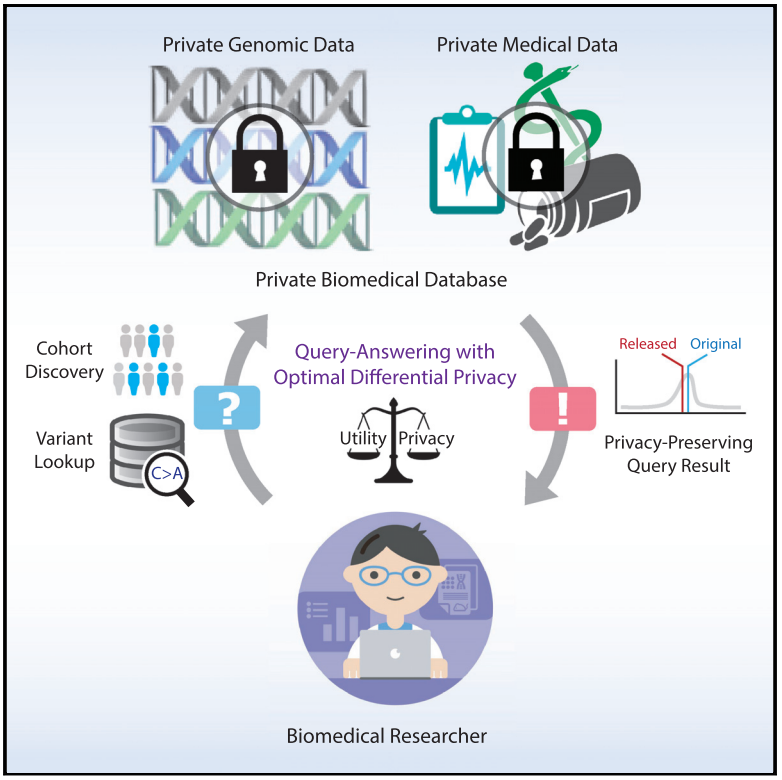
Cell Systems

Article

Privacy-Preserving Biomedical Database Queries with Optimal Privacy-Utility Trade-Offs

Graphical Abstract



Highlights

- We present privacy-preserving mechanisms for answering biomedical database queries
- We enable more accurate cohort discovery and variant lookup over existing methods
- Accuracy of our method is provably optimal under the theory of differential privacy

Authors

Hyunghoon Cho, Sean Simmons, Ryan Kim, Bonnie Berger

Correspondence

hhcho@broadinstitute.org (H.C.), bab@mit.edu (B.B.)

In Brief

Large-scale biomedical databases often provide interactive query-answering systems, which allow researchers to utilize the scientific insights from these data without accessing sensitive individual-level data. However, privacy concerns remain, as the query results can still leak sensitive information about individuals in the database. We introduce privacy-preserving mechanisms for answering key biomedical queries, including cohort discovery and variant lookup. Our method provably minimizes the loss of accuracy for any desired level of privacy. Our results demonstrate enhanced accuracy over existing methods and illustrate real-world use cases.

Cho et al., 2020, Cell Systems *10*, 408–416
May 20, 2020
https://doi.org/10.1016/j.cels.2020.03.006

CellPress

CellPress
OPEN ACCESS
Focus on RECOMB

Cell Systems

Article

Privacy-Preserving Biomedical Database Queries with Optimal Privacy-Utility Trade-Offs

Hyunghoon Cho,[1,3,5,*] Sean Simmons,[1,4,5] Ryan Kim,[2,3] and Bonnie Berger[1,3,4,6,*]
[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[2]Harvard University, Cambridge, MA 02138, USA
[3]Computer Science and AI Laboratory, MIT, Cambridge, MA 02139, USA
[4]Department of Mathematics, MIT, Cambridge, MA 02139, USA
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: hhcho@broadinstitute.org (H.C.), bab@mit.edu (B.B.)
https://doi.org/10.1016/j.cels.2020.03.006

SUMMARY

Sharing data across research groups is an essential driver of biomedical research. While interactive query-answering systems for biomedical databases aim to facilitate the sharing of aggregate insights without divulging sensitive individual-level data, query answers can still leak private information about the individuals in the database. Here, we draw upon recent advances in differential privacy to introduce query-answering mechanisms that provably maximize the utility (e.g., accuracy) of the system while achieving formal privacy guarantees. We demonstrate our accuracy improvement over existing approaches for a range of use cases, including cohort discovery, variant lookup, and association testing. Our new theoretical results extend the proof of optimality of the underlying mechanism, previously known only for count queries with symmetric utility functions, to more general utility functions needed for key biomedical research workflows. Our work presents a path toward interactive biomedical databases that achieve the optimal privacy-utility trade-offs permitted by the theory of differential privacy.

INTRODUCTION

The fast accumulation of biomedical datasets around the globe, including personal genomes and medical records, hold immense potential to advance biology and medicine. However, most of these data are held in isolated data repositories (e.g., different hospitals or biobanks); sharing these data across repositories is often infeasible due to data privacy concerns. A pressing challenge is to develop systems that allow researchers to jointly leverage large data collections across multiple sites, in order to gain more accurate and refined biomedical insights crucial to realizing the vision of personalized health.

In response to this need, several large-scale databases in both the medical and genomics communities have developed interactive query-answering systems in order to allow external researchers and clinicians to utilize their databases in a limited and controlled fashion (Lemke et al., 2010; Saeed et al., 2011; Weber et al., 2009; Lowe et al., 2009; Murphy et al., 2011; Fiume et al., 2019). For example, medical data repositories, such as i2b2 and STRIDE (Lowe et al., 2009; Murphy et al., 2011), allow researchers designing clinical studies to query how many patients in the database satisfy a given set of criteria prior to requesting data access or recruiting patients, a workflow commonly known as cohort discovery. In addition, recently emerging genomic "beacon" services (Fiume et al., 2019) allow users to query whether or not a given genetic variant is observed in the database, a workflow we refer to as variant lookup. The Beacon project by the global alliance for genomics and health (GA4GH) (Global Alliance for Genomics and Health, 2016) has helped launch a network of over 100 beacons to date around the world (Fiume et al., 2019). These interactive systems are poised to play a key role in driving data-sharing efforts in the biomedical community.

Yet, despite the limited scope of allowed queries and the fact that only aggregate-level information is shared, query-answering systems can still leak sensitive information about the underlying individuals (Homer et al., 2008; Vinterbo et al., 2012; Shringarpure and Bustamante, 2015; Raisaro et al., 2017). One could, for example, ask for the number of 25-year-old females on a certain medication who do not have a particular disease. If the answer returned is zero, we know that any 25-year-old female in the database who is on that medication has that disease, a fact the patient might wish to keep private. Moreover, given access to an individual's genotype, a small number of queries to a beacon server would be sufficient to reveal whether the individual is included in the database (Raisaro et al., 2017; Shringarpure and Bustamante, 2015). This information could potentially be detrimental to the individual if the underlying cohort represents a group of individuals with sensitive characteristics (e.g., a rare medical condition).

To address these privacy concerns, existing systems and studies have attempted to improve individual privacy in query-

answering systems by perturbing query results with a small amount of noise in order to reduce sensitivity to the underlying individuals (Murphy and Chueh, 2002; Raisaro et al., 2017; Wieland et al., 2008). However, these efforts either lack rigorous theoretical guarantees of privacy or introduce an excessive amount of noise into the system, limiting their effectiveness in practice. For instance, i2b2 and STRIDE add truncated Gaussian noise to the number of subjects matching a given filter without consideration of formal models of privacy. Recent studies (Raisaro et al., 2019; Vinterbo et al., 2012) have proposed applying the theoretical framework of differential privacy (Dwork et al., 2006, 2015) to cohort discovery (see Method Details). As we demonstrate in our work, existing methods perturb the results more than is necessary to achieve a desired level of privacy. For beacon servers, although real-world systems are yet to adopt a standard protocol to remedy privacy risks, Raisaro et al. (2017) have explored potential attacks on beacon servers and provided a set of risk-mitigation strategies. Unfortunately, their techniques are based on a simplified model of genotype distribution, which make them vulnerable to sophisticated attacks that exploit the data patterns not captured by the model (e.g., deviations from Hardy-Weinberg equilibrium), as we describe in our results.

Here, we build upon recent advances in differential privacy (DP) to introduce query-answering systems with formal privacy guarantees, while ensuring that the query results are as accurate as theoretically possible. We focus on three types of queries: (1) cohort discovery, (2) variant lookup, and (3) chi-squared association tests built upon count queries. We empirically demonstrate the accuracy improvements of our proposed DP mechanisms for each query type. Furthermore, we provide case studies on how our optimal DP mechanisms could be used to enable data sharing with privacy, following the real-world workflows in published studies for cohort discovery and variant lookup. We also newly illustrate how a user's ability to choose a prior belief over the true answer can further boost the accuracy of the query results in our DP framework, without incurring additional privacy cost. To aid the reader's understanding, we provide additional background and define key vocabulary of our work in the Primer (Box 1) and Glossary (Box 2), respectively.

Our methods leverage the truncated α-geometric mechanism (α-TGM), previously developed for a limited class of count queries in a theoretical context (Ghosh et al., 2012), to obtain a differentially private result for each query type. Our key theoretical advances include showing that α-TGM, combined with a post-processing step performed by the user, provably maximizes the expected utility (encompassing accuracy) of the system for a broad range of user-defined notions of utility in both cohort discovery and variant lookup workflows. Notably, the optimality of α-TGM was previously known for only count queries with symmetric utility functions, which are insufficient for workflows we typically encounter in biomedical databases (Vinterbo et al., 2012). We newly extend this result to a more general class of utility functions, including asymmetric functions, thereby answering an open question posed in the original publication of α-TGM (Ghosh et al., 2012). Asymmetric utility is a desirable notion in cohort discovery applications, where overestimation is often more desirable than underestimation (Vinterbo et al., 2012). Moreover, our generalized notion of utility newly enabled us to prove the optimality of our DP mechanism for variant lookup queries. Our work shows how one can leverage the theory of DP to protect the privacy of individuals in biomedical query-answering systems, while simultaneously maximizing the benefits of data sharing for science.

RESULTS

Overview of Our System

Here, we describe the overall workflow of our optimal DP mechanism for biomedical database queries (Figure 1). Suppose the database D consists of data from n individuals, $d_1, \ldots, d_n$. First, the user chooses a desired privacy level, denoted by ϵ, and a query. For both cohort discovery and variant lookup scenarios, the query can be represented using a predicate s, which evaluates to true (1) or false (0) for each individual d_i in the database, indicating whether he or she matches the desired criteria for the cohort or has a variant of interest. The user is interested in $x := \sum_{i=1}^{n} s(d_i)$ for cohort discovery and $\mathbf{1}\{x>0\}$ for variant lookup, where $\mathbf{1}\{\cdot\}$ is the indicator function. The user submits (ϵ, s) to the database server. Next, the server releases a differentially private statistic $\tilde{x}$ of x to the user, using the truncated α-geometric mechanism (with $\alpha = \exp(-\epsilon)$; see Method Details). Then, the user chooses a prior distribution $\pi(x)$ over the true count x, representing his or her belief about the underlying data distribution, and a loss function $\ell(x,y)$, representing how undesirable obtaining a query result y is, given x, where $y \in \{0, \ldots, n\}$ for cohort discovery and $y \in \{0, 1\}$ for variant lookup. Based on the chosen π and ℓ, the user maps $\tilde{x}$ to the optimal choice of y, which the user interprets as the answer to their query. We refer to this final step as local post-processing by the user, which can be performed an arbitrary number of times for different choices of π and ℓ, without affecting privacy guarantees. Our scheme satisfies ϵ-DP and provably minimizes (maximizes) the expected loss (utility) of the query result among all stochastic mechanisms that achieve ϵ-DP, for both cohort discovery and variant lookup scenarios. A more thorough description of our mechanisms, including theoretical analyses of their optimality, are provided in Method Details.

Regarding the computational cost of our system, the overhead for the server is negligible as it randomly samples one additional number per query. Communication is identical to the underlying system without DP, except for the inclusion of a privacy parameter ϵ in the query. Local post-processing by the user takes $O(n^2)$ time per query for a database of size n in general; however, the structure of loss functions in our application scenarios can be exploited to reduce the complexity to $O(n\log(n))$ for cohort discovery and $O(n)$ for variant lookup (Method Details). Using these optimizations, which are incorporated into our software, we achieve post-processing times of less than a second for both query types on a standard laptop computer, even when n is a million. Thus, our methods incur a minimal computational overhead overall compared to query-answering systems without DP.

Differentially Private Cohort Discovery with Maximum Utility

We first evaluated the utility of our proposed approach for cohort discovery. For baseline, we compared our approach to the exponential mechanism proposed by Vinterbo et al., 2012, and the Laplace mechanism used in the MedCo framework of Raisaro et al., 2019. Note that the exponential mechanism takes a utility

Box 1. Primer

Query-answering systems for biomedical databases can allow researchers to obtain aggregate insights from the data without accessing sensitive individual-level data, thus broadening the scientific impact of these data resources. For example, large-scale patient registries that allow users to see how many patients in the database meet certain criteria can facilitate the design of research studies, such as clinical trials. Similarly, genomic data repositories that allow users to retrieve information about specific genetic variants can help researchers tap into insights offered from larger or more diverse datasets. However, these systems do not ensure adequate protection of privacy in general for the individuals in the database, as it is often possible to construct a set of queries that reveal information about a specific individual.

In this work, we are interested in designing query-answering systems with rigorous privacy guarantees based on the theoretical framework of DP. The idea behind DP is that, if we have a dataset from which we want to release some statistic, it should be difficult to tell the difference between the statistic calculated on our dataset and the statistic calculated on a dataset that differs from ours by exactly one individual. DP achieves this property by adding a controlled amount of statistical noise to the shared data in such a way that it ensures the statistical impact of any particular individual in the database is smaller than the desired level.

To use DP mechanisms in practice, one needs to balance the trade-off between privacy and utility. Note that adding less noise to the data is better in terms of ensuring the usefulness of the system, yet adding more noise leads to increased levels of privacy. This trade-off can be formalized through a utility function that quantifies how useful or desirable a particular perturbed query result is, given the true query result. The overall effectiveness of the system can then be measured by its expected utility at a given privacy level, which averages the utility function over different true query results as well as over the randomness in the DP mechanism.

Building upon recent results in DP, our work introduces optimal DP mechanisms that provably maximize the expected utility of the system for answering key types of biomedical database queries, including count queries for determining the size of a potential study cohort in a medical database (referred to as cohort discovery) and membership queries for finding out whether a genetic variant is included in a genomic database (variant lookup).

function as input, which we set to the negative of the loss function used in our mechanism. This linear mapping is motivated by the existing theory showing that exponential mechanism approximately maximizes the expected utility (McSherry and Talwar, 2007), which corresponds to minimizing the expected loss in our setting. In addition, because neither the exponential nor the Laplace mechanisms have a notion of prior distribution over the true count, we set our prior distribution to uniform, which represents a neutral prior.

As expected, our approach achieves the lowest expected loss across all values of privacy parameter ϵ for both symmetric and asymmetric loss of functions (Figure 2). In fact, our theoretical results suggest any mechanism that achieves ϵ-DP cannot perform better than our approach on an average in both symmetric and asymmetric settings (Method Details). Comparing the probability distribution over the query result, we see that our approach results in a more concentrated probability mass near the true count. It is worth noting that, with asymmetric loss, the mode of the distribution over the query result does not align with the true count, which is a result of the skewness of the loss function.

The fact that our margin of improvement over the Laplace mechanism is smaller than that over the exponential mechanism can be attributed to the fact that the Laplacian distribution, from which the perturbed count is sampled in the Laplace mechanism, closely approximates the geometric distribution used in our approach, which could be viewed as a discrete version of the Laplacian distribution. However, the Laplace mechanism is unable to tailor its answer to the user's desired notion of utility (captured by the loss function) or prior belief over the data, resulting in a greater loss in utility in more sophisticated settings. For example, our experiments with asymmetric loss shows that

Box 2. Glossary

ϵ-DP	A theoretical notion of privacy, which states that the probability of observing a certain outcome of the system, conditioned on the underlying database, does not differ by more than a multiplicative factor of $\exp(\epsilon)$ for any two databases that differ by exactly one individual. Non-negative real number ϵ is referred to as the privacy parameter, where smaller values of ϵ give stronger guarantees of privacy.
Utility (loss) function	A function that quantifies how desirable (undesirable) a certain outcome of the system is to the user. In this work, this function takes the true query result and a perturbed query result (from a DP mechanism) as input and outputs a real number.
Prior distribution	Probability distribution over a random variable, which is typically unobserved, expressing one's belief about the underlying distribution from which the variable is sampled. In this work, this term refers to the distribution over the true query result.
Cohort discovery	The workflow of determining how many individuals in a given database meet the desired criteria for inclusion in a study.
Variant lookup	The workflow of determining whether a genetic variant of interest is observed in a given (subset of the) database.

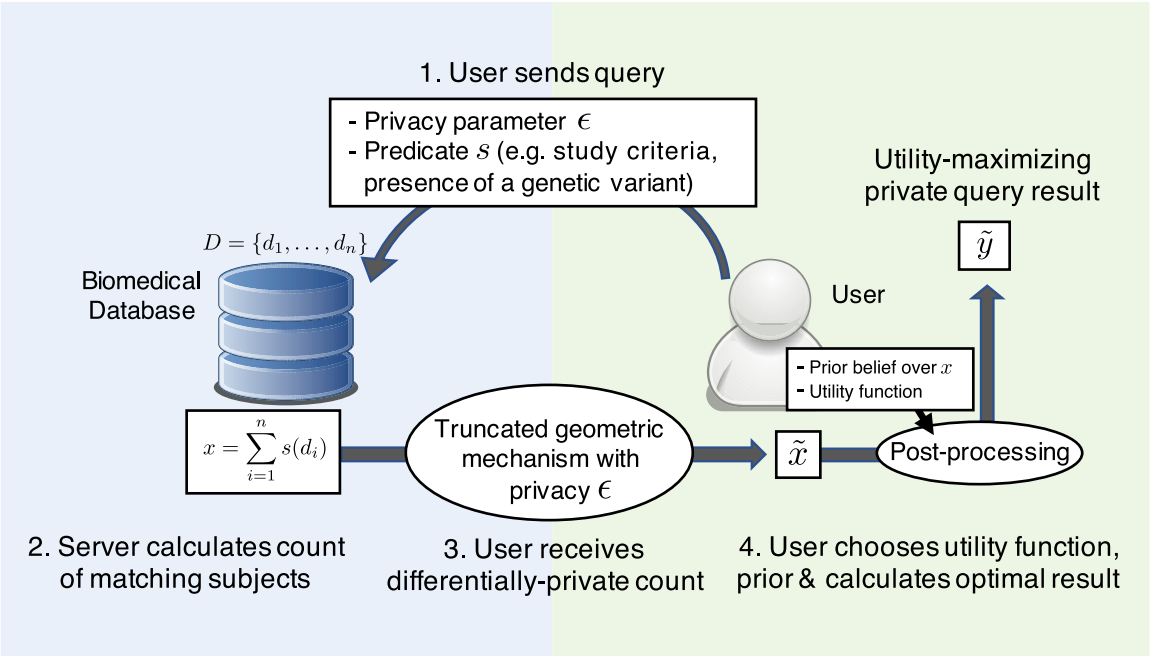


Figure 1. Workflow of Our Optimal Differential Privacy Mechanisms for Biomedical Queries

Given a count or membership query from the user, the database returns a differentially private count of individuals matching the query using the generic truncated α-geometric mechanism (Ghosh et al., 2012). The user locally transforms the result based on his or her chosen generic prior belief over the true count and a utility function. As we showed, the final result provably maximizes expected utility over all differential privacy mechanisms with the same privacy level.

the Laplace mechanism resorts to a suboptimal output distribution centered around the true count, leading to a more pronounced performance difference between our approach and the Laplace mechanism compared with the symmetric case (Figures 2B and 2C).

A key aspect of our approach is that it allows the user to incorporate his or her belief about the underlying data distribution. For example, if a user is interested in the number of individuals with a disease that is known to be extremely rare, then it would be desirable to take advantage of this prior knowledge about the disease to further improve the accuracy of the query result, rather than assuming that every possible answer is equally likely (uniform prior). To demonstrate this capability, we performed an experiment where the query distribution is concentrated on small numbers (e.g., <5 out of 1,000). We then evaluated the expected loss of our mechanism based on different choices of prior distributions of varying skewness, ranging from uniform to one that is highly skewed toward the low counts (Figure S1). Our results show that, indeed, adopting a prior distribution that is better aligned with the true result further reduces the expected loss of our system.

Differentially Private Variant Lookup with Maximum Utility

Variant lookup queries are becoming increasingly relevant for the biomedical community given the growing number of genomic beacon servers. Previously proposed privacy risk-mitigation strategies for beacons (Raisaro et al., 2017) have aspects that are reminiscent of DP (e.g., the notion of privacy budget), yet their privacy guarantees were based upon a simplified statistical model of genotype distributions in a population. As a result, the proposed strategies do not necessarily provide protection against attacks that take advantage of data patterns that lie outside of the model. For example, based on the 1,000 Genomes Project dataset (2015), we observed that selectively targeting genetic variants that deviate from Hardy-Weinberg equilibrium can lead to greater leakage of private information than captured by the previously proposed privacy budget accounting scheme (Raisaro et al., 2017) (Figure S2). Thus, theoretical frameworks, such as DP offer a valuable, more thorough alternative to mitigating privacy risks in beacon servers.

Our theoretical results enabled us to design an optimal DP mechanism for variant lookup queries that maximizes a user-defined notion of utility for any desired privacy level (Method Details). In our experiments, we consider two types of loss functions shown in Figure 3. Linear loss puts a linearly increasing penalty on answering that the variant is not found as its true count in the database grows (i.e., $\ell(c, 0) = c$ for $c>0$). Uniform loss equally penalizes all incorrect query answers regardless of how many times the query variant was observed in the database. In both cases, to additionally penalize false positive findings, the penalty for answering that the variant is found in the database when it is not (i.e., $\ell(0, 1)$) can be set to a higher value than one, which controls the desired balance between false positive and false negative findings. Note that our framework allows users to choose their own loss function as long as it satisfies a straightforward assumption that the wrong answer is not preferred to the correct answer (Method Details).

To construct a realistic prior distribution on the true count x, we leverage the allele frequency (AF) breakdown of real-world variant lookup queries submitted to the ExAC browser

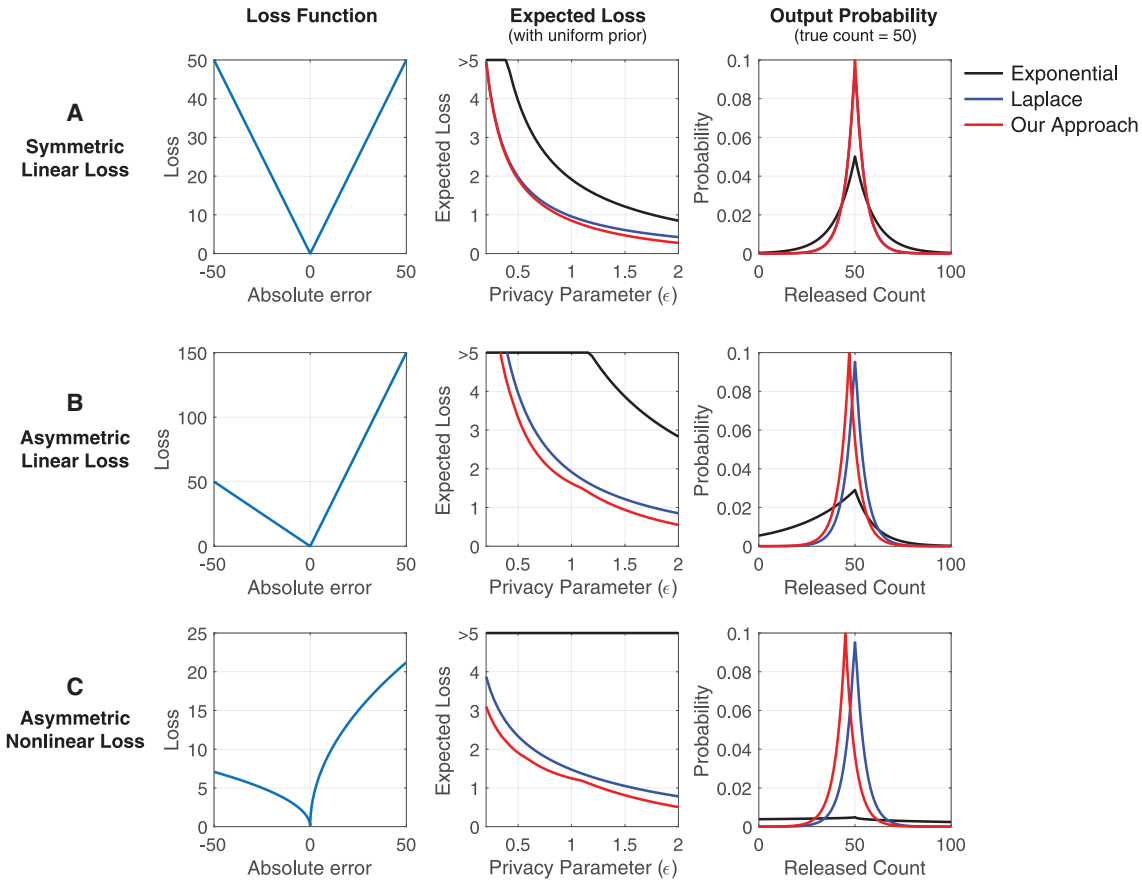


Figure 2. Our Approach Improves the Utility of Medical Cohort Discovery with Differential Privacy

We compared the performance of our optimal DP mechanism for count queries to the exponential (Vinterbo et al., 2012) and Laplace (Raisaro et al., 2019) mechanisms for different choices of loss functions (rows). We considered the following parameter settings for Vinterbo et al. (2012)'s loss function described in Method Details: (A) $\alpha_- = \alpha_+ = \beta_- = \beta_+ = 1$; (B) $\alpha_- = \alpha_+ = 1$, $\beta_- = 1$, and $\beta_+ = 3$; and (C) $\alpha_- = \alpha_+ = 0.5$, $\beta_- = 1$, and $\beta_+ = 3$. In each row, the subfigures show the shape of loss function (left), expected loss over a range of privacy parameters ϵ (center), and a sample probability distribution over the private query result, with a true count of 50 and $\epsilon = 0.2$ (right). We used the uniform prior for our mechanism. Overall, our approach reduces the expected loss while maintaining the same level of privacy. See also Figure S1.

(Karczewski et al., 2017) over a period of 12 weeks, provided by Raisaro et al. (2017) (Figure 3A). Assuming a uniform distribution within each AF category defined by Raisaro et al. (2017), we transformed the prior over AF into a prior over the number of individuals in the database with the query variant, which we used both in our optimal mechanism and for averaging the results over the real-world query distribution.

We compared our approach to standard DP techniques, including the exponential and the Laplace mechanisms. For the exponential mechanism, we set the utility to be the negative of our loss function and release a binary answer directly based on the true count x of the query variant. For the Laplace mechanism, we first use it to obtain a privacy-preserving count $\tilde{x}$ of x, then apply a threshold $\mathbf{1}\{\tilde{x}>0\}$ to obtain a binary answer. Both these approaches represent a straightforward application of the existing techniques.

Our results show a significant improvement in utility over existing mechanisms across a range of privacy parameters ϵ, with respect to linear loss (Figure 3B). In the case of uniform loss, our expected loss minimization problem reduces to minimization of the overall error probability of the system, since the loss function evaluates to one for the incorrect answer and zero for the correct answer in all cases. Under this setting, our mechanism also achieves smaller error probabilities compared to the existing DP techniques (Figure 3C).

Differentially Private Association Tests with Improved Accuracy

Count queries used in cohort discovery represent a fundamental operation that can enable a range of downstream analyses in a privacy-preserving manner, including association tests and histogram visualizations. Given that our optimal DP framework improves the accuracy of count queries while achieving the same level of privacy, it is expected that these improvements will also translate into accuracy improvements in downstream tasks. Here, we set out to empirically demonstrate this idea using the χ^2 association test as an example, which is a foundational technique for testing independence between pairs of categorical variables (e.g., clinical, genomic, and demographic characteristics of individuals).

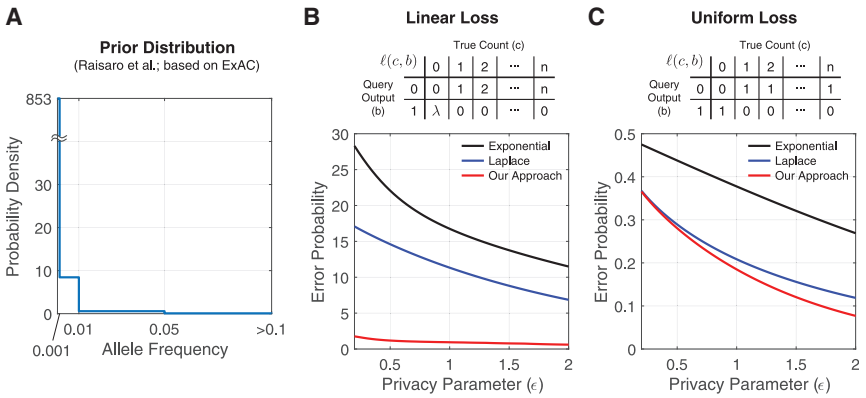


Figure 3. Our Approach Improves the Utility of Genomic Variant Lookup with Differential Privacy

Using the allele frequency distribution of variant lookup queries submitted to the ExAC database (Raisaro et al., 2017) as our prior distribution (A), we evaluated the expected loss of our optimal differential privacy mechanism for variant lookup with two different loss functions shown in tables: (B) linear loss, which employs a linearly increasing penalty for negatively answering the query as the number of observations in the database grows and an increased penalty for false positives parameterized by λ, and (C) uniform loss, where any error that results in a flipped query result incurs the same amount of penalty, in which case the expected loss can be interpreted as error probability. Overall, our approach achieves the lowest expected loss compared to existing differential privacy mechanisms.

To this end, we generated a large collection of sample 2-by-2 contingency tables with varying degrees of dependence between the two binary variables (one represented by the rows and the other by the columns). Assuming these tables as ground truth, we then obtained a differentially private version of each table by separately invoking differential private count query on each of the four cells with privacy parameter ϵ. (Note that due to the parallel composition property of differential privacy (Method Details), this overall procedure satisfies ϵ-DP, instead of 4ϵ.) We then computed the χ^2 association statistic based on both the true table and the perturbed table for comparison. As expected, we observed that the differentially private χ^2 statistics calculated based on our optimal mechanism more accurately matched the ground truth as opposed to other baseline mechanisms (Figure 4), though the Laplace mechanism achieves comparable performance due to its similarity to our mechanism under symmetric linear loss we adopted in this experiment. These results illustrate the broad impact of our approach beyond cohort discovery and variant lookup workflows. It would be interesting to see how these count query-based approaches compare to other DP mechanisms tailored for chi-squared tests (Uhlerop et al., 2013) and whether one can design an optimal mechanism for calculating chi-squared statistics based on our techniques.

Case Studies

Here, we provide case studies that illustrate how our DP mechanisms could be used to enhance privacy in key data-sharing workflows in biomedicine.

First, to demonstrate cohort discovery with privacy, we follow the study of Kullo et al. (2016), which investigated the effectiveness of genetic risk score (GRS) in improving health-related outcomes associated with coronary heart disease (CHD). Kullo et al. (2016) set out to recruit study participants in the Mayo Clinic Biobank who meet the following criteria: "age 45–65 years, non-Hispanic White ethnicity, no history of atherosclerotic cardiovascular disease, not on statins, at intermediate risk for CHD (10 year CHD risk 5%–20%), and residents of Olmsted County Minnesota." In addition, the study required at least a hundred subjects in both high GRS (odds ratio $\geq$1.1) and low-GRS (<1.1) groups in order to achieve sufficient power in statistical analysis. Fortunately, 2,026 subjects in the biobank met the inclusion criteria, and Kullo et al. (2016) were able to enroll 216 subjects ($\sim$100 in each GRS group) through targeted recruitment. This study led to an important finding that the patient's knowledge of GRS can bring tangible improvements to their physiological health—in this case a reduction in low-density lipoprotein cholesterol (LDL-C) levels, a key biomarker of CHD risk.

Our methods enable the design of query-answering systems that allow researchers like Kullo et al. (2016) to determine whether a sufficient number of individuals in a database meet the desired inclusion criteria, while protecting the privacy of individuals. Note that, given the potentially specific nature of inclusion criteria, answering arbitrary queries without any measures for privacy protection could leak sensitive information about the individuals; for example, in the above scenario, one may infer that a specific individual has a high GRS for CHD by carefully designing the queries. In Table 1, we report differentially private query results returned by our mechanism for different values of the true count. For large counts such as 2,000 (representative of Kullo et al. (2016)'s scenario), even a small value of privacy parameter ($\epsilon = 0.05$) leads to a highly accurate result (e.g., 1,998). Smaller counts require more noise to be added for privacy in general, but the results are still accurate enough for the user to make informed decisions. For instance, when the true count is 100, a perturbed query result of 88 ($\epsilon = 0.05$) would give Kullo et al. (2016) enough information to consider broadening the inclusion criteria or finding other patient registries in order to recruit enough participants.

To illustrate the variant lookup workflow, we consider a recent study by Velmeshev et al. (2019), which explored cell-type-specific gene expression patterns in autism spectrum disorder (ASD) patients. In an effort to identify genetic determinants that underlie pathological changes in gene expression, Velmeshev et al. (2019) performed whole exome sequencing of the ASD patients and reported high-confidence variants with a likely role in the disease, leveraging a series of stringent filters (e.g., low AF in existing databases). Despite the fact that many of these variants were located in genes with known association with ASD, a deeper understanding of their functional impact on ASD remains elusive. One approach to gaining further insight into a set of poorly understood variants is to look up their previous occurrences (if

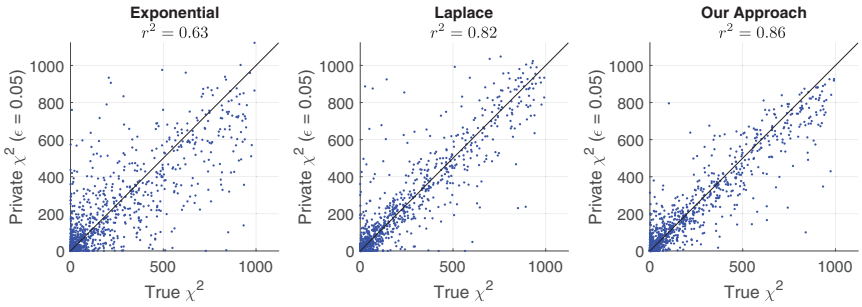


Figure 4. Our Approach Improves the Accuracy of Chi-Squared Association Tests with DP

We simulated datasets each including 1,000 paired observations of two binary random variables with varying strengths of association. We considered a differentially private (DP) scheme for releasing the chi-squared association statistic between the two variables for each dataset, where we use DP count queries to construct a 2 × 2 contingency table, based on which the statistic is computed as post-processing. The subfigures show the agreement between the DP statistics and the ground truth for different choices of DP count query mechanisms. We optimized our mechanism with respect to uniform prior and a symmetric linear loss function. Our approach achieves the best accuracy overall, comparable to the Laplace mechanism yet considerably more accurate than the exponential mechanism.

any) in other databases, in order to transfer existing knowledge or to prioritize further analysis. For instance, in the case of Velmeshev et al. (2019), one may be interested in finding out whether any of the identified variants have been previously observed in a large-scale biobank of ASD patients (e.g., SPARK; SPARK Consortium, 2018). However, answering such type of queries in an unrestricted fashion does not respect the privacy of individuals in the database, because a small set of queries could reveal whether someone is in the database, thus disclosing his or her disease status.

Our DP mechanism allows researchers to obtain answers to variant lookup queries while protecting the privacy of individuals. To illustrate, we queried the top ASD variants reported by Velmeshev et al. (2019) against the ClinVar database (Landrum et al., 2020) to check whether or not each variant has previously been annotated by experts to be pathogenic. Table 2 shows example query results returned by our differentially private variant lookup mechanism ($\epsilon = 0.2$ for each query), along with the number of expert assessments in ClinVar that marked each variant as pathogenic. Overall, our mechanism faithfully returned the majority of query results (11 out of 17); note that some amount of error is necessary in order to achieve DP. Analogous to the cohort discovery setting, our results are more accurate for variants that have more occurrences in the database, which reveal less information about a specific individual. Based on these results, one may prioritize the variants with a positive query response in follow-up experiments. Although the ClinVar database in this example can be queried without restriction in practice, we envision that an analogous workflow will be useful for querying variants against a growing number of disease-specific databases that are considered more sensitive.

Table 1. Our Approach Leads to Informative and Privacy-Preserving Query Results for Cohort Discovery

Enrollment Criteria	True Count	Differentially Private Count $\epsilon = 0.2$	Differentially Private Count $\epsilon = 0.05$	Proceed with Study? (Count $\geq$ 200)
Age 45–65, non-hispanic white, no history of CHD, not on statins, 10-year CHD risk 5%–20%, and lives in Olmsted County Minnesota	2,000	1,999	1,998	yes
	1,000	998	991	yes
	500	495	444	yes
	250	246	214	yes
	100	98	73	no

Example output of our optimal mechanism for a range of possible true query results in the cohort discovery workflow of Kullo et al. (2016). Our results accurately inform the user whether the available sample size exceeds the desired threshold of 200, while achieving DP. We used uniform prior and asymmetric loss function of Vinterbo et al. (2012) with $\beta_+ = 2$ and $\beta_- = \alpha_+ = \alpha_- = 1$. CHD, coronary heart disease.

DISCUSSION

A key aspect of our mechanisms is that they allow the user to decide their own desired definitions of utility function and prior distribution to obtain a query result that is optimal with respect to their chosen definitions. Although our mechanisms support a broad class of utility functions and prior distributions (enabled by our generalized proof of optimality), we expect the example choices in our experiments to be reasonable options for many scenarios. Specifically, we suggest Vinterbo et al. (2012)'s parametric loss function (Figure 2) and uniform or parametric decay prior (Figure S1) for cohort discovery; and linear or uniform loss function and empirical query distribution as prior for variant lookup (Figure 3). In addition, the user may benefit from an interactive resource that allows them to easily explore the consequence of different choices of these components, similar to the one provided by Vinterbo et al. (2012) for their parametric loss function. It is worth noting that, unlike Vinterbo et al. (2012)'s proposal, our mechanisms allow the user to try out any number of utility and prior distribution without incurring additional privacy cost, due to the fact these components are needed only during local post-processing of the query result.

Given the optimality of our schemes, our results illustrate the theoretical boundaries of leveraging DP for privacy risk mitigation in biomedical query-answering systems. Nevertheless, due to the stringent nature of DP, the smallest possible error probabilities for beacons achieved by our optimal mechanism could still be too high for real-world systems to endure without significantly restricting their use. This potential limitation may be attributed to the fact that the queries submitted to beacons are highly skewed toward rare variants, which are also the most sensitive data from the perspective of DP, thus inflating the overall error probability. Our optimality proofs suggest that in order to surmount this challenge, we will require fundamental shifts in theory to expand the scope of DP to allow more permissive frameworks that remain useful in practical scenarios.

Table 2. Our Approach Leads to Informative and Privacy-Preserving Query Results for Variant Lookup

Query Variant			# of Occurrences in Clinical Database	Differentially Private Lookup Result (ϵ = 0.2)
Genomic Position	Gene	Translation Impact		
chr2:166054637	SCN1A	NA	7	yes*
chr22:40350018	ADSL	missense	5	yes*
chr19:12655693	MAN2B1	NA	4	yes*
chr22:18918421	PRODH	missense	3	yes*
chr6:156871638	ARID1B	stop gain	2	yes*
chr4:6301815	WFS1	missense	2	no
chr22:18918380	PRODH	missense	2	yes*
chr20:5302677	PROKR2	missense	1	no
chr20:5302677	PROKR2	missense	1	no
chr19:13298946	CACNA1A	missense	0	no*
chr6:156778581	ARID1B	missense	0	no*
chr11:6632354	DCHS1	missense	0	no*
chr8:99832630	VPS13B	missense	0	no*
chr11:6632354	DCHS1	missense	0	yes
chr16:29813701	PRRT2	missense	0	yes
chr9:6604646	GLDC	missense	0	no*
chr12:2585485	CACNA1C	missense	0	yes

Example output of our variant lookup mechanism for testing whether the putative autism-associated variants reported by Velmeshev et al. (2019) have previously been labeled pathogenic in the ClinVar database. Our mechanism correctly responds to 11 out of 17 queries (marked by asterisks), notably with increased accuracy for those with higher occurrence in the database. We used Raisaro et al.'s variant query distribution (Raisaro et al., 2017) as prior and linear loss with $\lambda = 2$ (see Figure 4).

Another solution may be hybrid systems that combine DP with traditional access control procedures to allow the sharing of highly sensitive query results based on trust, while securing and facilitating other types of queries that are more amenable to DP. In particular, our DP techniques achieve high accuracy for count query results that are sufficiently large (e.g., greater than 50) and membership queries for which the underlying count is similarly large; both cases will become increasingly common as biomedical databases grow in size.

There are several interesting methodological directions that merit further research. It may be possible to design optimal DP mechanisms for more complex tasks beyond count and membership queries, including emerging applications of DP in federated machine learning (Abadi et al., 2016). We also plan to explore better ways to address the setting where each user submits a potentially large number of queries; although current work considers each query independently, one could exploit the structure of common biomedical queries to develop more effective methods to compose multiple queries, such that the overall privacy cost is smaller than that obtained by general-purpose composition techniques (Kairouz et al., 2017). Lastly, in line with the work of Raisaro et al. (2019), we need further efforts to incorporate DP mechanisms into federated systems that enable collaborative analysis across a group of entities with isolated datasets, who are unable to directly pool the data due to regulatory constraints or conflicting interests (e.g., Cho et al., 2018; Hie et al., 2018). Together, these efforts will help us bring cutting-edge advances in DP to real-world biomedical data-sharing platforms in order to empower researchers while enhancing protection for individuals.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - Review of Differential Privacy
  - Previous Methods for Cohort Discovery and Variant Lookup with Privacy
  - Our Approach: Utility-Maximizing Differential Privacy Mechanisms
  - Implementation Details
  - Optimal Choice of Privacy Parameter for Exponential Mechanisms
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.03.006.

ACKNOWLEDGMENTS

This work is accepted for oral presentation at RECOMB 2020. H.C. is supported by Eric and Wendy Schmidt through the Schmidt Fellows Program at Broad Institute. S.S. is supported through the Stanley Center for Psychiatric Research. When this work began, S.S. and H.C. were partially supported by NIH GM081871 (to B.B.).

AUTHOR CONTRIBUTIONS

All authors conceived the methodology. H.C., S.S., and B.B. performed the theoretical analyses. H.C. and R.K. performed the computational experiments. B.B. guided the research. All authors analyzed results and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 4, 2020
Revised: February 26, 2020
Accepted: March 25, 2020
Published: April 22, 2020

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Aaronson, S., and Rothblum, G.N. (2019). Gentle measurement of quantum states and differential privacy. arXiv https://arxiv.org/abs/1904.08747.

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security 2016, pp. 308–318.

Cho, H., Wu, D.J., and Berger, B. (2018). Secure genome-wide association analysis using multiparty computation. Nat. Biotechnol. *36*, 547–551.

Dankar, F.K., and El Emam, K. (2013). Practicing differential privacy in health care: a review. Trans. Data Privacy *6*, 35–67.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). The reusable holdout: preserving validity in adaptive data analysis. Science *349*, 636–638.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, *vol 3876*, S. Halevi and T. Rabin, eds. (Springer), pp. 265–284.

Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S.O.M., Brookes, A.J., Carey, K., Lloyd, D., Goodhand, P., et al. (2019). Federated discovery and sharing of genomic data using Beacons. Nat. Biotechnol. *37*, 220–224.

Gardner, J., Xiong, L., Xiao, Y., Gao, J., Post, A.R., Jiang, X., and Ohno-Machado, L. (2013). Share: system design and case studies for statistical health information release. J. Am. Med. Inform. Assoc. *20*, 109–116.

Ghosh, A., Roughgarden, T., and Sundararajan, M. (2012). Universally utility-maximizing privacy mechanisms. SIAM J. Comput. *41*, 1673–1693.

Global Alliance for Genomics and Health (2016). GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science *352*, 1278–1280.

Hie, B., Cho, H., and Berger, B. (2018). Realizing private and practical pharmacological collaboration. Science *362*, 347–350.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., and Craig, D.W. (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLoS Genet. *4*, e1000167.

Jiang, X., Sarwate, A.D., and Ohno-Machado, L. (2013). Privacy technology to support data sharing for comparative effectiveness research: a systematic review. Med. Care *51*, S58–S65.

Kairouz, P., Oh, S., and Viswanath, P. (2017). The composition theorem for differential privacy. IEEE Trans. Inform. Theory *63*, 4037–4049.

Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al. (2017). The exac browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. *45*, D840–D845.

Kullo, I.J., Jouni, H., Austin, E.E., Brown, S.A., Kruisselbrink, T.M., Isseh, I.N., Haddad, R.A., Marroush, T.S., Shameer, K., Olson, J.E., et al. (2016). Incorporating a genetic risk score into coronary heart disease risk estimates: effect on low-density lipoprotein cholesterol levels (the MI-GENES clinical trial). Circulation *133*, 1181–1188.

Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: improvements to accessing data. Nucleic Acids Res. *48*, D835–D844.

Lemke, A.A., Wu, J.T., Waudby, C., Pulley, J., Somkin, C.P., and Trinidad, S.B. (2010). Community engagement in biobanking: experiences from the emerge network. Genomics Soc. Policy *6*, 50.

Lowe, H.J., Ferris, T.A., Hernandez, P.M., and Weber, S.C. (2009). STRIDE-an integrated standards-based translational research informatics platform. AMIA Annu. Symp. Proc. *2009*, 391–395.

Machanavajjhala, A., He, X., and Hay, M. (2017). Differential privacy in the wild: A tutorial on current practices & open challenges. Proceedings of the 2017 ACM International Conference on Management of Data 2017, pp. 1727–1730.

McSherry, F., and Talwar, K. (2007). Mechanism design via differential privacy. 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07) 7, pp. 94–103.

Murphy, S.N., and Chueh, H.C. (2002). A security architecture for query tools used to access large biomedical databases. Proc. AMIA Symp. *2002*, 552–556.

Murphy, S.N., Gainer, V., Mendis, M., Churchill, S., and Kohane, I. (2011). Strategies for maintaining patient privacy in i2b2. J. Am. Med. Inform. Assoc. *18*, i103–i108.

Raisaro, J.L., Tramèr, F., Ji, Z., Bu, D., Zhao, Y., Carey, K., Lloyd, D., Sofia, H., Baker, D., Flicek, P., et al. (2017). Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. J. Am. Med. Inform. Assoc. *24*, 799–805.

Raisaro, J.L., Troncoso-Pastoriza, J.R., Misbach, M., Sousa, J.S., Pradervand, S., Missiaglia, E., Michielin, O., Ford, B., and Hubaux, J.P. (2019). Medco: enabling secure and privacy-preserving exploration of distributed clinical and genomic data. IEEE ACM Trans. Comp. Biol. Bioinform. *16*, 1328–1341.

Saeed, M., Villarroel, M., Reisner, A.T., Clifford, G., Lehman, L.W., Moody, G., Heldt, T., Kyaw, T.H., Moody, B., and Mark, R.G. (2011). Multiparameter intelligent monitoring in intensive care ii: a public-access intensive care unit database. Crit. Care Med. *39*, 952–960.

Shringarpure, S.S., and Bustamante, C.D. (2015). Privacy risks from genomic data-sharing beacons. Am. J. Hum. Genet. *97*, 631–646.

Simmons, S., Sahinalp, C., and Berger, B. (2016). Enabling privacy-preserving GWASs in heterogeneous human populations. Cell Systems *3* (1), 54–61.

SPARK Consortium (2018). Spark: a us cohort of 50,000 families to accelerate autism research. Neuron *97*, 488–493.

Uhlerop, C., Slavković, A., and Fienberg, S.E. (2013). Privacy-preserving data sharing for genome-wide association studies. J. Priv. Confid. *5*, 137–166.

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D.H., and Kriegstein, A.R. (2019). Single-cell genomics identifies cell type–specific molecular changes in autism. Science *364*, 685–689.

Vinterbo, S.A., Sarwate, A.D., and Boxwala, A.A. (2012). Protecting count queries in study design. J. Am. Med. Inform. Assoc. *19*, 750–757.

Vu, D., and Slavkovic, A. (2009). Differential privacy for clinical trial data: preliminary evaluations. In IEEE International Conference on Data Mining Workshops 2009 (IEEE), pp. 138–143.

Weber, G.M., Murphy, S.N., McMurry, A.J., Macfadden, D., Nigrin, D.J., Churchill, S., and Kohane, I.S. (2009). The shared health research information network (shrine): a prototype federated query tool for clinical data repositories. J. Am. Med. Inform. Assoc. *16*, 624–630.

Wieland, S.C., Cassa, C.A., Mandl, K.D., and Berger, B. (2008). Revealing the spatial distribution of a disease while preserving privacy. Proc. Natl. Acad. Sci. USA *105*, 17608–17613.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
ExAC database query distribution	Raisaro et al., 2017	Table 2; [https://doi.org/10.1093/jamia/ocw167]
Rare variants from the exome sequencing of autism patients	Velmeshev et al., 2019	Data S5; [https://doi.org/10.1126/science.aav8130]
Software and Algorithms		
A MATLAB implementation of our differential privacy mechanisms	This paper	https://github.com/hhcho/priv-query

LEAD CONTACT AND MATERIALS AVAILABILITY

This study did not generate new materials. Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Bonnie Berger (bab@mit.edu).

METHOD DETAILS

Review of Differential Privacy

Differential privacy (DP) (Dwork et al., 2006, 2015) is a theoretical framework for sharing aggregate-level information about a dataset while limiting the leakage of private information about individuals in the dataset. The idea behind differential privacy is that, if we have a dataset from which we want to release some statistic, it should be difficult to tell the difference between the statistic calculated on our dataset and the statistic calculated on a dataset that differs from ours by exactly one individual. Differential privacy achieves this property by adding a controlled amount of statistical noise to the shared data in such a way that it ensures the statistical impact of any particular individual in the database is smaller than the desired level. Originally developed in cryptography, the theory of differential privacy has found many applications and theoretical connections in diverse domains, including population genetics (Simmons et al., 2016), privacy-preserving machine learning (Abadi et al., 2016), robust statistics (Dwork et al., 2015), and quantum computing (Aaronson and Rothblum, 2019), and has recently been adopted to protect private individuals in real-world systems, e.g., by Google and the United States Census Bureau (Machanavajjhala et al., 2017).

More formally, assume that we have a dataset, denoted X, and we want to release some statistic, denoted $f(X)$, that has been calculated on our data set. This statistic, however, may not preserve the privacy of the individuals in X. As such, we will instead release a perturbed statistic, denoted $F(X)$, which approximates $f(X)$ while still achieving a certain level of privacy. This level of privacy is measured by a parameter ϵ>0, where the closer to zero ϵ is, the more privacy is retained. The goal is that, for every pair (X, X') of "neighboring" datasets (i.e., X and X' are of the same size and differ by exactly one individual), we have that for any possible outcome y in the image of F,

$$P(F(X)=y) \leq \exp(\epsilon) \cdot P(F(X')=y)$$

Any F that satisfies this property is said to be ϵ-differentially private. Intuitively, this property ensures that it is statistically hard to distinguish $F(X)$ from $F(X')$, thereby ensuring that no one individual loses too much privacy when $F(X)$ is released.

One can extend this framework to protect multiple queries using the following composition properties: Given a sequence of k statistics $F=(F_1, \ldots, F_k)$ where each statistic F_i is ϵ_i-differentially private, the overall algorithm F is $\sum_{i=1}^{k} \epsilon_i$-differentially private (sequential composition). When each of the statistics in F is computed based on a disjoint subset of the individuals in the dataset, F is $\max_{i \in \{1,\ldots,k\}} \epsilon_i$-differentially private (parallel composition). More advanced composition techniques with tighter privacy bounds on the overall mechanism have been proposed (Kairouz et al., 2017). These tools enable database owners to assign a privacy "budget" to each individual user and keep track of the combined privacy level ϵ throughout the user's interaction with the database, which reflects how much private information is revealed to the user overall.

Previous Methods for Cohort Discovery and Variant Lookup with Privacy

Key use cases of biomedical query-answering systems include cohort discovery and variant lookup. In cohort discovery, clinical researchers ask how many patients in the medical record database meet a certain condition. This information can help researchers to design studies and assess their feasibility without having to obtain sensitive patient data first. In variant lookup, researchers or physicians ask whether a given genetic variant is observed in the individuals in the database. This query can facilitate the matching of patients with similar genetic determinants of disease across different hospitals, or improve the phenotypic characterization of

genetic variants of interest. Here we review existing proposals for mitigating privacy risks associated with these two types of queries.

Cohort discovery with privacy

Existing cohort discovery systems such as i2b2 and STRIDE (Lowe et al., 2009; Murphy et al., 2011) use approaches to protecting privacy that give no real privacy guarantees, whereby they release perturbed counts for some users instead of raw counts. Both of these methods work by adding truncated Gaussian noise to the query results, which unfortunately do not provide formal privacy guarantees. In order to remedy this issue, Vinterbo et al. (2012) suggested a method to produce differentially private answers to count queries. Although their work was not the first attempt to use differential privacy in a medical research context, it is the first we are aware of to do so as a way to improve medical count queries (Vu and Slavkovic, 2009; Gardner et al., 2013; Dankar and El Emam, 2013; Jiang et al., 2013).

We briefly review the approach of Vinterbo et al. (2012) here. In a nutshell, the authors assume that there is a database consisting of n patients, and they want to know the number of people in that database meeting a certain condition. To accomplish this goal, they introduce a loss function ℓ defined by

$$\ell(x,y) = \begin{cases} \beta_+(y-x)^{\alpha_+} & \text{if } y \geq x, \\ \beta_-(x-y)^{\alpha_-} & \text{otherwise,} \end{cases} \qquad \text{(Equation 1)}$$

where $\alpha_+, \alpha_-, \beta_+, \beta_-$ are parameters given by the user, y is an integer in the range $[r_{min}, r_{max}]$ to be released by the mechanism, and x is the true count. This loss function measures the loss of approximating x with y.

Note that ℓ has sensitivity (i.e., maximum change caused by substituting a single individual) $\Delta\ell = \max(\beta_+, \beta_-)$. Therefore, if we define a random function X_ω such that, for a given value x, $P(X_\omega(x) = y)$ is proportional to $\exp\{-\omega\ell(x,y)\}$ for all integers $y \in [r_{min}, r_{max}]$, then X_ω is $2\Delta\ell\omega$-differentially private. This approach is commonly known as the exponential mechanism (McSherry and Talwar, 2007). Given the above parameters and ϵ, Vinterbo et al.'s mechanism (Vinterbo et al 2012) returns an estimate of x given by $X_\omega(x)$, where $\omega = \epsilon/(2\Delta\ell)$. Note that this result is ϵ-differentially private by the standard analysis of the exponential mechanism (McSherry and Talwar, 2007), though a more thorough analysis can show that this is not a tight bound on ϵ, which we provide in a later section of Method Details.

A recently proposed framework called MedCo (Raisaro et al., 2019) allows multiple hospitals to collectively answer cohort discovery queries without sharing the underlying databases. MedCo uses the Laplace mechanism to achieve ϵ-differential privacy, whereby the system adds noise from the Laplacian distribution with scale parameter $1/\epsilon$ to the true count, rounds it to the nearest integer, then releases the result to the user. Unlike Vinterbo et al.'s approach, this framework does not allow users to choose their own notion of utility; instead, it approximately corresponds to a loss with $\beta_+ = \beta_-$ and $\alpha_+ = \alpha_- = 1$.

Variant lookup with privacy

Shringarpure and Bustamante (2015) were the first to demonstrate a membership inference attack on genomic beacon services, which we refer to as the SB attack. In their attack scenario, it is assumed that the attacker has access to (a subset of) the target individual's genotypes. The attacker repeatedly queries the beacon with the individual's data to infer whether or not the individual is included in the database. More precisely, the likelihood ratio over the query responses between the case where the database includes the individual and the case where it does not is used to statistically test whether the individual is indeed in the database. Depending on how rare the queried variants are, it has been shown that a small number of queries are sufficient for the SB attack to succeed.

Recently, Raisaro et al. (2017) further explored this privacy threat for emerging beacon services, and proposed a set of risk mitigation strategies, which include: (i) only answering yes when the query variant is observed at least twice; (ii) answering incorrectly with probability ϵ for rare variants; and (iii) keeping track of the cumulative likelihood ratio in the SB attack for each individual and suppressing a user once a threshold has been reached. However, these strategies were specifically designed for and analyzed based on the SB attack, which assumes certain properties about the data distribution, e.g., that every variant satisfies Hardy-Weinberg equilibrium. As a result, although Raisaro et al.'s strategies are reasonable first steps, they do not guard against more sophisticated attacks that exploit data patterns not captured by the underlying model (Figure S2). Differential privacy techniques, on the other hand, are agnostic to how the individual-level data is distributed and thus enable more rigorous guarantees of privacy.

Our Approach: Utility-Maximizing Differential Privacy Mechanisms

We introduce differential privacy (DP) mechanisms for cohort discovery and variant lookup problems that achieve provably optimal trade-offs between privacy and utility. Our approaches build upon the truncated α-geometric mechanism for the count query problem (Ghosh et al., 2012), which is well-studied in a theoretical context.

Count queries for cohort discovery

Here, we briefly describe previous work (Ghosh et al., 2012) and then turn to our generalizations of it.

α-geometric mechanism. Let $x \in \{0\} \cup \mathbb{Z}^+$ be the true result of a count query. The α-geometric mechanism (for $\alpha \in (0,1)$) takes x as input and releases a perturbed value $y = x + \Delta \in \mathbb{Z}$, where

$$P(\Delta = \delta) = \frac{1-\alpha}{1+\alpha}\alpha^{|\delta|}.$$

This mechanism achieves $\ln(1/\alpha)$-differential privacy, because changing a single database entry changes x by at most one; the likelihood of any observed value of y differs by at most a multiplicative factor of $1/\alpha = e^{\ln(1/\alpha)}$ for neighboring x and x' that differ by one.

Truncated α-geometric mechanism (α-TGM). Now consider a mechanism that takes the output y from the α-geometric mechanism and truncates it to the range of possible results for a count query given a database of n entries to release $z = \min\{\max\{0,y\},n\}$. This is called the truncated α-geometric mechanism. Due to the post-processing property of differential privacy, which states that any additional computation on differentially private statistics does not cause additional privacy loss, this mechanism is also $\ln(1/\alpha)$-differentially private.

Optimality of α-TGM for count queries with symmetric utility. Let $\ell(x,y)$ be a loss function that captures the disutility (i.e., how undesirable an outcome is to the user) of releasing a result y when the true result of the count query is x. We are interested finding a ϵ-DP mechanism that maximizes (minimizes) the expected utility (disutility) for any given ϵ. Note that we can parameterize the DP mechanism by the conditional probability distribution $q(y|x)$. Formally, the utility-maximization problem is given by

$$\begin{array}{ll}\text{minimize}_{q(y|x)} & \mathbb{E}_{x\sim\pi(x)}\left[\mathbb{E}_{y\sim q(y|x)}[\ell(x,y)]\right] \\ \text{s.t.} & \alpha\cdot q(y|x) \le q(y|x'),\ \forall y,x,x' : |x-x'|=1\end{array}$$

where $\alpha = e^{-\epsilon}$ and $\pi(x)$ denotes the prior belief over the true count $x\in\{0,\ldots,n\}$.

The following theorem summarizes the main result of Ghosh et al. (2012):

Theorem 1. (Ghosh et al., 2012; adapted). *Given any prior belief $\pi(x)$ and a symmetric loss function $\ell(x,y) = f(|x-y|)$ for a monotonically increasing f, a utility-maximizing ϵ-DP mechanism for the count query problem is obtained by applying a (π,ℓ)-dependent post-processing to the output of $\exp(-\epsilon)$-TGM.*

This is a striking result in that the core component of the optimal mechanism (i.e., α-TGM) is agnostic to the user's chosen prior distribution and loss function. This theorem suggests that given a desired privacy level ϵ, the user can obtain a generic query response from the database based on the α-TGM and locally apply post-processing to obtain a tailored response that is provably optimal in terms of utility among all ϵ-DP mechanisms for this task. Furthermore, this scheme allows the user to locally try out different choices of π and ℓ without consuming additional privacy budget (which roughly corresponds to the number of queries a user is permitted). Thus, the theoretical results of Ghosh et al. (2012) immediately provide an improved DP mechanism for cohort discovery compared to existing proposals (Vinterbo et al., 2012; Raisaro et al., 2019).

The optimal post-processing step is given by minimizing the expected loss under the posterior belief over the true count, conditioned on the output of the α-TGM. Formally, let $q_{TGM}(z|x;\alpha,n)$ be the conditional output distribution of α-TGM for a true count x based on a dataset of size n. Conditioned on the value of z, the posterior belief over x can be expressed as:

$$q(x|z;\pi,\alpha,n) \propto \pi(x) q_{TGM}(z|x;\alpha,n).$$

Next, define a map $T : [n] \mapsto [n]$ as

$$T(z;\pi,\ell,\alpha,n) = \operatorname{argmin}_y \sum_x q(x|z;\pi,\alpha,n)\ell(x,y), \qquad \text{(Equation 2)}$$

which represents the loss-minimizing guess for x conditioned on z. Finally, the ϵ-DP query response with maximum expected utility with respect to (π,ℓ) is given by $T(z;\pi,\ell,\exp(-\epsilon),n)$, where z is the output of the $\exp(-\epsilon)$-TGM.

Our contribution: extending the optimality of α-TGM to more general utility functions. A key limitation of Ghosh et al.'s scheme (Ghosh et al., 2012) is that Theorem 1 applies to only loss functions that are symmetric around the true query result. It has been previously noted that asymmetric utility functions are useful in cohort discovery (Vinterbo et al., 2012). For example, overestimating the amount of resources required to perform a clinical study due to overestimating the number of individuals who can be enrolled is more favorable than underestimating it only to encounter a resource bottleneck during the study.

To this end, we newly generalize Theorem 1 to a more general class of utility functions, notably including both symmetric and asymmetric functions, thereby broadly enabling the use of α-TGM for cohort discovery applications. In fact, Ghosh et al. posed the case of asymmetric utility as an open question in their publication (Ghosh et al., 2012), which we resolve in our work. Note that we generalize the utility even further to allow a different choice of utility for every possible value of the true count x, a result that we will return to in the following section on variant lookup. Our generalized theorem is as follows.

Theorem 2. *Given any prior belief $\pi(x)$ and a loss function $\ell(x,y) = f_x(y-x)$, where f_x is quasiconvex and attains its minimum at zero for all x, a utility-maximizing ϵ-DP mechanism for the count query problem is obtained by applying a (π,ℓ)-dependent post-processing to the output of $\exp(-\epsilon)$-TGM.*

The proof is included in the next section. Note that the (π,ℓ)-dependent post-processing step is identical to the scheme we previously described for the symmetric case, except with expanded support for a broader range of loss functions the user can choose from.

Proof of Theorem 2: optimality of truncated α-geometric mechanism for count query with more general loss functions. Here, we answer the open question posed in Ghosh et al. (2012) about the optimality of α-geometric mechanism for asymmetric loss functions in the affirmative. In addition, we further generalize this result to loss functions that are separately defined for each possible value of the true count.

We first provide a sketch of the original proof. Ghosh et al. formulates the problem of finding the optimal differential privacy mechanism as a linear program (LP), whose objective to be minimized is the expected loss of the mechanism with respect to the given prior and loss function, and the constraints ensure that the chosen mechanism is valid (i.e. based on proper probability distributions) and satisfies ϵ-differential privacy. They introduce a combinatorial object called the "signature" Σ of the mechanism, which succinctly

characterizes the set of LP constraints that are bound (tightly satisfied) at the given solution. Ghosh et al. (2012) then fully characterize the signature of optimal solutions, then show that, given a post-processing step (i.e. remap), truncated α-geometric mechanism (α-TGM) can be transformed into a mechanism that has an identical signature as the optimal solution. Finally, they complete the proof of optimality by showing that the set of binding constraints captured by the optimal signature is large enough for the solution to be unique, which implies the equivalence of the optimal solution and α-TGM with optimal remap. We refer the readers to the original paper for details of this proof.

The only step in the original proof that relies on the properties of the loss function is in the following lemma. Thus, it is sufficient to prove that this lemma holds for more general loss functions.

Lemma (Ghosh et al., 2012; Lemma 5.5). *For every user u with a full-support prior and a strictly legal loss function, every optimal direct mechanism for u has a unimodal signature.*

Note that *full-support* prior refers to prior distribution with nonzero probability mass everywhere. Ghosh et al. define (strictly) *legal* loss function $\ell(i,j)$ for input $i\in\{0,\ldots,n\}$ and output $j\in\{0,\ldots,n\}$ as a (strictly) monotone increasing function that depends only on $|i-j|$. A count query mechanism is parametrized by a matrix X whose elements x_{ij} for $i,j\in\{0,\ldots,n\}$ represent the probability of releasing a count j given the true count i. A *direct* mechanism refers to a mechanism for which identity remap is the optimal remap for minimizing expected loss (i.e. it is optimal to take the output "at face value").

Signature Σ of a count query mechanism with privacy parameter α (as in α-TGM) is defined by a n-by-$(n+1)$ matrix, whose element $\sigma_{ij}\in\{\uparrow,\downarrow,S,0\}$ is determined as

$$\sigma_{ij}=\begin{cases}\uparrow & \text{if } x_{ij},x_{(i+1)j}>0 \text{ and } x_{(i+1)j}=\alpha x_{ij},\\ \downarrow & \text{if } x_{ij},x_{(i+1)j}>0 \text{ and } x_{(i+1)j}=x_{ij}/\alpha,\\ S & \text{if } x_{ij},x_{(i+1)j}>0 \text{ and } \alpha x_{ij}<x_{(i+1)j}<x_{ij}/\alpha,\\ 0 & \text{if } x_{ij}=x_{(i+1)j}=0,\end{cases}$$

for $i\in\{0,\ldots,n-1\}$ and $j\in\{0,\ldots,n\}$. A signature is called unimodal, if each row begins with some number of $\downarrow$ entries (possibly none), followed by zero or one S entries, followed by some number of $\uparrow$ entries (possibly none).

Below we prove that the above lemma still holds even if we expand the scope of legal loss functions to those that can be represented using a pair of monotone increasing functions $\ell_{\leq}^{(i)},\ell_{\geq}^{(i)}:\{0\}\cup\mathbb{Z}^{+}\mapsto\mathbb{R}$ for every $i\in\{0,\ldots,n\}$, whose values coincide at zero (i.e., $\ell_{\leq}^{(i)}(0)=\ell_{\geq}^{(i)}(0)$), and

$$\ell(i,j)=\begin{cases}\ell_{\geq}^{(i)}(i-j) & i\geq j,\\ \ell_{\leq}^{(i)}(j-i) & i\leq j.\end{cases}$$

In other words, this expanded class of loss functions include any function that monotonically increases in value as we change the output j in either direction starting from the true count i, without requiring that it is (i) symmetric ($\ell_{\leq}^{(i)}=\ell_{\geq}^{(i)}$) or (ii) identically shaped for different values of i ($\forall i,\ \ell_{\geq}^{(i)}=\ell_{\geq}$ and $\ell_{\leq}^{(i)}=\ell_{\leq}$ for some $\ell_{\geq},\ell_{\leq}$). Note that for the proof of the above lemma it is acceptable to assume strict monotonicity of loss functions; Ghosh et al. presents a limiting argument to discharge this assumption later in their proof, which applies analogously to our generalized loss function.

Lemma. *For every user u with a full-support prior and a loss function who is legal by our generalized definition, every optimal direct mechanism for u has a unimodal signature.*

The proof of the above lemma using our generalized definition of legal loss function is as follows. Given a full-support prior $\pi\succ 0$ and strictly legal loss function $\ell=\{(\ell_{\leq}^{(i)},\ell_{\geq}^{(i)})\}_{i=0}^{n}$ for a database of size n, let X be an optimal direct mechanism with signature Σ. As in the original proof, we want to show that Σ is unimodal. In order to prove this, it suffices to show that there is no row h and columns $k<m$ such that $\sigma_{hk}\in\{S,\uparrow\}$ and $\sigma_{hm}\in\{S,\downarrow\}$.

Assume there exist such h, k, and m. We first show that $k<h<m$ by deriving contradiction from $h\leq k$ or $h\geq m$. Assume $h\leq k$ (the $m\leq h$ case follows an analogous argument). The facts that $\sigma_{hk}\in\{S,\uparrow\}$ and $\sigma_{hm}\in\{S,\downarrow\}$ imply the following inequalities:

$$\alpha x_{hk}<x_{(h+1)k}\leq\frac{x_{hk}}{\alpha}\ \text{and}\ \alpha x_{hm}\leq x_{(h+1)m}<\frac{x_{hm}}{\alpha}.$$

Following the same argument given in the original proof, this means that we can find a small $\lambda>0$ such that defining a revised mechanism X' where $x'_{im}=(1-\lambda)x_{im}$ and $x'_{ik}=x_{ik}+\lambda x_{im}$ for all $i\in\{0,1,\ldots,h\}$, while keeping the remaining entries the same as X, results in a feasible mechanism for the given problem. Note that the difference in expected loss between the two mechanisms is given as

$$L(X')-L(X)=\sum_{i=0}^{h}\lambda\pi_i x_{im}(\ell(i,k)-\ell(i,m))$$

Given that $h\leq k$ and the initial assumption $k<m$, we have

$$\ell(i,k)-\ell(i,m)=\ell_{\leq}^{(i)}(k-i)-\ell_{\leq}^{(i)}(m-i)<0$$

for all $0\leq i\leq h$, using the strict monotonicity of $\ell_{\leq}^{(i)}$. Thus, X' achieves a strictly lower expected loss than X (given $\lambda>0$ and $\pi\succ 0$), which is a contradiction. An analogous line of reasoning, which instead involves moving the probability mass in rows $i\in\{h+1,\ldots,n\}$ of X and the monotonicity of $\ell_{\geq}^{(i)}$, shows contradiction for $h\geq m$. Thus, $k<h<m$.

Another implication of the above result is that for any tuple (h', k', m') where $k'<m'$, and $h' \leq k'$ or $h' \geq m'$, either $\sigma_{h'k'} = \downarrow$ or $\sigma_{h'm'} = \uparrow$. We now use this fact to derive contradiction for the case where $k<h<m$. Our assumption states that $\sigma_{hk} \in \{S, \uparrow\}$ and $\sigma_{hm} \in \{S, \downarrow\}$. Applying the above logic to tuples (h, k, h) and (h, h, m), we obtain that $\sigma_{hh} = \uparrow$ from the former tuple, and $\sigma_{hh} = \downarrow$ from the latter tuple, which together gives a contradiction. Thus, all possible values of h result in a contradiction, proving the original statement of the lemma. ∎

Membership queries for variant lookup

We newly show that α-TGM can be used to obtain an optimal DP mechanism for the variant lookup problem.

Our optimal differential privacy mechanism for variant lookup based on α-TGM. Formally, the problem can be described as follows. Let X be a dataset of n individuals represented as $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$. Given a user-provided predicate $q : \mathcal{X} \mapsto \{0, 1\}$, we define membership query as the task of calculating

$$f_q(X) = \mathbf{1}\left\{\sum_{i=1}^{n} q(x_i) > 0\right\},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Our goal is to answer queries of this form while preserving the privacy of individuals. Here we will restrict our attention to stochastic mechanisms that consider the total count $c = \sum_{i=1}^{n} q(x_i)$ as input and probabilistically output a binary answer for the query. However, analogous to the results of Ghosh et al. (2012) for count query, it can be shown this restriction is without loss of generality; i.e., our mechanism is optimal even among mechanisms that may depend on more fine-grain data patterns other than the count.

Akin to the count query setting, in order to quantify the usefulness of different DP mechanisms for membership queries, we introduce a loss function $\ell(c, b)$, representing how undesirable it is to output an answer $b \in \{0, 1\}$ given a true count $c \in \{0, \ldots, n\}$, as well as a prior belief $\pi(c)$ over the true count.

Our main result. The following theorem summarizes our main result that the optimal DP mechanism for membership queries, which achieves the minimum expected loss with respect to a particular choice of π and ℓ, is obtained by a generic application of α-TGM followed by a local post-processing step by the user; only the latter step depends on π and ℓ.

Theorem 3. *Given any prior belief $\pi(c)$ and a loss function $\ell(c, b)$ satisfying $\ell(0, 0) \leq \ell(0, 1)$ and $\ell(c, 0) \geq \ell(c, 1)$ for $c > 0$, a utility-maximizing ϵ-DP mechanism for the membership query problem is obtained by applying a (π, ℓ)-dependent post-processing step to the output of $\exp(-\epsilon)$-TGM.*

The proof is provided in the next section. The optimal post-processing step of the above theorem proceeds as follows. Given a specific choice of π and ℓ, the user first transforms ℓ into the corresponding loss function ℓ' in the count query setting as

$$\ell'(c, y) = \begin{cases} \ell(c, 0) & \text{if } y = 0, \\ \ell(c, 1) & \text{if } y > 0, \end{cases} \qquad \text{(Equation 3)}$$

for $c, y \in \{0, \ldots, n\}$. Let z be the output of $\exp(-\epsilon)$-TGM returned by the database. The user uses (π, ℓ') to obtain the loss-minimizing guess for the count, which is given by $T(z; \pi, \ell', \exp(-\epsilon), n)$ (see Equation 2). Finally, the user thresholds this number to obtain a binary answer for the membership query, given by $\mathbf{1}\{T(z; \pi, \ell', \exp(-\epsilon), n) > 0\}$. Note that the application of α-TGM using our transformed loss ℓ' is known to be optimal only given our generalized notion of utility we achieved in the previous section (Theorem 2); $\ell'(c, y)$ cannot be expressed as a monotonically increasing $f(|c - y|)$ as required by Theorem 1, nor is it sufficient to drop only the symmetry assumption and consider quasiconvex $f(c - y)$ with minimum at zero—we require the flexibility to set a different loss function $f_c(c - y)$ for each value of c.

Proof of Theorem 3: optimality of truncated α-geometric mechanism for membership query. Let $c \in \{0, \ldots, n\}$ be the number of individuals in a database matching the predicate of a given membership query (e.g. presence of a genetic variant). Let $\pi(c)$ be a user-defined prior belief over c, and $\ell(c, b)$ be a user-defined loss function representing the disutility of receiving a membership query result $b \in \{0, 1\}$ when the true count is c. As stated in the theorem, assume ℓ satisfies $\ell(0, 0) \leq \ell(0, 1)$ and $\ell(c, 0) \geq \ell(c, 1)$ for $c > 0$.

The expected loss minimization problem for ϵ-differentially private membership query, parameterized by a conditional probability distribution $q(b|c)$, can be expressed as follows.

$$\begin{aligned} \text{minimize}_{q(b|c)} \quad & \mathbb{E}_{c \sim \pi(c)}\left[\mathbb{E}_{b \sim q(b|c)}[\ell(c, b)]\right] \\ \text{s.t.} \quad & q(b|c) \leq e^{\epsilon} \cdot q(b|c'), \ \forall b, c, c' : |c - c'| = 1 \end{aligned} \qquad \text{(Equation 4)}$$

We want to show that an optimal solution $q(b|c)$ for the above problem is given by the $\exp(-\epsilon)$-TGM with a (π, ℓ)-dependent post-processing.

Now consider a transformed loss function ℓ', defined as

$$\ell'(c, y) = \begin{cases} \ell(c, 0) & \text{if } y = 0, \\ \ell(c, 1) & \text{if } y > 0, \end{cases}$$

for $c, y \in \{0, \ldots, n\}$. First, note that this loss function satisfies the conditions of Theorem 2; for each value of c, we have that $\ell'^{(i)}(y - i) := \ell'(i, y)$ is monotone increasing in both directions from zero, given our initial assumptions about ℓ. Therefore, we can invoke Theorem 2 using (π, ℓ') to obtain that there is a (π, ℓ')-dependent post-processing T such that, given the output z of a differentially private release of the true count c based on the $\exp(-\epsilon)$-TGM, $T(z)$ represents the output of an optimal ϵ-DP scheme that is an optimal solution to the following count query problem.

$$\begin{array}{ll}\text{minimize}_{q'(y|c)} & \mathbb{E}_{c\sim\pi(c)}\left[\mathbb{E}_{y\sim q'(y|c)}[\ell'(c,y)]\right] \\ \text{s.t.} & q'(y|c)\leq e^{\epsilon}\cdot q'(y|c'),\ \forall y,c,c' : |c-c'|=1\end{array} \quad \text{(Equation 5)}$$

We next prove that the problems Equation 4 and 5 are in fact interchangeable. First, note that any feasible solution q' of Equation 5, can be mapped to a feasible solution q of Equation 4 by setting

$$q(0|c) := q'(0|c)$$
$$q(1|c) := \sum_{y=1}^{n} q'(y|c),$$

which is equivalent to post-processing y as $b = \mathbf{1}\{y>0\}$. If y is ϵ-DP, then we have that b is also ϵ-DP, which implies feasibility of the mapped q. Furthermore, because $\ell'(c,y) = \ell(c,1)$ for all $y>0$, we have that

$$\mathbb{E}_{b\sim q(b|c)}[\ell(c,b)] = \mathbb{E}_{y\sim q'(y|c)}[\ell'(c,y)] \quad \text{(Equation 6)}$$

for our construction of q based on q'. Therefore, for any feasible solution of Equation 5, a corresponding solution exists for Equation 4 that achieves the same objective value.

Next, we consider the reverse direction. Given any feasible solution q of Equation 4, let q' be a solution for Equation 5 constructed as follows:

$$q'(0|c) := q(0|c),$$
$$q'(y|c) := \frac{1}{n} q(1|c),\ \forall y\in\{1,\ldots,n\}.$$

This is akin to post-processing b by setting $y=0$ if $b = 0$, and $y \sim \text{Unif}(\{1,\ldots,n\})$ if $b = 1$. Therefore, if b is ϵ-DP then y is also ϵ-DP, proving that the above q' is a feasible solution for Equation 5. Because $q(1|c) = \sum_{y=1}^{n} q'(y|c)$ by construction, Equation 6 still holds. This proves that for any feasible solution for Equation 4, there exists a feasible solution for Equation 5 with the same objective value.

Now putting the two directions together, we have that an optimal solution for Equation 5, represented by $T(z)$, can be post-processed as $\mathbf{1}\{T(z)>0\}$ to obtain an optimal solution for Equation 4. This is because, if we assume a better solution for Equation 4 exists, then we can map it to Equation 5 to obtain a better solution than $T(z)$, which is a contradiction. Note that $\mathbf{1}\{T(z)>0\}$ can be viewed as a post-processing of the $exp(-\epsilon)$-TGM output z, where T depends on the user-provided π and ℓ. This concludes the proof of the theorem. ∎

Implementation Details

Although our proposed mechanisms incur a negligible computational overhead on the database server (random sampling of one additional number per query from a geometric distribution), the local post-processing by the user naïvely takes $O(n^2)$ for a database of size n (see Equation 2), which can be burdensome for large n. Fortunately, both our use cases (cohort discovery and variant lookup) are based on loss functions with a special structure that allows more efficient computation.

First, in cohort discovery workflows, we are primarily interested in a loss function that can be expressed in terms of the difference between the true count and the perturbed count. In other words, the summation term of Equation 2 can be expressed as a convolution operation between the two functions q and ℓ, as follows:

$$\sum_{x} q(x|z;\pi,\alpha,n)\ell(y-x).$$

Since both x and y are discrete, we can use the standard fast Fourier transform (FFT)-based convolution algorithm to compute this expression for all values of y in $O(n\log(n))$ time, a significant reduction from $O(n^2)$.

Second, in the case of variant lookup, Equation 2 uses a transformed loss function which is identical for any $y>0$ (Equation 3). Thus, the $O(n)$-time summation over x need to be performed for only two values of y (0 and 1), resulting in an overall complexity of $O(n)$ instead of $O(n^2)$.

Optimal Choice of Privacy Parameter for Exponential Mechanisms

In this section, we show that the standard application of exponential mechanism to count queries is suboptimal in terms of the privacy-utility trade-off and provide our modified algorithm for choosing the optimal privacy parameter. Recall that, given a loss function $\ell(x,y)$ for true count x and perturbed result y, exponential mechanism samples $y = X_\omega(x)$ according to probability proportional to $exp\{-\omega\ell(x,y)\}$ with privacy parameter ω. Standard analysis shows that this mechanism is $2\Delta\ell\omega$-differentially private, where $\Delta\ell$ denotes the sensitivity of the function ℓ. In the analysis below, we consider the loss function defined in Equation 1 introduced by Vinterbo et al. (2012).

Here, we follow the intuition that the larger the privacy parameter ω in the exponential mechanism, the more accuracy is achieved. The question then becomes whether we can choose a larger ω than that given by the standard analysis of the exponential mechanism while still guaranteeing privacy. Specifically, we search for a larger ω that still ensures ϵ-differential privacy. From the standard

exponential mechanism, we know that if $\omega \leq \epsilon/(2\Delta\ell)$, then X_ω is ϵ-differentially private. Moreover, it is easy to check that X_ω is not ϵ-differentially private if $\omega \geq \epsilon/\Delta\ell$. Therefore, we perform a binary search on the interval $\left[\frac{\epsilon}{2\Delta\ell}, \frac{\epsilon}{\Delta\ell}\right]$ to find the largest ω such that X_ω is ϵ-differentially private.

Formally, we choose a parameter k (which dictates how long the binary search continues) and use a binary search to find the largest ω in the set

$$\left\{\frac{\epsilon}{2\Delta\ell}, \frac{(k+1)\epsilon}{2k\Delta\ell}, \ldots, \frac{(2k-1)\epsilon}{2k\Delta\ell}, \frac{\epsilon}{\Delta\ell}\right\},$$

such that X_ω is ϵ-differentially private.

This procedure would requires us to check, for a given ω, if X_ω is ϵ-differentially private. Naïvely, this means checking

$$P(X_\omega(x) = y) \leq \exp(\epsilon) P(X_\omega(x+1) = y)$$

and

$$P(X_\omega(x) = y) \leq \exp(\epsilon) P(X_\omega(x-1) = y)$$

for each y and x.

However, this naïve method suffers from overwhelming computational burden. Given the range of possible outcomes of our exponential algorithm, $y \in [r_{min}, r_{max}]$, we are able to show that it suffices to check

$$P(X_\omega(r_{min}) = r_{min}) \leq \exp(\epsilon) P(X_\omega(r_{min}+1) = r_{min})$$

and

$$P(X_\omega(r_{max}) = r_{max}) \leq \exp(\epsilon) P(X_\omega(r_{max}-1) = r_{max}).$$

This result allows us to greatly speed up the computation, thereby making our method run in minimal time.

Algorithm 1 Our optimized exponential mechanism for count queries

Require: $x, \epsilon, \alpha_-, \alpha_+, \beta_+, \beta_-, r_{\max}, r_{\min}, k$

Ensure: ϵ-differential privacy

$$S = \left\{\left(1 + \frac{i}{k}\right)\frac{\epsilon}{2\Delta\ell}\right\}_{i \in \{0,\ldots,k\}} = \{\omega_0, \ldots, \omega_k\}$$

for $i = 0, \ldots, k$ **do**

$X_i = X_{\omega_i}$

end for

For $i = 0, \ldots, k$ **do**

$$a_i = \frac{P(X_i(r_{max}) = r_{max})}{P(X_i(r_{max}-1) = r_{max})}$$

$$b_i = \frac{P(X_i(r_{min}) = r_{min})}{P(X_i(r_{min}+1) = r_{min})}$$

$$c_i = \max\{a_i, b_i\}$$

end for

Let i_0 be the largest i so that $c_i \leq \exp(\epsilon)$ (found by binary search)

return $X_{i_0}(x)$

To prove this algorithm is differentially private, we first need to prove the following:

Theorem 4. *Let* $\alpha_+, \alpha_- \in [0, 1]$, $\beta_+ > 0$, $\beta_- > 0$ *and*

$$\ell(x, y) = \begin{cases} \beta_+ (y - x)^{\alpha_+} & \text{if } y \geq x, \\ \beta_- (x - y)^{\alpha_-} & \text{otherwise.} \end{cases}$$

Let $X_\omega(x)$ *be a random variable defined such that* $P(X_\omega(x) = y)$ *is proportional to* $\exp(-\omega\ell(x, y))$ *for all integers* $x \in [r_{\min}, r_{\max}]$ *and* $y \in [0, n]$. *Then we have that, if*

1. $\omega \leq \min\left\{\frac{\epsilon}{b_+}, \frac{\epsilon}{b_-}\right\} = \frac{\epsilon}{\Delta\ell}$,
2. $P(X_\omega(r_{\max}) = r_{\max}) \leq \exp(\epsilon) P(X_\omega(r_{\max} - 1) = r_{\max})$, *and*
3. $P(X_\omega(r_{\min}) = r_{\min}) \leq \exp(\epsilon) P(X_\omega(r_{\min} + 1) = r_{\min})$,

then $X_\omega(x)$ is ϵ-differentially private.

Proof of Theorem 4. Let $Z_x = \sum_{y=r_{\min}}^{r_{\max}} \exp(-\omega \ell(x,y))$. By definition, we have

$$P(X_\omega(x)=y) = \frac{\exp(-\omega \ell(x,y))}{Z_x}.$$

For a given y, we are interested in $\frac{P(X_\omega(x)=y)}{P(X_\omega(x')=y)}$ and $\frac{P(X_\omega(x')=y)}{P(X_\omega(x)=y)}$ where $x' = x+1$. There are three cases to consider: (i) $x \geq r_{max}$, (ii) $x < r_{\min}$, and (iii) $x \in [r_{\min}, r_{\max}-1]$.

First, let $x < r_{\min}$. Then $Z_x < Z_{x+1}$, which implies

$$\frac{P(X_\omega(x+1)=y)}{P(X_\omega(x)=y)} = \exp(-\omega\beta_+((y-x-1)^{\alpha_+} - (y-x)^{\alpha_+}))\frac{Z_x}{Z_{x+1}}$$

$$\leq \exp(\omega\beta_+((y-x)^{\alpha_+} - (y-x-1)^{\alpha_+})).$$

Since $\alpha_+ \leq 1$, $y-x > y-x-1 \geq 0$, we see that $(y-x)^{\alpha_+} - (y-x-1)^{\alpha_+} \leq 1$. Thus, the above ratio is at most $\exp(\omega\beta_+) \leq \exp(\epsilon)$ as desired. On the other hand, consider

$$\frac{P(X_\omega(x)=y)}{P(X_\omega(x+1)=y)} = \exp(-\omega\beta_+((y-x)^{\alpha_+} - (y-x-1)^{\alpha_+}))\frac{Z_{x+1}}{Z_x}.$$

Note that $\exp(-\omega\beta_+((y-x)^{\alpha_+} - (y-x-1)^{\alpha_+})) \leq 1$, thus the above ratio is less than Z_x/Z_{x+1}. Since ℓ has sensitivity bounded by $\max\{b_-, b_+\}$, we have $Z_x/Z_{x+1} \leq \exp(\epsilon)$, which gives us

$$\frac{P(X_\omega(x)=y)}{P(X_\omega(x+1)=y)} \leq \exp(\epsilon).$$

Following an analogous argument, when $x \geq r_{\max}$ we have that $\frac{P(X_\omega(x)=y)}{P(X_\omega(x+1)=y)} \leq \exp(\epsilon)$ and $\frac{P(X_\omega(x+1)=y)}{P(X_\omega(x)=y)} \leq \exp(\epsilon)$.

Finally, we consider the case when $r_{\min} \leq x < r_{\max}$. Note that

$$Z_{x+1} = Z_x - \exp(-\omega\ell(x, r_{\max})) + \exp(-\omega\ell(x+1, r_{\min}))$$

$$= Z_x - \exp(-\omega\beta_+(r_{\max}-x)^{\alpha_+}) + \exp(-\omega\beta_-(x-r_{\min}+1)^{\alpha_-}).$$

If $y \geq x$, then $\ell(x,y) > \ell(x+1,y)$. In this case, given $\omega \leq \frac{\epsilon}{\Delta\ell}$, we get

$$\frac{P(X_\omega(x)=y)}{P(X_\omega(x+1)=y)} = \exp(\omega(\ell(x+1,y) - \ell(x,y)))\frac{Z_{x+1}}{Z_x}$$

$$\leq \frac{Z_{x+1}}{Z_x} \leq \exp(\epsilon).$$

On the other hand, if $y > x$, then $\exp(\omega(\ell(x+1, y) - \ell(x, y))) \leq \exp(\omega\beta_+)$. The equality holds when $y = x+1$. Note that $\exp(-\omega\beta_+(r_{\max}-x)^{\alpha_+})$ is increasing in x, while $\exp(-\omega\beta_-(x-r_{\min}+1)^{\alpha_-})$ is decreasing in x. Thus, letting

$$L(x) = -\exp(-\omega\beta_+(r_{\max}-x)^{\alpha_+}) + \exp(-\omega\beta_-(x-r_{\min}+1)^{\alpha_-}),$$

there exists $x_0 \in (r_{\min}, r_{\max})$, where $L(x) \leq 0$ if $x \geq x_0$ and $L(x) > 0$ if $x < x_0$.

Thus, if $x \geq x_0$, we have $\frac{Z_{x+1}}{Z_x} \leq 1$, which implies

$$\frac{P(X_\omega(x)=y)}{P(X_\omega(x+1)=y)} \leq \exp(\omega\beta_+)\frac{Z_{x+1}}{Z_x} \leq \exp(\omega\beta_+) \leq \exp(\epsilon).$$

If $x < x_0$, then $Z_x < Z_{x+1}$, so $Z_x \geq Z_{r_{min}}$. Therefore,

$$\frac{Z_{x+1}}{Z_x} = 1 + \frac{-\exp(-\omega\beta_+(r_{\max}-x)^{\alpha_+}) + \exp(-\omega\beta_-(x-r_{\min}+1)^{\alpha_-})}{Z_x}$$

$$\leq 1 + \frac{-\exp(-\omega\beta_+(r_{\max}-r_{\min})^{\alpha_+}) + \exp(-\omega\beta_-(r_{\min}-r_{\min}+1)^{\alpha_-})}{Z_{r_{\min}}}$$

$$= \frac{P(X_\omega(r_{\min})=r_{\min})}{P(X_\omega(r_{\min}+1)=r_{\min})} \leq \exp(\epsilon),$$

which completes the proof that $P(X_\omega(x)=y) \leq \exp(\epsilon)P(X_\omega(x+1)=y)$ for all x and y. A symmetric argument shows that $P(X_\omega(x+1)=y) \leq \exp(\epsilon)P(X_\omega(x)=y)$ for all x and y. Thus, X_ω is ϵ-differentially private. ∎

Corollary 4.1. *Our modified exponential mechanism for count queries (Algorithm 1) is ϵ-differentially private.* (A Python implementation of this method is available at: https://github.com/seanken/DP_count.)

Furthermore, if we fix k and let ω_ϵ be the ω parameter chosen by Algorithm 1, then because $\omega_\epsilon \geq \frac{\epsilon}{2\Delta\ell}$, we have the following corollary which shows that our optimization improves the utility of the exponential mechanism.

Corollary 4.2. *For any given ϵ and c,*

$$P(\ell(x, X_{\omega_\epsilon}(x)) \leq c) \geq P\left(\ell\left(x, X_{\frac{\epsilon}{2\Delta\ell}}(x)\right) \leq c\right).$$

DATA AND CODE AVAILABILITY

A MATLAB implementation of our DP mechanisms and scripts for reproducing our results are available at: https://github.com/hhcho/priv-query.