for study design must be perturbed to provide patient privacy. We have presented a practical approach to extending current query tools to provide provable privacy guarantees that at the same time allows users to tailor system responses to suit their needs and preferences. In consequence, the presented approach yields both increased control of privacy risks as well as usefulness to the end user. The approach also serves as an example of a service that can be incorporated in an enterprise-wide system for tracking privacy and accountability.

Acknowledgments We thank the associate editor and anonymous reviewers for their comments.

Contributors SAV and ADS developed the proposed method. AAB helped with creation of the use cases and with testing and feedback. All authors contributed to the writing of the manuscript.

Funding This work was supported by NIH grants R01 LM07273, UL1RR031980, U54 HL108460 and CALIT2 at UC San Diego.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement This work does not produce data.

REFERENCES

1. **Dick RS,** Steen EB. *The Computer Based Patient Record: An Essential Technology for Health Care*. Washington, DC: The National Academies Press, 1991.
2. **Sweeney L.** K-anonymity: a model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems* 2002;**10**:557—70.
3. **Machanavajjhala A,** Kifer D, Gehrke J, *et al*. L-diversity: privacy beyond k-anonymity. *ACM Trans Knowl Discov Data* 2007;**1**:3.
4. **Li N,** Li T, Venkatasubramanian S. T-closeness: privacy beyond k-anonymity and l-diversity. *Proceedings of ICDE* 2007:106—15. doi:10.1109/ICDE.2007.367856
5. **Malin B.** K-unlinkability: a privacy protection model for distributed data. *Data Knowl Eng* 2008;**64**:294—311.
6. **Benitez K,** Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc* 2010;**17**:169—77.
7. **El Emam K,** Brown A, AbdelMalik P. Evaluating predictors of geographic area population size cut-offs to manage re-identification risk. *J Am Med Inform Assoc* 2009;**16**:256—66.
8. **Winkler WE.** Masking and re-identification methods for public-use microdata: overview and research problems. In: Domingo-Ferrer J, Torra V, eds. *Privacy in Statistical Databases*. New York: Springer, 2004:519.
9. **Sweeney L.** Weaving technology and policy together to maintain confidentiality. *J Law Med Ethics* 1997;**25**:98—110.
10. **Narayanan A,** Shmatikov V. Robust de-anonymization of large sparse datasets. *Proceedings of 2008 IEEE Symposium on Security and Privacy* 2008:111—25.
11. **Ganta SR,** Kasiviswanathan SP, Smith A. Composition attacks and auxiliary information in data privacy. *Proceedings of 14th ACM SIGKDD* 2008:265—73. doi:10.1145/1401890.1401926
12. **Dinur I,** Nissim K. Revealing information while preserving privacy. *PODS '03: Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. New York: ACM, 2003:202—10. doi:10.1145/773153.773173
13. **Dwork C,** McSherry F, Nissim K, *et al*. Calibrating noise to sensitivity in private data analysis. *3rd IACR TCC* 2006:486—503. doi:10.1007/11681878_14
14. **Institute of Medicine.** *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. http://www.iom.edu/Reports/2009/Beyond-the-HIPAA-Privacy-Rule-Enhancing-Privacy-Improving-Health-Through-Research.aspx (accessed 1 Mar 2012).
15. **Fost N,** Levine RJ. The dysregulation of human subjects research. *JAMA* 2007;**298**:2196—8.
16. **Gostin LO,** Nass S. Reforming the HIPAA privacy rule: safeguarding privacy and promoting research. *JAMA* 2009;**301**:1373—5.
17. **Armstrong D,** Kline-Rogers E, Jani SM, *et al*. Potential impact of the HIPAA privacy rule on data collection in a registry of patients with acute coronary syndrome. *Arch Intern Med* 2005;**165**:1125—9.
18. **Murphy SN,** Mendis ME, Berkowitz DA, *et al*. Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc* 2006:1040. PMID: 15911725.
19. **Lowe HJ,** Ferris TA, Hernandez PM, *et al*. STRIDE—an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc* 2009:391—5.
20. **Murphy SN,** Chueh H. A security architecture for query tools used to access large biomedical databases. *AMIA, Fall Symposium 2002* 2002:552—6. PMID: 12463885.
21. **McSherry F,** Talwar K. Mechanism design via differential privacy. *FOCS* 2007:94—103. doi:10.1109/FOCS.2007.66
22. **Dwork C,** Yekhanin S. New efficient attacks on statistical disclosure control mechanisms. *Advances in Cryptology—CRYPTO 2008*. 2008:469—80. doi:10.1007/978-3-540-85174-5_26
23. **Roth A,** Roughgarden T. Interactive privacy via the median mechanism. *Proceedings of the 42nd ACM Symposium on Theory of Computing*. New York: ACM, 2010:765—74. doi:10.1145/1806689.1806794
24. **Blum A,** Liqett K, Roth A. A learning theory approach to non-interactive database privacy. *Proceedings of the Annual ACM Symposium on Theory of Computing*. New York: ACM, 2008:609—17. doi:10.1145/1374376.1374464
25. **McSherry FD.** Privacy integrated queries: an extensible platform for privacy: preserving data analysis. *SIGMOD '09: Proceedings of the 35th SIGMOD International Conference on Management of Data*. New York: ACM, 2009:19—30. doi:10.1145/1559845.1559850
26. **Chaudhuri K,** Monteleoni C, Sarwate AD. Differentially private empirical risk minimization. *J Mach Learn Res* 2011;**12**:1069—109.
27. **Dwork C,** Lei J. Differential privacy and robust statistics. *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*. New York: ACM, 2009:371—80. doi:10.1145/1536414.1536466