

PROTECTING PATIENT DATA CONFIDENTIALITY USING DIFFERENTIAL
PRIVACY

By

Denny Guang-Yeu Lee

A CAPSTONE THESIS

Presented to the Department of Medical Informatics and Clinical
Epidemiology

and the Oregon Health & Science University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Biomedical Informatics

December 2008

School of Medicine
Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Capstone Project of

Denny Guang-Yeu Lee

PROTECTING PATIENT DATA CONFIDENTIALITY USING DIFFERENTIAL
PRIVACY

Has been approved

Judith R. Logan, MD, MS

Table of Contents

Acknowledgements	5
Abstract	6
Introduction	8
Background	10
Importance of Privacy for Healthcare Research	10
Privacy Concerns and Policies	11
Privacy Techniques	13
Auditing.....	13
Suppression of data with limited participation	13
Limit the number of queries to a dataset	14
Privacy Algorithms	15
Protecting Privacy by k-anonymity	16
Protecting Privacy by Differential Privacy	18
Defining Privacy and Differential Privacy.....	18
Revisiting the Governor William Weld case.....	20
Acceptability of the effects of privacy algorithms	21
Privacy-Integrated Queries (PINQ)	23
The Application of PINQ to Healthcare.....	30
Methods	31

Analysis Example.....	33
Results	35
Higher values (n), differences statistically significant.....	35
Lower n with no statistical significance	40
Higher n with no statistical significance	44
Discussion	46
Conclusion	49
Works Cited	51
Appendix	58
Appendix 1: Differential Privacy concept	58
Appendix 2: Application of Differential Privacy Case Study	68
Appendix 3: C# Code Example	72

Acknowledgements

I want to thank Dr. Judith Logan for her guidance and patience for my Capstone thesis. As well, I would like to thank Cynthia Dwork and Frank McSherry for their continued mentoring on the mathematics and mechanisms of privacy.

Abstract

From the breakthrough research of Latanya Sweeney we have learned that the secondary use of healthcare data may be privacy-revealing. Common techniques for ensuring privacy focus more on security and validation and are vulnerable to privacy attacks. Yet, it is only through the secondary use of healthcare data that we can ultimately achieve Elias Zerhouni's vision to *"Transform Medicine from Curative to Preemptive"*. After all, it is the analysis and sharing of data that is crucial to understanding patterns within the healthcare population. But we cannot do this unless we can ensure the privacy of the individual patient.

In this paper we review some of the common methods used to ensure privacy, including the use of privacy algorithms, and emphasize the use of differential privacy algorithms and the application of differential privacy via Privacy Integrated Queries (PINQ). Differential privacy protects the patient by adding exponentially distributed random noise to the results of a query against a data set. Exponentially distributed random noise has some interesting properties that provide privacy guarantees. Within the framework of PINQ, one can apply the differential privacy algorithm, specify the amount of accuracy (epsilon) desired, and PINQ translates this to the units of privacy that it can guarantee (i.e. providing risk enforcement). The concern, especially within healthcare datasets which are typically small in size, is that the additional of random noise, while providing privacy guarantees, will significantly reduce statistical accuracy.

By applying differential privacy using PINQ against a healthcare dataset, we were able to successfully alleviate this concern. By analyzing historical data, we narrowed down the range of candidate epsilon values. We then replicated the statistical tests performed prior to perturbation at the bounds of the 95% confidence interval to determine the ideal epsilon value. It is important that values with higher sample sizes have a lower epsilon value so that we could

in turn apply a higher epsilon value to the values with smaller sample sizes. Altogether this allows us to find and publish the ideal epsilon value that ensures statistical accuracy while providing privacy guarantees.

Introduction

Based on an analysis of the 1990 United States Census, 87% of the United States population is uniquely identifiable by the three attributes of zip code, date of birth, and gender (Sweeney, k-anonymity: a model for protecting privacy). Latanya Sweeney's breakthrough research clearly depicts that the de-identification, masking, or hiding of publicly available information does not adequately protect the privacy of an individual. Sweeney obtained masked medical data from the Group Insurance Commission of Massachusetts responsible for state employees' health insurance. The masked medical data contained only what was considered non-identifiable information such as ethnicity, visit date, diagnosis, procedure, medication, total charge, zip code, date of birth, and gender. This masked medical data set was compiled following the Limited Data Set directive as noted within the Health Information Portability and Accountability Act or HIPAA (United States Department of Health and Human Services). For a cost of \$20, Sweeney then obtained the publicly available Cambridge, MA voter list. The voter list provided personal information but was limited only to voting information, address, date of birth, and gender. Illustrated in the Venn diagram in Figure 1, each of the two data sources by itself revealed little information.

Recall that 87% of the US population is uniquely identifiable by only the three attributes of zip code, date of birth, and gender. In this case, there were only six people in Cambridge, MA who had the same date of birth as Massachusetts Governor William Weld and only three of those were men. Of those men, the Governor was the only person to live in his particular zip code. Through these two pieces of information joined only by gender, date of birth, and zip code, Sweeney was able to identify and reveal the medical records of the Governor.

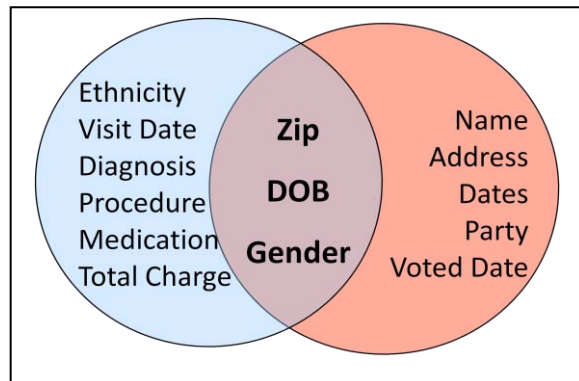


Figure 1: Patient Re-identification via the three attributes. Medical data is on the left, voting records

Due to the importance of patient privacy, this paper will study the use of Privacy Integrated Queries (PIQ) which applies a differential privacy algorithm to a data set. While there are unique perturbations of the data with this technique, our goal is to demonstrate its efficacy and build an initial methodology to use differential privacy (via PIQ) on healthcare datasets to allow analysts to have secondary use of the data and protect the subjects at the same time. This research project will serve as the capstone project for Mr. Denny Lee in the Master of Biomedical Informatics program at OHSU and has been approved by that institution's Institutional Review Board.

Background

[Importance of Privacy for Healthcare Research](#)

In addition to better documentation, accountability, and clinical operational benefits, a key benefit of the use of electronic medical records is to lower the barriers to the secondary use of data that is collected and primarily used in clinical settings, including its use for research purposes. However, privacy concerns remain large. Consider that over 80% of the US consumers will cite privacy concerns as a major reason why they are slow to adopt the use of a personal health record (Markle Foundation). After all, while sharing of medical data is beneficial, the rights of individual to privacy protection must not be violated (Buckovich, Rippen and Rozen; Korn). Unless we can ensure privacy for the individual patient, we will not have the necessary infrastructure to analyze and understand the data.

The vision of the director of the National Institutes of Health, Dr. Elias A. Zerhouni, for the future paradigm of medicine is to *“Transform Medicine from Curative to Preemptive”* (Zerhouni). It is apparent that to fulfill this vision of healthcare prediction, we will need the ability for secondary use of data. It will become necessary to apply advanced techniques to mine patient health records so that we can find patterns in the population. For example, by analyzing many patient health records it was possible to determine that mean arterial pressure provides a strong indicator of stroke (Fogelholm, Avikainen and K) while pulse pressure forewarns of a heart attack (Vaccarino, Holford and Krumholz). If we lose the ability for secondary use of data, we will no longer be able to make these discoveries that can help us predict and prevent health issues from occurring.

We cannot stress enough the importance of patient privacy, as we cannot benefit from the analysis of patient data unless patients believe their privacy is respected and institutions trust

that their data is secure. While we are doing more to ensure that our data is secure, this effort has not been necessarily translated to privacy. At the same time, it should not be presumed that secondary use of data and electronic medical records automatically means that one's privacy is revealed. There is anecdotal evidence that suggests that electronic medical records may still offer better protection than traditional paper records due to features such as auditing trails, the ability to filter data, and electronically sign documents to ensure consistency (Barrows and Clayton). For example, the [Cancer Bioinformatics Grid](#) (caBIG) from the National Cancer Institute has developed a grid of servers and institutions across the country that allow the sharing of key cancer research between institutions so researchers can more easily leverage each other's work. They have even provided guidance on how to build a more secure environment by way of their project (Langella, Hasgints and Oster). It can be seen that most of the focus on privacy and security occur in high profile cases such as the Kaiser Permanente Internet Patient portal security breach (Collmann and Cooper). While this focus on securely providing secondary use of data is required and admirable, as noted in the Governor Weld case security does not necessarily mean privacy.

[Privacy Concerns and Policies](#)

Over the last few years, there has been a concerted effort to provide the policies to address patient's privacy concerns. There are now legal definitions of privacy (Nissenbaum) and regulations such as those resulting from the Health Insurance Portability and Accountability Act of 1996 (HIPAA) provide very explicit instructions on what information can and cannot be revealed in different scenarios (US Department of Health and Human Services (Office for Civil Rights)). Businesses such as Microsoft® provide detailed statements (e.g., "Microsoft Online Privacy Statement") about the personal information collected, uses of that information, and offering users the choice to not share/expose their data for secondary uses (MSN). Academic

institutions, such as Oregon Health & Science University, have requirements in addition to HIPAA standards stating the permitted uses and disclosures associated with protected health information (OHSU Healthcare System (Administrative Policy Manual)). Yet, as seen from the Governor Weld case, the masking of data in compliance with the HIPAA Limited Data Set rules and explicit statements of use and disclosures may not necessarily insure privacy. After all, HIPAA may provide the policies to protect the privacy of medical records but it wasn't designed to facilitate or limit the secondary use of medical data. This is a stark reminder that there should be additional policies or mechanisms to reduce threats to privacy and confidentiality breaches (Annas).

While there are many privacy policy frameworks, none provide actual techniques or methodologies to ensure privacy. Currently, there is a lack of coherent policies and standard "good practices" for the desired secondary use of data (Safran, Bloomrosen and Hammond), i.e. we lack the actual information on how to properly implement the proposed frameworks. For example, if you refer to the [HIMSS Privacy and Security Toolkit](#) (HIMSS), you will notice a lack of in-depth guidance on how to ensure privacy. The Markle Foundation has provided a verbose *Privacy Notice to Consumers* as part of their overall *Common Framework for Networked Personal Health Information* (Markle Foundation) and the *Model Privacy Policies and Procedures for Health Information Exchange* within the same framework (Markle Foundation). But again, while we may now have effusive policies and models on privacy, there is little discussion of how to actually implement it. One of the more complete books on the topic of privacy is that of JC Cannon's *What Developers and IT Professionals Should Know* (Cannon) but it is very much information technology (IT) centric and will require a very good healthcare IT consultant and staff to help bridge the gap between the IT techniques described and their application to healthcare.

Privacy Techniques

While there is a lack of standard best practices for privacy, this has not prevented individuals and institutions from attempting to put together their own privacy methodologies. From an informal survey of privacy advocates and privacy officers, a theme has emerged on the types of privacy techniques that are commonly employed, including auditing, suppression of data with limited participation, limiting the number of queries to a dataset, and application of privacy algorithms.

[Auditing](#)

An audit trail ensures that all users and actions to patient data are recorded. While auditing medical data is important, it is not a technique that prevents confidentiality breaches. It is important to ensure accountability with an audit trail, as it may act as a deterrent for persons with malicious intent or who are just snooping. But auditing data only provides the ability to determine who and how the breach occurred after it had already happened. If a breach of patient privacy occurs, the Office of Health and Human Services can ask any number of questions during their investigation, as can be seen in the article *HIPAA audit: The 42 questions HHS might ask* (Vijayan).

[Suppression of data with limited participation](#)

A common technique believed to ensure privacy is applying the rule that if the number of records meeting a query criterion falls below a defined threshold, then the results are not made available. An example of this approach might be the question, “How many patients of this clinic have three children?” where, if less than 4 patients have less than three children (4 being the threshold value), you do not display the results. The first problem with this approach is that

healthcare data is often comprised of small populations, which could result in frequent suppression of query results and consequent impairment of investigators to perform data analysis. In addition, this method can be broken. After asking a limited number of questions, for example, it may become possible to determine what that threshold is (if it is not already publicized) which will provide you a definitive insight into a small population albeit in a rather round-about way. In addition, to attack the above question – you need only ask the questions:

- Question 1: For how many patients is there a record of the number of their children?
- Question 2: How many patients have more than 3 children?
- Question 3: How many patients have less than three children?

Therefore, to answer “How many patients have 3 children?” you need only calculate Answer 1 – Answer 2 – Answer 3.

[Limit the number of queries to a dataset](#)

As can be seen from the above example, the ability to ask an unlimited number of questions may provide one with the ability to break any privacy protecting technique. Therefore, a common technique to assure privacy is to limit the number of questions a user is allowed to ask. By itself, this technique does not provide any privacy protections unless you have an extremely low threshold for the number of questions allowed to be asked (in the above example, you would only allow two questions). In combination with other privacy algorithms, it can be a helpful tool, however, because it prevents one from being able to reverse engineer the privacy algorithm applied to your data set using "brute force".

Privacy Algorithms

Within the privacy research and IT worlds, there are a number of privacy techniques that have been designed and employed only to later find limiting issues. These techniques include (but are not limited to):

- Disambiguation of data by applying algorithms to blank or null out certain fields in your data so that no row in the table is unique (Fischetti and Salazr; Øhrn and Ohno-Machado). Further research and analysis has determined that this methodology can be successfully attacked (Drieseitl, Vinterbo and Ohno-Machado).
- Trusted Third Parties (TTP) that separate the researcher and the original patient data with the full power to encrypt and decrypt the data. This method was pioneered by deCODE Genetics and ensures that no personal identifiers associated with medical data ever reach the analyst (Gulcher, Kristjánsson and Gudbjartsson). However, as noted from the Governor Weld case, there are other forms of attack to which TTP is susceptible (Malin, An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future). The same paper also provides information on the susceptibility of other privacy protection algorithms including *de-identification*, *de-nominalization*, and *SEMITRUST*.
- Cell Suppression is a technique in which individual columns or cells of data are suppressed thus preventing an attacker from identifying a patient because the data is missing key pieces of information (Sweeney, Achieving k-anonymity privacy protection using generalization and suppression; Su and Ozsoyoglu). Anonymization by way of cell suppression does not necessarily prevent inferences being made of that data (Ohno-Machado, Vinterbo and Drieseitl).

More recently, there have been efforts to mathematically define and formulate attacks against privacy as well as providing privacy guarantees (Chawla, Dwork and McSherry, Toward Privacy in Public Databases; Malin and Sweeney, How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems). The area that has received the most attention so far is that of k-anonymity.

Protecting Privacy by k-anonymity

A widely used approach for privacy protection is that of k-anonymity (Sweeney, k-anonymity: a model for protecting privacy). In addition, there are k-anonymity derivatives such as l-diversity (Machanavajjhala, Gehrke and Kifer), t-closeness (Li, Li and Venkatasubramanian), and m-invariance (Xiao and Tao) as well as are other suppression methodologies built on top of k-anonymity (Sweeney, Guaranteeing anonymity when sharing medical data, the datafly system; Wellner, Huyck and Mardis) to allow data release (Li, Wang and Jin).

The key fundamental of k-anonymity is the ability to measure privacy risk. For example, consider a dataset consisting of four fields: gender, birth date, zip code and person (Figure 2). When one denotes that this data set is anonymized to a value of $k=2$, this translates to a risk threshold of $1/k = 0.50$. If you were to view this data by a set of selected attributes (e.g. gender, birth date, and zip code), there must be at least two records for each combination of gender, birth date, and zip code that exist in the database. If there are attribute combinations within the dataset that have less than two records, then these records would not be sufficiently anonymized.

Gender	Birth date	Zip Code	Person
F	1/1/1970	98001	Person A
F	1/1/1970	98001	Person B
M	2/2/1971	98002	Person C
M	2/2/1971	98002	Person D
F	3/3/1972	98003	Person E
M	4/4/1973	98004	Person F
F	5/5/1974	98005	Person G

$k = 1$ { (rows 5-7)
 $k = 2$ { (rows 1-4)

Figure 2: Sample data set where the first four rows are anonymized to $k = 2$ while the last three rows are not sufficiently anonymized

Observe that the top four records in the example in Figure 2 have two sets of $k = 2$ data while the bottom 3 records have the value of $k = 1$. To ensure that your released data has a $k = 2$ anonymization, you will need to suppress the last three rows and only provide the top four rows for analysis. The concern with this approach is that the suppression of data will skew the analysis of your data because a significant proportion of the population may not be included in this new data set. To help mitigate the effects of this skew, the measurement error (information loss) can be calculated by the sample ratio and the level of suppression. El Emam and Dankar's *Protecting Privacy Using k-Anonymity* (El Emam and FK) extends this concept by the application of hypothesis testing prior to algorithmic application as well providing a categorization of risk and risk assessment tool.

Even with all of these tools, k -anonymity provides the ability to measure risk as opposed to providing privacy guarantees. In addition, the skew in data distribution especially within a healthcare setting of small sample populations may cause statistical bias. From more recent research, it has also been determined that k -anonymity and its derivatives are vulnerable to composition attacks and the risk assessment of $1/k$ may not be correct (Ganta, Kasiviswanathan and Smith).

Note that the purpose of this paper is not to judge the efficacy of k-anonymity nor its derivatives. However, based on the desire to have privacy guarantees without the statistical bias due to data suppression, the application of differential privacy algorithms to healthcare data presents an enticing alternative.

Protecting Privacy by Differential Privacy

[Defining Privacy and Differential Privacy](#)

A formal definition of privacy can be described as protection from being brought to the attention of others (Gavison). But to understand and provide guarantees of privacy, i.e. “guarantee that one will not be brought to attention of others”, it was important first to define privacy mathematically. The works of *Privacy-Preserving Data Mining in Vertically Partitioned Databases* (Dwork and Nissim, *Privacy-Preserving Data Mining in Vertically Partitioned Databases*) and *Practical Privacy: The SuLQ Framework* (Blum, Dwork and McSherry) expanded the mathematical statements abstracting the concept of a database, adversary, and define (mathematically) adversarial success. By having a mathematical definition of privacy, it became possible to mathematically define privacy attacks and protection. It is through these mathematical definitions that the technique of differential privacy was first described.

The basic principal of differential privacy is that privacy is guaranteed because nothing more about the individual can be learned when her information is in the database than when it is not.

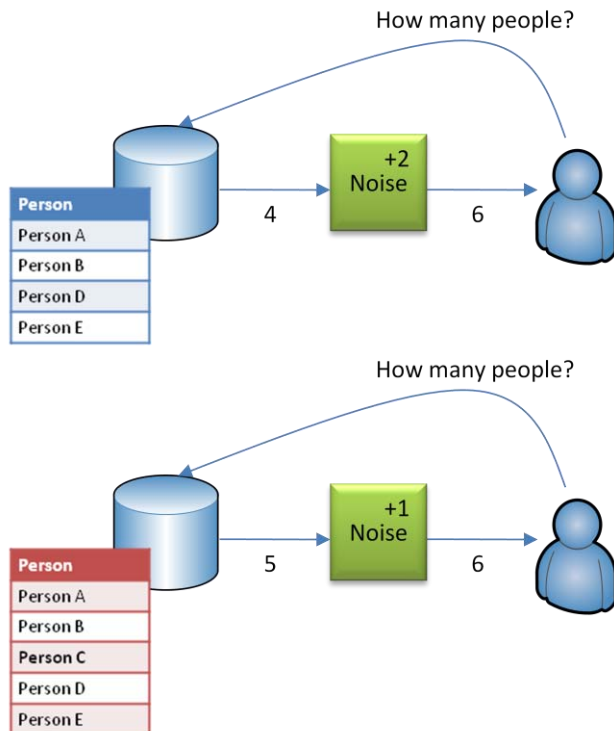


Figure 3: With differential privacy, one cannot learn more about an individual (Person C) whether she is in or not in the database

There are several works on differential privacy (Dwork, McSherry and Nissim; Blum, Dwork and McSherry; Chawla, Dwork and McSherry, On Privacy-Preserving Histograms; McSherry and Talwar, Mechanism design via differential privacy; Shuchi, Dwork and McSherry) which note end-to-end differential privacy guarantees on different statistical analyses without customization or needing to work in isolation. That is, differential privacy can work within the context of other analyses while keeping these guarantees (McSherry, Privacy Integrated Queries).

To do this, random noise generated by an exponential distribution is applied to the outgoing result. Using the example shown in Figure 3, the correct answer to the question "How many people are in the dataset?" is four. The addition of random exponential noise in this particular case adds a value of +2 which results in the answer of 6. In a query of a similar dataset, where

the correct answer is 5, with addition of exponential noise of +1, the result is also 6. Therefore, the person asking the question cannot learn any more information about the individual whether she is in (second case) or not in (first case) the dataset. This is a simple analogy of the impact of random noise to the result. In reality, the amount of noise is truly random hence the value applied and the corresponding final answer to the user may be quite different. For more in-depth information on the sanitization concept, please refer to Appendix 1: Differential Privacy concepts.

[Revisiting the Governor William Weld case](#)

Recall that to attack aggregate data, one need only to ask enough questions to drill down to a specific individual with very distinct attributes. In the Governor Weld case, the first question was the number of people who had his date of birth (answer is 6). The next question was the number of those people who were male (answer is 3). The final question is the number of those people who lived in his 5-digit zip code (answer is 1). In this example, only three questions were required for the author, Sweeney, to drill down to Governor Weld and his medical records (Sweeney, k-anonymity: a model for protecting privacy). However, if we were to add random exponential noise (i.e. some integer value between $-\infty$ to $+\infty$) to consistently change the above values, it would be impossible to drill down to this one person. For example, if the noise values to be applied to these questions were that of (-2, +1, +6), then the results of the above questions would be:

Question	Original Answer	Noise Values	Noisy Answer
How many patients with Governor Weld’s date of birth	6	-2	4
How many of those users are male	3	+1	4
How many of those users are in Governor Weld’s Zip Code	1	+6	7

When looking at the results from the application of differential privacy , the additional noise has made it impossible to drill down to the single individual, therefore the data is “privacy preserving”. In this particular example, the noise is large, too large for an investigator to discover any meaningful understanding of trends and patterns. The amount of added noise must be calibrated such that enough noise is added to the system that the answer changes but that the results of statistical analyses applied to the data do not change. In addition, to prevent attacks to the noise algorithm itself, more noise needs to be added as more questions are asked. The efficacy toward privacy guarantees of this methodology is not in question as this has already been amply researched and verified (Chawla, Dwork and McSherry, On Privacy-Preserving Histograms; McSherry, Privacy Integrated Queries) . The usefulness of the methodology in working environments, however, is still to be determined.

[Acceptability of the effects of privacy algorithms](#)

One of the key concerns of this approach is the acceptability of adding random noise to the results of the data. As noted in a pilot qualitative case study referenced in the Carnegie Mellon University / Microsoft Research Mindswap (Lee), the key to accepting the use of privacy algorithms and/or the addition of noise to the data is the users’ trust of the data. For more information, please refer to **Error! Reference source not found.** While we have a better understanding of the perspective of business analysts, the question arises as to how clinical researchers will react to additional noise being added to their research data. The same pilot qualitative study on the perceptions of clinical researchers (Lee) noted the following six themes:

1. **Statistical Accuracy:** Ensuring statistical precision and consistency is the most important perspective towards the application of privacy algorithms. Because of this, the bulk of this paper focuses on ensuring that the application of differential privacy to healthcare

data will still result in the same statistical outcomes as when differential privacy was not applied.

2. **Understanding the privacy algorithm:** Most clinical researchers are willing to accept the mathematics and science behind privacy algorithms without necessarily understanding them. While this increases acceptability, the potential problem is that one will implement privacy algorithms without understanding them and then incorrectly state a guarantee of privacy.
3. **Can get back to the original data:** It is a very important requirement that healthcare researchers can get back to the original data if so required. Some privacy algorithms perturb data so while guaranteeing patient privacy, it is impossible to get back to the patient. As differential privacy only affects the results, not the original data, this is not a concern.
4. **Understanding the purpose of the privacy algorithms:** Most educated healthcare professionals either understand issues surrounding patient privacy or would quickly understand it by providing case studies such as the Governor Weld case to make it more apparent. But it will be important to provide well worded text and/or confidence intervals, including publishing of the epsilon values, below a chart or report that has privacy algorithms applied.
5. **Management ROI:** The secondary use of patient data involves Institutional Review Boards (IRB), privacy and security steering committees, and HIPAA compliance. Therefore, the return-on-investment for the extra steps for security and privacy validation revolves around providing patient privacy guarantees that reduce the legal expenses attributed to revelation or perceived revelation of patient data.

6. **Protecting Patient Privacy:** It will become possible for researchers to use differential privacy to ensure patient privacy provided there is guidance on how to use it.

Differential privacy satisfies the latter five themes; it is the purpose of this paper to ascertain if differential privacy satisfies the theme of **statistical accuracy**.

Privacy-Integrated Queries (PINQ)

PINQ is a software application framework that applies the mathematics of differential privacy. It is built on top of Language-Integrated Query (LINQ) which is a .NET framework that allows developers and IT professionals to query any data source using the same methods (<http://msdn.microsoft.com/en-us/library/bb308959.aspx>). This allows users to apply differential privacy without understanding the mathematics behind it. For example, with LINQ you can write a .NET framework method to ask a question from a database (e.g. query a SQL database) and it will provide you with the answer (e.g. results from a SQL query to that database). With PINQ, you ask the same question but it will give you an answer with differential privacy applied, meaning that there will be some exponential noise added to your results. How much noise is used depends on the epsilon value, ϵ , which the user provides to PINQ. The larger the epsilon value, the less noise is applied to the result set, which means it is more accurate but potentially more privacy revealing. Conversely, the smaller the epsilon value, the more noise is applied to the result set which means there is more privacy but less accuracy.

IOP	# of patients
08-09	1
10-11	2
12-13	17

14-15	20
16-17	43
18-19	63
20-21	57
22-23	23
24-25	7
26-27	2
28-29	0
30-31	2
32-33	1

Table 1: Intraocular pressure ranges based on the Woolson dataset (Woolson)

To best see the effect of PINQ to a dataset, consider its use against a publicly available healthcare dataset, in this case a dataset containing the intraocular pressure for 238 patients visiting an ocular disease clinic (Woolson). A common way to view the dataset is in terms of intraocular pressure ranges as noted in Table 1. When applying differential privacy with an epsilon value of $\epsilon = 0.1$, an example of the result set that might be returned is shown in Table 2.

Note this is only one of many possible results; since differential privacy by PINQ is the application of random noise (by way of exponential distribution), the results can be different every time. The application of this random noise can result in wide ranges of variation. The higher the epsilon value, the smaller the degree of variation and the more accurate the values are to the real value.

IOP	# of patients (real value)	# of patients (differential privacy applied)
08-09	1	-1.089174486
10-11	2	0.332053694

12-13	17	16.71256706
14-15	20	18.67770913
16-17	43	62.75784853
18-19	63	79.93521335
20-21	57	44.60007283
22-23	23	28.04627043
24-25	7	-14.62243625
26-27	2	15.67088707
28-29	0	19.38616653
30-31	2	9.491168864

Table 2: Intraocular pressure ranges with the differential privacy applied at $\epsilon = 0.1$

This variation is best exemplified in Figure 4 where the actual value (blue) and six separate query executions at the set epsilon value (orange) are shown graphically side-by-side. At $\epsilon = 0.01$, the actual values are barely visible as the variation of differential privacy values is larger than the actual value. But at $\epsilon = 0.1$, the actual values are more visible and the magnitude of variation is not as significant. At $\epsilon = 1.0$, there is little variation as most of the values are very close to the actual values. Recall that the application of noise is based on an exponential distribution which means that there is an exponential (as opposed to linear) difference between values of 0.1 and 1.0.

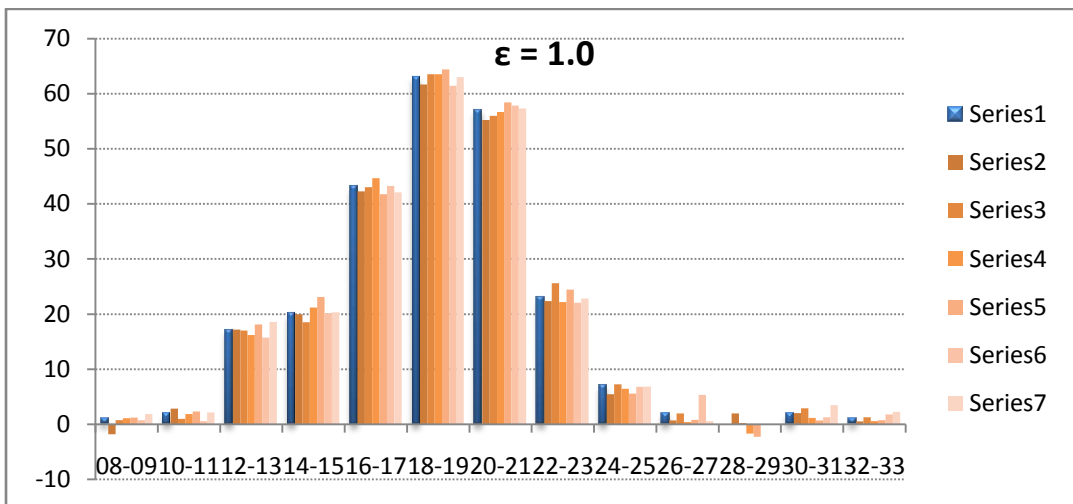
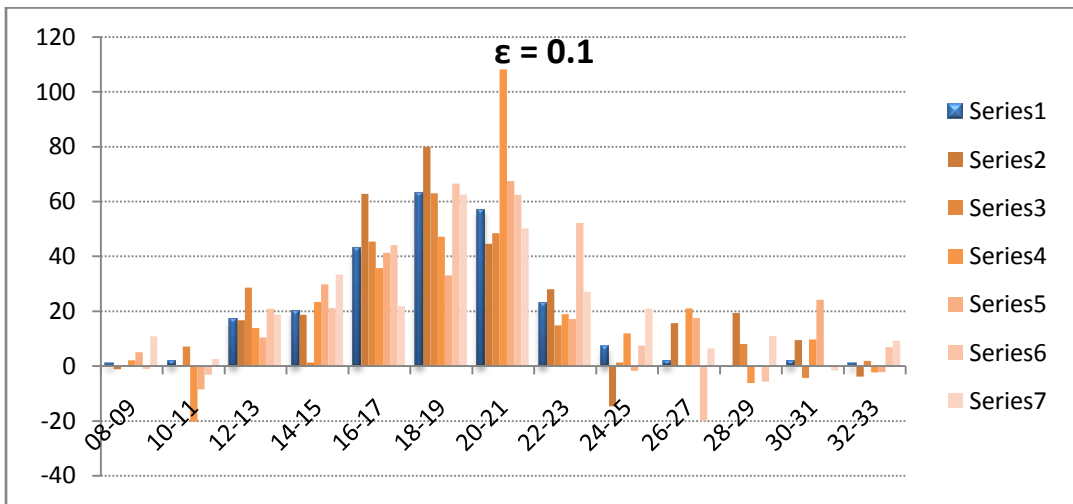
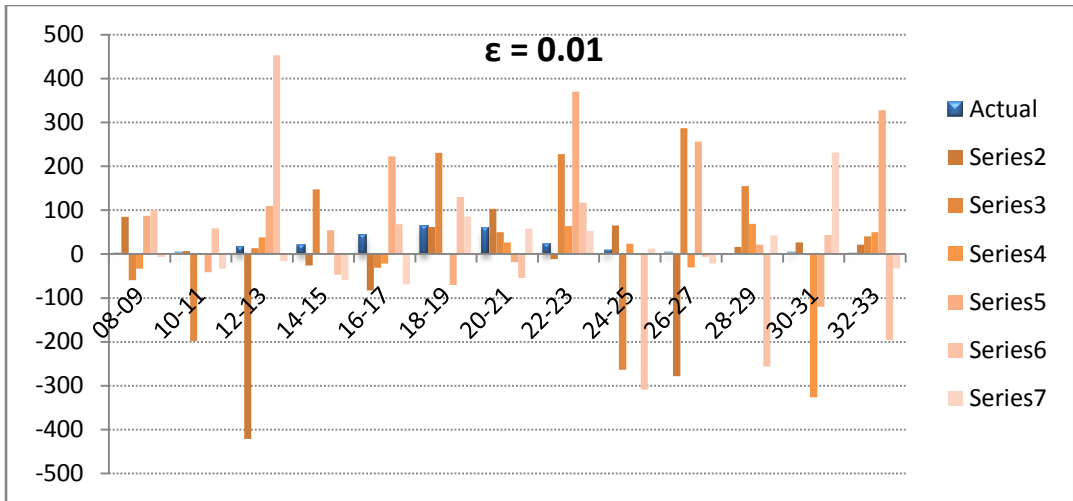


Figure 4: IOP ranges with various values of epsilon applied (differential privacy)

When there is a need for accurate values, the analysis can employ a higher epsilon value (e.g. $\epsilon = 1.0$) but the problem with this approach is that consistently providing accurate values will result in consistently having low privacy guarantees (i.e. the analysis may be privacy revealing). The goal is to provide enough accuracy in the result set to retain statistical significance while ensuring enough noise has been added to provide privacy guarantees. One solution is to introduce a gating mechanism to prevent investigators from asking too many high accuracy questions. The investigator could then ask a mix of questions with high accuracy (low privacy) and low accuracy (high privacy) epsilon values to meet the desired goals of statistical significance and patient privacy within an allowed total value of accuracy. To provide the privacy context, one should publish the epsilon values with the results.

Figure 5 shows the conceptual model of how PINQ gates the amount of privacy revelation when applying the exponential noise derived from the differential privacy algorithm. The gating mechanism is analogous to an ATM withdrawal from your bank except that instead of money, privacy units are withdrawn. In this example, a user is allowed use up to 20 privacy units (PrU) a day. This isn't the same as 20 questions a day because some questions are more complex than others. Questions that require filtering logic and associations to multiple data sources are more expensive. As well, the higher the epsilon value (i.e. lower privacy guarantees) the more expensive it is to ask these questions.

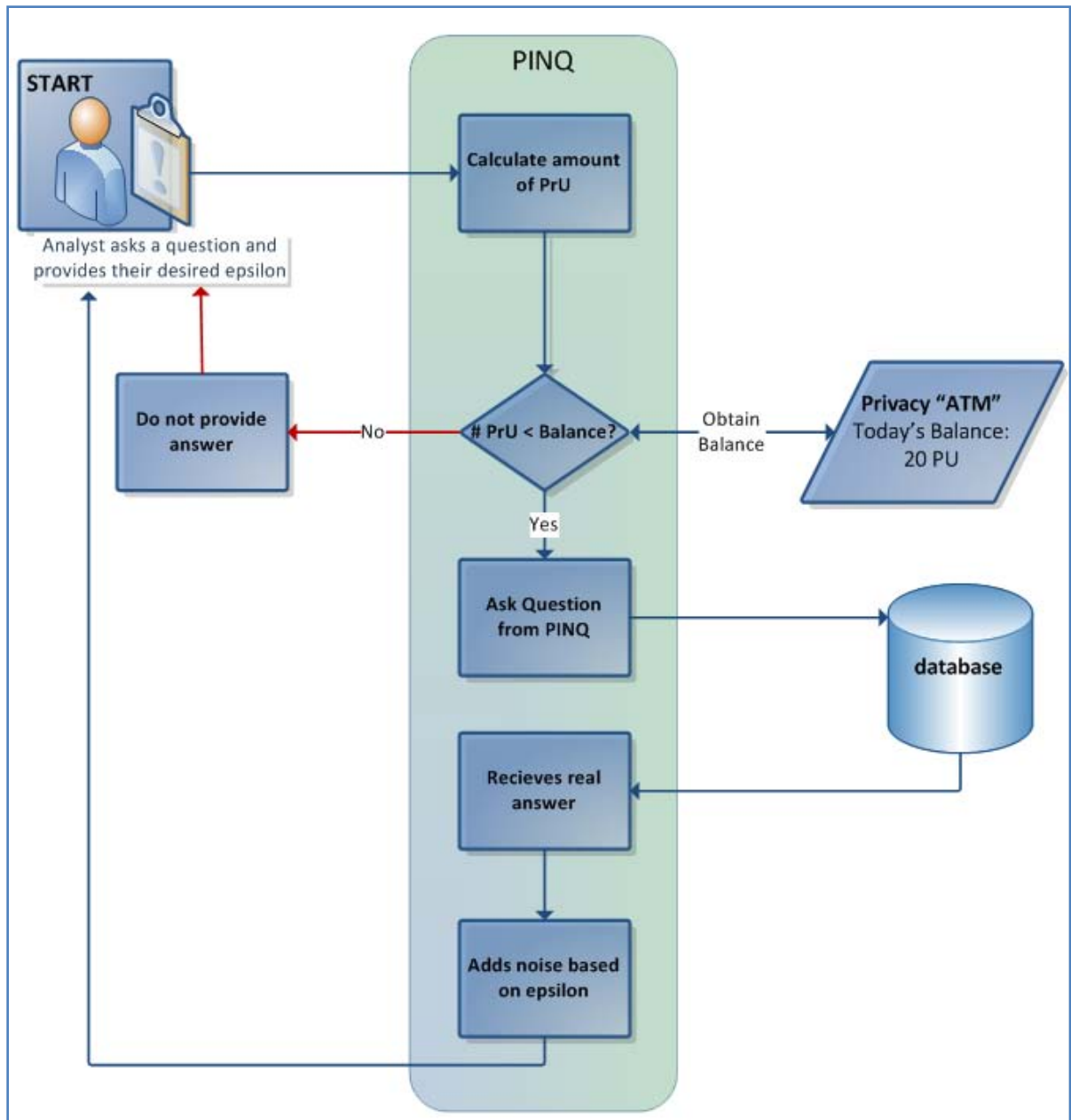


Figure 5: PINQ Privacy Gating Mechanism

The scenario proceeds with an analyst asking a question with their desired epsilon value (e.g. $\epsilon = 0.01$). When submitted to PINQ, it will calculate the amount of privacy units it takes to calculate this question (e.g. 0.05 PrU). Then it will compare that with the balance with the

amount you have that day (e.g. 20 PrU). Provided the privacy units it takes to answer this question is less than the available balance, PINQ can then submit the question to the data source (e.g. database), obtain the real answer, apply the exponential noise based on the epsilon value applied, provide the answer to the analyst and deduct the privacy unit amount from the available balance. If the question is very complex and/or the epsilon value is too high, the amount of calculated privacy units may be too high and be greater than the balance available. If this occurs, then PINQ will simply return without an answer. By design it will not indicate the reason for not providing the answer (e.g. epsilon too high, question too complex, not enough balance, etc.) as if revealed it may make PINQ more vulnerable to attack. For examples of how to use PINQ including the logging of privacy units, please refer to **Appendix 3: C# Code**

Example.

Determining the amount of privacy units an analyst is allowed to use on a daily basis and the amount of epsilon a user should apply for their questions will differ with different situations. Recall that the amount of epsilon that should be applied has an impact on the statistical significance of the answers you receive. Therefore, we will need to first provide guidance on how to determine the right amount of epsilon for a healthcare dataset. This in turn will ultimately lead to determining the right amount of daily privacy units. Determining the right amount of daily privacy units is out of scope for this paper as it will need to involve the study of different analyst scenarios (trusted researcher vs. journalist vs. third party analyst, etc.)

The Application of PINQ to Healthcare

The goal of the remaining portion of this paper is to examine the use of differential privacy on healthcare data. In particular, the objective is to devise a methodology to determine the correct values of epsilon for healthcare datasets. Because healthcare datasets typically involve smaller numbers of records, the concern is that epsilon values will need to be high to ensure statistical significance. It is important to determine, if possible, the right value of epsilon with enough variability to ensure the desired privacy guarantees and yet enough accuracy to retain statistical accuracy.

Methods

To determine statistical applicability of differential privacy to a healthcare dataset, we chose to apply the differential privacy algorithm (using PINQ) to the results of recently published research performed here at OHSU (Lieberman, Holub and Moravec). Data gathered by the Clinical Outcomes Research Initiative (CORI) through their national research network of gastrointestinal endoscopists was analyzed to determine if there is evidence of racial differences in the frequency and location of large or neoplastic colonoscopic polyps. The final result set is shown in Table 3.

Statistical difference in the two study groups (Caucasian non-Hispanic and African American non-Hispanic) was determined in the published analysis using Pearson's Chi-Square Test. Successful application of the differential privacy algorithm should, therefore, result in a dataset in which there is inaccuracy in the values (by addition of random noise) yet accuracy in the statistical results, i.e. the same significance can be found on comparison of the groups when analyzed with Pearson's Chi-Square Test.

A subset of this data was chosen for analysis as shown in Table 4. These three rows are representative of three possible scenarios: 1) higher values (n) with the difference in the groups found to be statistically significant, 2) lower n with no statistical significance, and 3) higher n with no statistical significance. Defining statistical significance as $p < 0.05$, it can be observed that African Americans with routine or average risk factors have a higher risk of advanced neoplasia compared to Caucasians ($p = 0.0002$). However, no difference is found in the other two subgroups, <50 and $50-59$ years of age ($p = 0.06$).

Characteristic	Caucasian non-Hispanic		African American non-Hispanic		p-value
	Total	With outcome	Total	With outcome	
Number (%)	80061	4964 (6.2)	5464	422 (7.7)	<.0001
By Screening group					
Routine/Average Risk	60380	3822 (6.3)	4366	339 (7.8)	.0002
Family history	19681	1142 (5.8)	1098	83 (7.6)	.02
By Age group					
<50	5279	222 (4.2)	421	26 (6.2)	.06
50 – 59	36400	1942 (5.3)	3100	190 (6.1)	.06
60 – 69	24125	1717 (7.1)	1378	145 (10.5)	<.0001
70 – 79	12194	934 (7.7)	501	54 (10.8)	.01
80+	2063	149 (7.2)	64	7 (10.9)	.26
By Age Category					
Age <60	41679	2164 (5.2)	3521	216 (6.1)	.02
Age ≥60	38382	2800 (7.3)	1943	206 (10.6)	<.0001
By Gender					
Female	38268	1769 (4.6)	2813	197 (7.0)	<.0001
Male	41793	3195 (7.6)	2651	225 (8.5)	.11
By Site type					
Community/HMO	64205	3910 (6.1)	4392	350 (8.0)	<.0001
Academic	8263	453 (5.5)	310	18 (5.8)	.81
VA/military	7593	601 (7.9)	762	54 (7.1)	.42

Table 3: Distribution of patients with Advanced Neoplasia

Characteristic	Caucasian non-Hispanic		African American non-Hispanic		p-value
	Total	With outcome	Total	With outcome	
By Screening group					
Routine/Average Risk	60380	3822 (6.3)	4366	339 (7.8)	.0002
By Age group					
<50	5279	222 (4.2)	421	26 (6.2)	.06
50 – 59	36400	1942 (5.3)	3100	190 (6.1)	.06

Table 4: Subset of queries from Table 3 that will be used for analysis

As noted above, the epsilon value within PINQ determines the amount of random exponential noise applied to a result set and therefore determines the degree of privacy protection. A low epsilon value provides higher privacy protection while a higher epsilon value provides more accuracy (i.e. the results are closer to the true values) but with less privacy guarantees. One aim

in this project is to define a methodology for selecting the best epsilon value, one which provides enough variability in the result set for privacy guarantees yet enough accuracy that the statistical findings in the result set will not be altered.

For this healthcare dataset, hypothesis testing using the Pearson’s Chi-Square test was performed for each value of epsilon. As the application of differential privacy will provide different results for each execution, we first performed 30 repetitions of the algorithm at each epsilon value. We then calculated the Average Query Error (LeFevre, DeWitt and Ramakrishnan; Kodeswaran and E) to narrow down the range of epsilon values that we will test. Average Query Error is defined as:

$$\text{Average Query Error} = \frac{\text{sum} (|\text{Real Value} - \text{PINQ Value}|)}{\text{Number of Queries}}$$

With the narrowed result set, we then calculated the average and standard deviation of the result set. These steps were repeated across a wide range of epsilon values.

[Analysis Example](#)

As an example of the analyses performed, consider the routine / average risk screening results. Table 5 is a contingency table from this data. Application of Pearson's Chi-square test gives $\chi^2 = 13.6975$, $p = 0.000215$.

Race	Outcome Variable		Total
	No Outcome (Total - Outcome)	Outcome	
Caucasian (Cau)	56558	3822	60380
African American (AA)	4027	339	4366
Total	60585	4161	64746

Table 5: Chi-Square Table for Routine/Average Risk screening group

Race	Outcome Variable		Total
	No Outcome (Total - Outcome)	Outcome	
Caucasian (Cau)	56395.16	3729.79	60124.95
African American (AA)	3890.91	261.50	4152.41
Total	60286.07	3991.29	64277.36

a.

Race	Outcome Variable		Total
	No Outcome (Total - Outcome)	Outcome	
Caucasian (Cau)	56676.45	3936.29	60612.74
African American (AA)	4151.25	436.25	4587.5
Total	60827.7	4372.54	65200.24

b.

Table 6: Contingency tables at a. ($\bar{x} - 2\sigma$) and b. ($\bar{x} + 2\sigma$)

Repeating the application of PINQ to this data provides as many possible values for each of the observations. At $\epsilon = 0.03$, with 30 repetitions of the algorithm, for example, the average and SD for Caucasian patients with the outcome of advanced neoplasia are $\bar{x} = 3833.04$ and $\sigma = 51.62$. This value of epsilon adds enough noise to assure privacy protections, but also (by introducing measurement error) modifies the statistical accuracy of the data. We consider the perturbation of the dataset to be statistically accurate if the values maintain statistical significance at ± 2 SD from the average. This means that 95% of the time that differential privacy is applied, the resulting (perturbed) dataset retains the statistical accuracy of the original dataset.

Tables 6a and 6b show the contingency table for the above example. Pearson's Chi-square test for Table 6a ($\bar{x} - 2\sigma$) results in $\chi^2 = 0.0441$, $p = 0.83$; for Table 6b ($\bar{x} + 2\sigma$), $\chi^2 = 61.1864$, $p = 5.19E-15$. Recall, the actual p-value is $p = 0.0002$ meaning that the difference in the outcome observations (7.8% of the African American population vs. 6.3% of the Caucasian population) is statistically significant. Because the ($\bar{x} - 2\sigma$) value is not statistically significant, we cannot use an $\epsilon = 0.03$ when applying differential privacy using PINQ. Other values of epsilon, however, may result in statistical accuracy.

Results

The results are organized by the three scenarios of Pearson’s Chi-Square tests performed to calculate the p-value to determine the statistical significance of the patient’s comparative risk of advanced neoplasia within the context of characteristic and ethnicity.

Higher values (n), differences statistically significant

The first group to undergo analysis is the screening group of routine / average risk which has a relatively high n. There is statistical significance indicating that African Americans have a higher risk than Caucasians for advanced neoplasia ($p = 0.0002$) as noted in Table 7.

There is a similar exponential curve pattern irrelevant of population size when viewing the Average Query Error for a range of epsilon values ($0.001 \leq \epsilon \leq 1.0$) for all four categories as noted in Figure 6. Whether dealing with a smaller (e.g. $n_{\{AA\ Outcome\}} = 338$) or a larger (e.g. $n_{\{Cau\ Total\}} = 60380$) population, the AQE curves imply that the epsilon value will need to be $\epsilon \geq 0.04$ to ensure statistical accuracy (the point of the where the curves converge and plateau).

When reviewing the average query error percentage ($AQE\% = AQE/n$) observed in Figure 7, it notes that sample size has an impact on the amount of noise applied (as opposed to the AQE graph in Figure 6). Because $n_{\{AA\ Outcome\}}$ is smaller in size as compared to the other three datasets, it requires a higher epsilon value (e.g. $\epsilon \geq 0.1$) before the graphs coalesce indicating a higher epsilon value will be required to ensure statistical accuracy.

Characteristic	Caucasian non-Hispanic		African American non-Hispanic		p-value
	Total	With outcome	Total	With outcome	
Routine/Average Risk	60380	3822 (6.3)	4366	339 (7.8)	.0002

Table 7: Routine / Average risk patients with Advanced Neoplasia

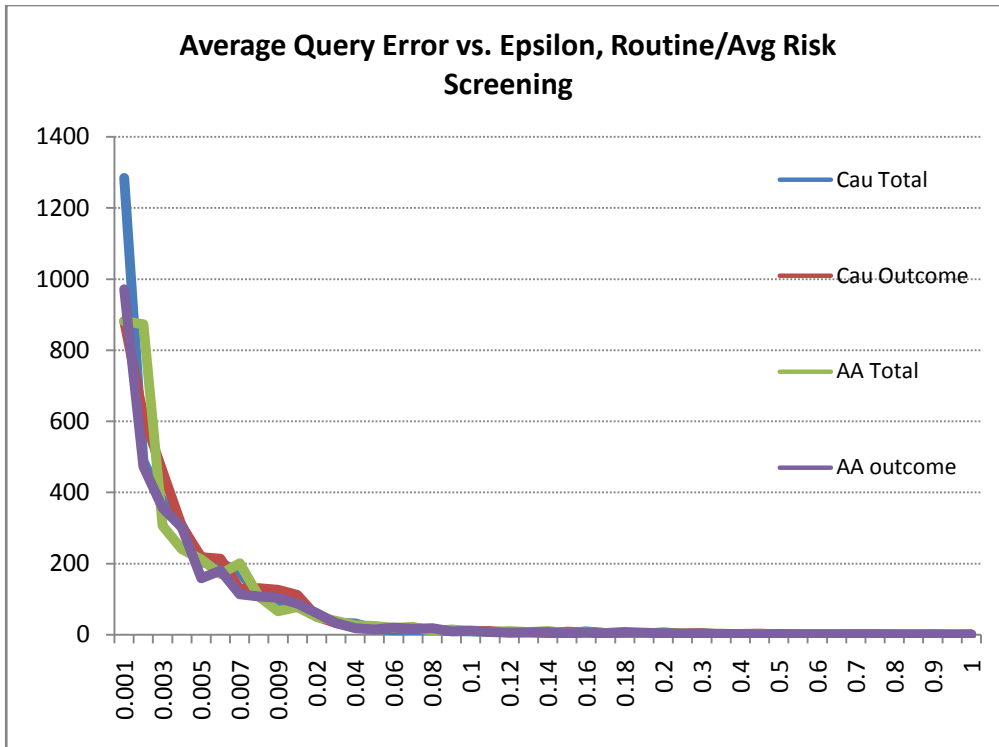


Figure 6: Average query error by epsilon for all routine/average risk screening groups

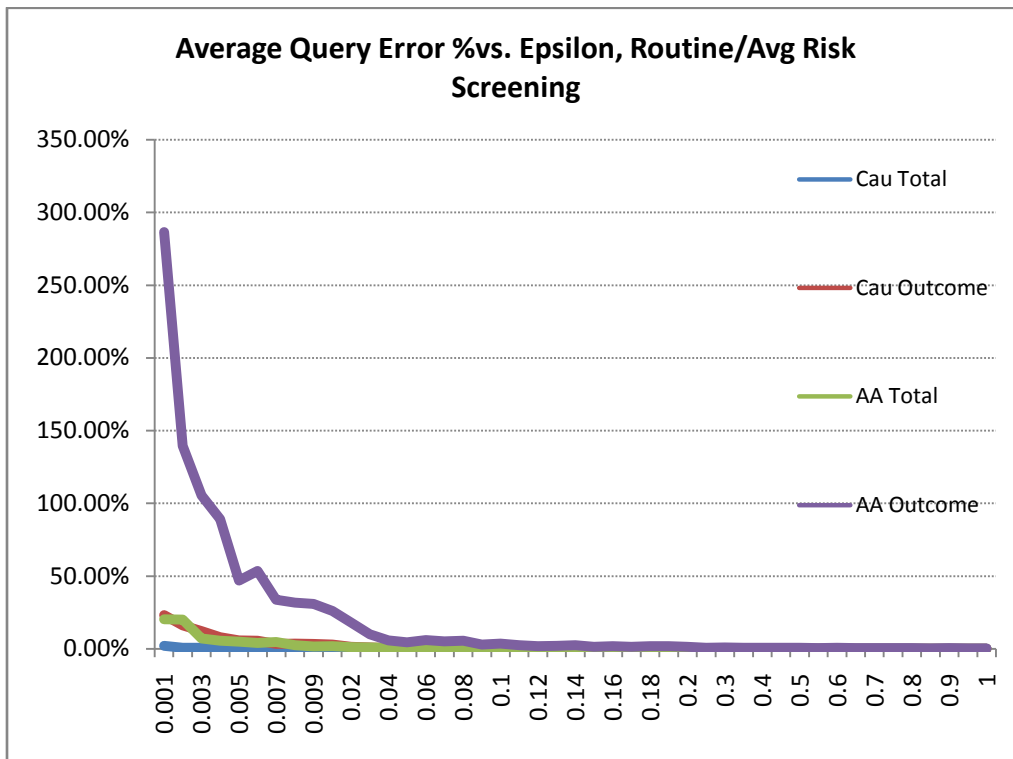


Figure 7: Average query error percentage by epsilon for all routine/average risk screening groups

	Cau Total	Cau Outcome	AA Total	AA Outcome
n	60380	3822	4366	339
AQE	33.908	33.3886	39.4536	33.2521
AQE%	0.06%	0.87%	0.90%	9.81%

Table 8: Focusing on the $\epsilon = 0.03$ values

To elaborate the difference between AQE and AQE%, the values at $\epsilon = 0.03$ can be seen in Table 8. While the AQE for all four values is approximately 33, the AQE% is quite different (AQE%_{AA Outcome} = 9.81%, AQE%_{Cau Total} = 0.06%).

Recall our goal is to find the right level of epsilon that can be applied in this scenario such that there is a high degree of variability while at the same time statistical significance (or insignificance) is unaffected. Based on the AQE graph in Figure 6, the degree of variability is less extreme at $\epsilon = 0.04$. But the AQE% graph in Figure 7 indicates a less extreme degree of variability at $\epsilon = 0.1$. To help determine which metric is the correct starting point to determine the right epsilon, we will start with $\epsilon = 0.04$ since it is lower.

Table 9 provides the list of p-values calculated on differential privacy applied data for the epsilon range of 0.04 to 1.0. The goal is to have the p-value close to the actual p-value ($p = 0.0002$) for the average and $\pm 2SD$ (\bar{x} , $\bar{x} - 2\sigma$, $\bar{x} + 2\sigma$). If you review $\epsilon = 1.0$ you'll notice that all three values are close to the actual p-value. But at this level of epsilon, there is very little noise applied to the data ($\max(\text{AQE}\%) = 0.29\%$). That is, at $\epsilon = 1.0$ you have a high degree of accuracy but there is very little privacy guarantee. Therefore, we will want to review other smaller epsilon value candidates. The epsilon range we will want to focus on is between 0.12 and 0.25 as noted in Table 9 because these values are close to the original p-value of $p = 0.0002$. Figure 8 provides this view within the context of a high/low bar graph.

epsilon	$P_{(\rho, \bar{x})+2\sigma_p}$	$P_{(\rho, \bar{x})-2\sigma_p}$	$P_{(\rho, \bar{x})}$	max(AQE%)
0.04	4.9E-07	0.286987	0.001906	5.64%
0.05	6.17E-09	0.059931	9.65E-05	4.40%
0.06	5.7E-11	0.202343	6.64E-05	5.93%
0.07	1.59E-08	0.112237	0.000228	5.11%
0.08	3.15E-09	0.235576	0.000304	5.58%
0.09	5.73E-06	0.012186	0.0004	2.84%
0.1	6.6E-07	0.02092	0.000249	3.50%
0.11	2.81E-06	0.002463	0.00011	2.37%
0.12	2.92E-05	0.002585	0.000314	1.71%
0.13	9.21E-06	0.003541	0.00023	2.07%
0.14	1.03E-05	0.006124	0.000335	2.46%
0.15	4.16E-05	0.001076	0.000226	1.22%
0.16	7.92E-06	0.00242	0.000171	1.75%
0.17	2.02E-05	0.001128	0.000168	1.27%
0.18	1.18E-05	0.003244	0.000242	1.79%
0.19	7.77E-06	0.002513	0.000175	1.67%
0.2	1.59E-05	0.001153	0.000153	1.22%
0.25	8.07E-05	0.000502	0.000206	0.63%
0.3	6.97E-05	0.001332	0.000324	0.90%
0.35	5.17E-05	0.000593	0.000182	0.75%
0.4	9.16E-05	0.000591	0.000238	0.68%
0.45	9.07E-05	0.000554	0.000229	0.55%
0.5	6.67E-05	0.000687	0.000222	0.63%
0.55	9.73E-05	0.000415	0.000204	0.45%
0.6	9.28E-05	0.000563	0.000234	0.58%
0.65	9.83E-05	0.000389	0.000198	0.47%
0.7	9.85E-05	0.000373	0.000194	0.35%
0.75	7.55E-05	0.000438	0.000186	0.47%
0.8	9.84E-05	0.000455	0.000215	0.46%
0.85	0.000166	0.000315	0.000229	0.21%
0.9	0.000101	0.000501	0.000229	0.42%
0.95	0.00015	0.000354	0.000231	0.28%
1	0.000138	0.000359	0.000224	0.29%

Table 9: p-value calculated on PINQ applied data - Avg, Avg+2Std, Avg-2Std, max(AQE%)

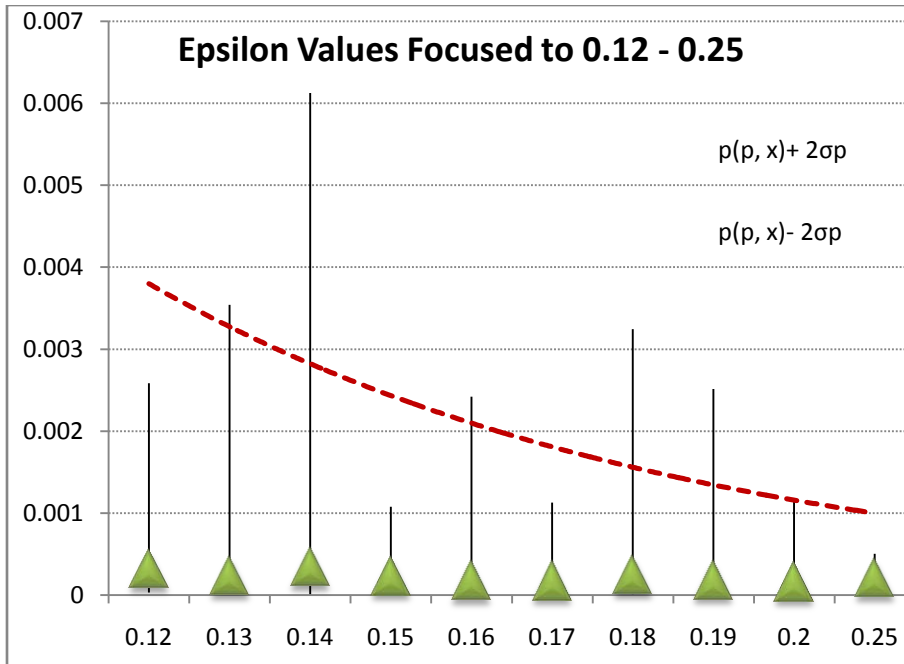


Figure 8: Epsilon values focused on 0.12 to 0.25

Within Figure 8, the triangle figures signify the average p-value (hovering at around 0.0002) against differential privacy applied data. The high/low bar provides the range of values found between +/- 2 standard deviations. The desire is to have a low epsilon value that has a high variation (i.e. longer high/low bar) while the average value is as close to $p = 0.0002$ as possible. The top three candidates would be epsilon values 0.14, 0.13, and 0.18. But since $\epsilon = 0.14$ has $p_{avg} = 0.00033$ and $\epsilon = 0.18$ has less variability, the best choice here is to use $\epsilon = 0.13$ with its higher variability, $p_{avg} = 0.00023$, and at +/- 2 standard deviations the p-value observations all note statistical significance.

Because PINQ is applying the differential privacy algorithm via random exponential noise uniformly across sample sizes, at smaller sizes there is larger amount of relative error that will have impact the statistical calculations that will be applied to this data. This observation also notes that to narrow down the range of epsilon candidates, one will need to use the AQE% metric ($\epsilon \geq 0.1$) as opposed to the AQE metric ($\epsilon \geq 0.04$) as the starting point.

Lower n with no statistical significance

The first group involving routine/average risk screening is an easier test because of the statistically significant outcome. In the case of the <50 age group, we will apply the same analysis logic but the solution is more complex because the differences between the ethnicities in this age group is not statistically significant ($p = 0.06$). If we apply too much error, the differential privacy applied results can more easily provide a false-positive (falsely indicating statistical significance). The actual values can be seen in Table 10. As observed in the previous test, to narrow down the range of candidate epsilon values, we will start with the AQE% graph as noted in Figure 9.

Characteristic	Caucasian non-Hispanic		African American non-Hispanic		p-value
	Total	With outcome	Total	With outcome	
<50	5279	222 (4.2)	421	26 (6.2)	.06

Table 10: By age group <50 with Advanced Neoplasia

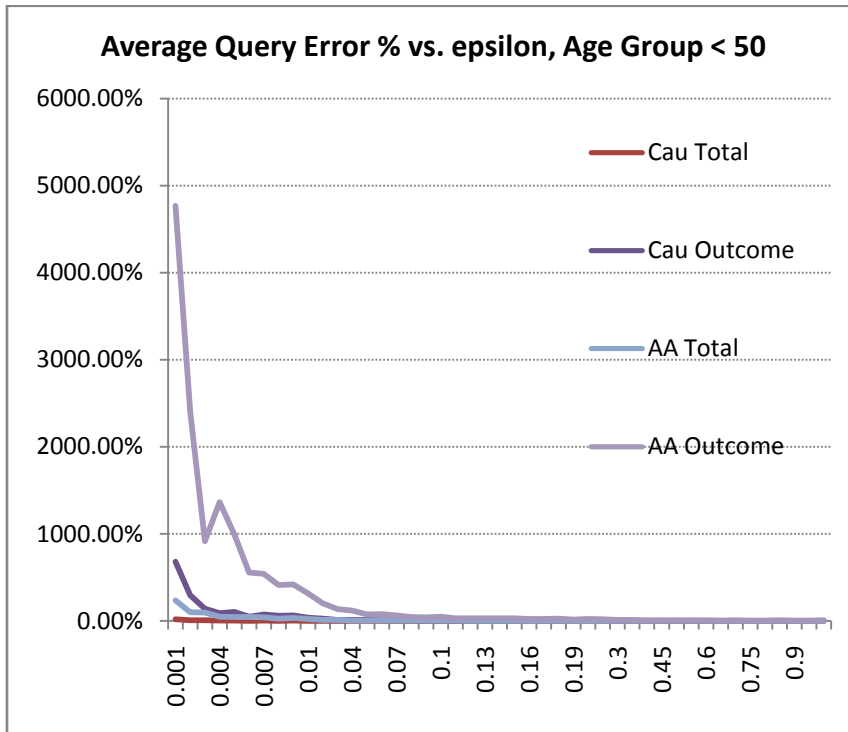


Figure 9: Average query error percentage by epsilon for age < 50 groups

Epsilon	$P_{(p, \bar{x})+2\sigma_p}$	$P_{(p, \bar{x})-2\sigma_p}$	$P_{(p, \bar{x})}$
0.4	0.058694814	0.302132392	0.0072022
0.45	0.089662818	0.283554003	0.023362266
0.5	0.107305992	0.437373461	0.016990851
0.55	0.090430474	0.337005333	0.017400007
0.6	0.091122496	0.611744416	0.005380644
0.65	0.070627906	0.261704467	0.014270468
0.7	0.097995276	0.663075723	0.005278488
0.75	0.083205255	0.183233662	0.034329218
0.8	0.06356584	0.170854703	0.020451077
0.85	0.09520205	0.332489944	0.019563679
0.9	0.057815042	0.21576649	0.011486068
0.95	0.085943971	0.357958722	0.013319093
1	0.085547165	0.223119552	0.027615096

Table 11: p-value calculated on PINQ applied data for <50 age group - Avg, Avg+2Std, Avg-2Std

Because of the smaller sample size ($n_{\{AA\ Outcome\}} = 26$) it will be soon observed that this larger AQE% will have a profound impact on the analysis of differential privacy applied datasets. Upon narrowing down the candidate range of epsilon values ($\epsilon \geq 0.4$), Table 11 provides a list of the different p-values derived from the \bar{x} , $\bar{x} - 2\sigma$, and $\bar{x} + 2\sigma$ based on thirty repeated runs.

All of the p-values for ($\bar{x} + 2\sigma$) and ($\bar{x} - 2\sigma$) for the entire range of epsilon values indicate lack of statistical significance. But even at $\epsilon = 1.0$, the \bar{x} value denotes statistical significance ($p = 0.0276$) due to the small population size ($n_{\{Age < 50, AA\ Outcome\}} = 26$). With so much noise added to the final result, the statistical outcome resulted in a false positive of statistical significance.

Since the AQE% variation is high between the four different values making up the contingency table, we altered the approach slightly so that we apply a different epsilon value to the {AA Outcome} category since it has the smallest sample size. In the previous test we had applied the same ϵ value to make the test and method easier. But it is apparent that with the difference in sample size and AQE% variation, it is necessary to use a larger epsilon value for {AA Outcome}.

ε	$\varepsilon_{\{AA, Outcome\}}$	$p_{(\rho, \bar{x})+2\sigma_p}$	$p_{(\rho, \bar{x})-2\sigma_p}$	$p_{(\rho, \bar{x})}$	$\max(AQE\%)$
0.18	0.9	0.023866	0.130846	0.056331	4.58%
0.18	0.95	0.026711	0.243806	0.08497	5.15%
0.18	1	0.055933	0.130216	0.084111	7.68%
0.18	1.05	0.056194	0.09544	0.071969	3.07%
0.18	1.1	0.064358	0.090271	0.075024	2.87%
0.18	1.15	0.056949	0.101357	0.074617	3.54%
0.18	1.2	0.054636	0.11429	0.077803	3.51%
0.18	1.25	0.04017	0.094349	0.060545	3.21%
0.18	1.3	0.026874	0.122391	0.057329	3.57%
0.18	1.35	0.059742	0.074593	0.065759	2.71%
0.18	1.4	0.07774	0.076671	0.076388	2.49%
0.18	1.45	0.046262	0.094568	0.065126	3.36%
0.18	1.5	0.075438	0.06578	0.069893	2.28%
0.18	1.55	0.07038	0.114788	0.08841	2.72%
0.18	1.6	0.082181	0.066207	0.073341	2.21%
0.18	1.65	0.057154	0.086465	0.069103	2.97%
0.18	1.7	0.081752	0.058565	0.069044	1.87%
0.18	1.75	0.098768	0.045819	0.068471	1.44%
0.18	1.8	0.072244	0.057398	0.064032	2.05%
0.18	1.85	0.084445	0.071142	0.076977	2.26%
0.18	1.9	0.108015	0.046341	0.072346	1.31%
0.18	2	0.079584	0.074867	0.076403	2.30%
0.18	2.5	0.096872	0.055402	0.073819	1.71%
0.18	3	0.118764	0.039313	0.07101	0.86%
0.18	3.5	0.117503	0.046768	0.07617	1.14%
0.18	4	0.119194	0.038475	0.07049	0.81%
0.18	4.5	0.10883	0.045748	0.072252	1.13%
0.18	5	0.117451	0.040754	0.071679	0.74%
0.18	6	0.126259	0.037945	0.072432	0.65%
0.18	7	0.134026	0.036274	0.073549	0.54%
0.18	8	0.133871	0.035984	0.073247	0.48%
0.18	9	0.135168	0.034845	0.072652	0.42%
0.18	10	0.140198	0.034902	0.074282	0.37%

Table 12: p-value calculated on PINQ applied data - \bar{x} , $\bar{x}+2\sigma$, $\bar{x}-2\sigma$, $\max(AQE\%)$ for different ε values for {AA, Outcome}; $\varepsilon = 0.18$ for all other categories.

To find the right epsilon value for the {AA Outcome} category, we needed to re-test using different epsilon values. Table 12 provides the list of p-values calculated on PINQ applied data for the epsilon range of 0.9 to 10.0 for {AA, Outcome} and $\epsilon = 0.18$ for the other three categories. The latter epsilon value was determined using the same technique but focused only on those three categories (as opposed to all four). Recall that we want all three metrics ($p_{(p, \bar{x})} + 2\sigma_{p, p_{(p, \bar{x})}} - 2\sigma_{p, p_{(p, \bar{x})}}$) to be as close to the actual value ($p = 0.06$) as possible without actually going under $p = 0.06$ since a lower value and a threshold of $p = 0.05$ that would falsely indicate statistical significance.

Narrowing down the range further by reviewing the original AQE% graph (Figure 9) and Table 12, we focused on the epsilon values for the {AA, Outcome} at $0.9 \leq \epsilon \leq 1.75$ as noted in Figure 10. The blue dashed bar within Figure 10 indicates that actual value of $p = 0.06$; any values that drop significantly below this bar ($p < 0.055$) will be excluded. As can be seen in Figure 10 and verified in Table 12, the average, average $\pm 2S D$ values at $\epsilon = 1.05$ for {AA, Outcome} and $\epsilon = 0.18$ for all other categories describe the same degree of statistical insignificance as the actual value. In addition to $\epsilon = 1.05$ being one of the lowest values and higher AQE% values (AQE% = 3.07%), it is also surrounded by two values that are also valid. Both $\epsilon = 1.0$ and $\epsilon = 1.1$ are acceptable as well increasing the chance that this value is valid as opposed to luck.

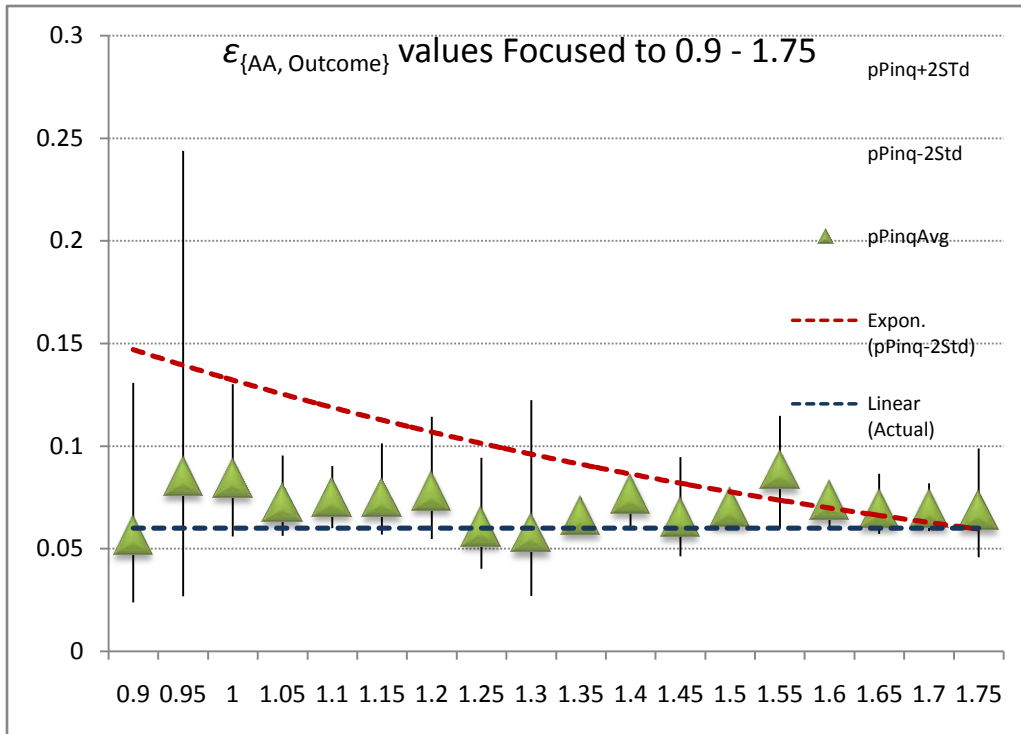


Figure 10: Epsilon values for {AA, Outcome} focused on 0.9 - 1.75

Higher n with no statistical significance

Based on the lessons learned from the first two scenario tests, we have determined a methodology to find the right balance of privacy and accuracy. But now that we have determined the template ϵ values, the question remains on whether we can apply those template values to a different dataset. Note that in Table 13 that the <50 and 50-59 age groups have a similar profile – ~6% of the African American population has the outcome with relatively small sample sizes. To determine if we can use the previous analyses as our template, we used the results from the Lower n with no statistical significance (Age Group: <50) analysis and executed the query runs using $\epsilon = 1.05$ for {AA Outcome} and $\epsilon = 0.18$ for all other categories for the Age Group 50-59 characteristic.

Characteristic	Caucasian non-Hispanic		African American non-Hispanic		p-value
	Total	With outcome	Total	With outcome	
<50	5279	222 (4.2)	421	26 (6.2)	.06
50 – 59	36400	1942 (5.3)	3100	190 (6.1)	.06

Table 13: By age group <50 and 50-59 with Advanced Neoplasia

As can be seen in Table 14, we successfully applied the template epsilon values from the <50 age group to the 50-59 age group with accurate results (i.e. statistical insignificance with $p \approx 0.06$).

This can be confirmed by reviewing all of the values for ϵ range of (0.7- 1.9) for {AA Outcome} as seen in Figure 11.

ϵ	$\epsilon_{\{AA, Outcome\}}$	$p_{(p, \bar{x})} + 2\sigma_p$	$p_{(p, \bar{x})} - 2\sigma_p$	$p_{(p, \bar{x})}$	P
0.18	1.05	0.06001	0.07167	0.06551	0.06

Table 14: p-values for average, avg +/- 2std at epsilon = 0.18 and epsilon = 1.05

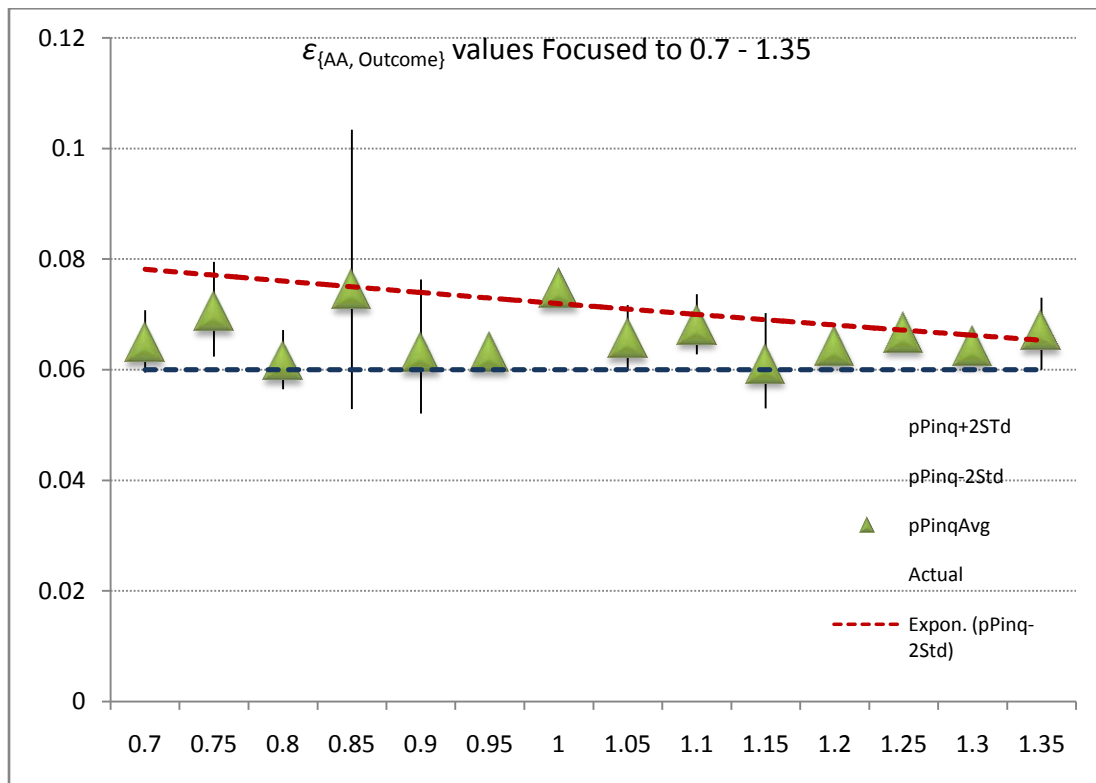


Figure 11: epsilon values focused to 0.7 - 1.35

Discussion

Since healthcare datasets typically employ smaller sample sizes, our concern was that the application of differential privacy in the form of exponentially distributed random noise would result in a high degree of measurement error. Because the noise is random, repeated questions have different amounts of noise applied resulting in different answers. We consider the perturbation of the dataset to be statistically accurate if the values maintain statistical significance at +/- 2 SD from the average. This means that 95% of the time that differential privacy is applied, the resulting (perturbed) dataset retains the statistical accuracy of the original dataset.

As demonstrated by the above tests and analysis against the CORI dataset, we can use PINQ to apply differential privacy against healthcare datasets to provide privacy guarantees and ensure statistical accuracy. For the first test concerning the Higher values (n) with statistical significance (Routine / Average Risk screening group), it was determined that African Americans have statistically significant higher risk ($p = 0.0002$) for advanced neoplasia as compared to Caucasians. By reviewing the various tests (AQE%, Pearson's Chi-Square test) at different degrees of epsilon, it was determined that the ideal epsilon value was $\epsilon = 0.13$. The Lower n with no statistical significance (<50 age group) was different in that while the African Americans had a higher risk than Caucasians for advanced neoplasia, the difference was not statistically significant ($p = 0.06$). As well, the {AA Outcome} sample size was relatively small ($n = 26$). Based on additional tests and analysis, it was determined that using the same epsilon for all questions would not provide statistically accurate results. The key lesson here was that for smaller sample sizes, a larger epsilon would need to be applied. In the case of our tests, we had applied $\epsilon = 0.18$ for the other three categories within the contingency table and experimented

with a range of epsilon values for the {AA Outcome} category. Ultimately, we were able to find an appropriate epsilon value, $\epsilon = 1.05$, for this category. In addition to having a lot of variability (e.g. relatively high standard deviation and AQE%) and statistically insignificant answers at the same magnitude as the non-PINQ applied data, we also chose a value where $\epsilon + 0.05$ and $\epsilon - 0.05$ were also statistically insignificant reducing the chance that the chosen epsilon value was arbitrary.

These tests allowed us to develop the process and methodology that analysts can use to determine their appropriate epsilon value. We had also wanted to see if we could apply these learnings to another group, i.e. apply the epsilon values for the Lower n with no statistical significance (<50 age group) to the Higher n with no statistical significance (50-59 age group). Using the same epsilon values ($\epsilon = 0.18$ for three categories, $\epsilon = 1.05$ for the {AA, Outcome} category), we obtained a similar set of p-values: $p_{(p, \bar{x})} = 0.066$, $(p_{(p, \bar{x}) - 2\sigma_p}) = 0.0716$, $(p_{(p, \bar{x}) + 2\sigma_p}) = 0.06$, and $p_{\text{Actual}} = 0.06$. While not included in this analysis, one could have also applied different levels of epsilon for each of the other three categories such that each category within the contingency table had a different ϵ value. More generically, one could simply apply a different ϵ value for every single question asked if so desired.

As observed from the above tests, we had achieved our goals of providing guidance on how to apply differential privacy using PINQ against a healthcare dataset and obtain statistically accurate results. The methodology involves the testing of different epsilon values against a known dataset and then applying that knowledge to an unknown dataset. Over time, with more analysts using differential privacy, it will be possible to build up additional guidelines to determine the correct levels of epsilon. Ultimately, knowing the right level of epsilon for your datasets will allow you to determine the appropriate level of “privacy unit” funds within your

privacy gating mechanism. It will take a combination of profiling the data (e.g. sample size, AQE%, etc.) and quantitatively measuring analyst access rights (e.g. trusted researcher vs. journalist) to eventually build the differential privacy application rules. Nevertheless, the results from these tests provide us a starting point.

Conclusion

Providing patient privacy is a key tenet to the future of healthcare – to transform medicine from curative to preemptive. It is the analysis and sharing of data that is fundamental to discovering patterns within the healthcare population. Yet, this cannot be achieved unless patients trust that their privacy is guaranteed and healthcare institutions have the processes and tools to implement these safeguards.

While there are many privacy techniques, many of them do not fulfill the six themes describing the acceptability of the effects of privacy algorithms by healthcare researchers. Many of these techniques provide a return on investment for management because they reduce the legal expenses associated with (perceived) patient data revelation. The perturbation effects of many algorithms provide the ability to get back to the original dataset when necessary. While many healthcare researchers do not understand (nor want to) the mathematics behind a privacy algorithm, they understand the issues surrounding patient privacy and accept some form of data perturbation. How differential privacy separates itself from most algorithms is that the mathematics behind it allows one to provide privacy guarantees, not just provide privacy risk. While the application of differential privacy, made easier by using PINQ, provides privacy enforcement, a cause for concern is the impact on statistical accuracy when one adds exponential random noise to the result set.

By using AQE% graphs and replicating and repeating the original statistical tests at the bounds of 95% confidence interval on PINQ applied datasets, we have provided a methodology to identify candidate epsilon values (the value the analyst sets that determines the accuracy of the result). By applying lower epsilon values for higher sample sizes and vice versa, we were able to

demonstrate that differential privacy can be applied to healthcare datasets to ensure privacy and provide statistically accurate results.

Works Cited

Annas, GJ. "Medical Privacy and Medical Research - Judging the New Federal Regulations." New England Journal of Medicine 346.3 (2002): 216-219.

Barrows, RD and PD Clayton. "Privacy, Confidentiality, and Electronic Medical Records." Journal of the American Medical Informatics Association 3.2 (1996): 139-148.

Blum, A, et al. "Practical Privacy: The SuLQ Framework." The 24th Symposium on Principles of Database Systems (PODS 2005). Baltimore: Association for Computing Memory, 2005. 128-138.

Buckovich, S, H Rippen and M Rozen. "Driving toward guiding principles: a goal for privacy, confidentiality, and security of health information." Journal of the American Medical Informatics Association 6 (1999): 122-133.

Cannon, JC. Privacy: What Developers and IT Professionals Should Know. Boston: Addison-Wesley Professional, 2004.

Chawla, S, et al. "On Privacy-Preserving Histograms." Uncertainty in Artificial Intelligence. Edinburgh: Association for Uncertainty in Artificial Intelligence, 2005.

Chawla, S, et al. "Toward Privacy in Public Databases." Second Thoery of Cryptography Conference (TCC 2005). Cambridge, MA: International Assocation for Cryptologic Research, 2005. 363-385.

Collmann, J and T Cooper. "Breaching the Security of the Kaiser Permanente Internet Patient Portal: the Organization Foundations of Information Security." Journal of the American Medical Informatics Association 14.2 (2007): 239-243.

Drieseitl, S, S Vinterbo and L Ohno-Machado. "Disambiguation Data: Extracting Information from Anonymized Sources." Journal of the American Medical Informatics Association 9 (Nov-Dec Supplement) (2002): S110-S114.

Dwork, C and K Nissim. "Privacy-Preserving Data Mining in Vertically Partitioned Databases." Advances in Cryptology—CRYPTO 2004. 2004. 528-544.

Dwork, C, et al. "Calibrating Noise to Sensitivity in Private Data Analysis." The Third Theory of Cryptography Conference (TTC 2006). New York: International Association for Cryptologic Research, 2006. 265-284.

El Emam, K and Dankar FK. "Protecting Privacy Using k-Anonymity." Journal of the American Medical Association 15.5 (2008): 627-637.

Fischetti, M and J Salazar. "Models and algorithms for the 2-dimension cell suppression problem in statistical disclosure control." Mathematical Programming 84 (1999): 283-312.

Fogelholm, R, S Avikainen and Murros K. "Prognostic Value and Determinants of First-Day Mean Arterial Pressure in Spontaneous Supratentorial Intracerebral Hemorrhage." Stroke 28 (1997): 1396-1400.

Ganta, SR, S Kasiviswanathan and A Smith. "Composition Attacks and Auxiliary Information in Data Privacy." ACM International Conference on Knowledge Discovery and Data Mining (SIG-KDD). Las Vegas: Association for Computing Machinery (ACM), 2008.

Gavison, R. Privacy and the Limits of the Law. Ed. DG Johnson and H Nissenbaum. Prentice Hall, 1995.

Gulcher, JR, et al. "Protection of privacy by third-party encryption in genetic research in Iceland." European Journal of Human Genetics 8.10 (2000): 739-742.

HIMSS. HIMSS Privacy and Security Toolkit. 2007.

<<http://www.himss.org/ASP/privacySecurityTree.asp?faid=78&tid=4>>.

Kodeswaran, P and Viegas E. "PINQ Analysis." TBD. TBD, 2009.

Korn, D. "Medical information privacy and the conduct of biomedical research." Academic Medicine 75 (2000): 963-968.

Langella, S, et al. "Sharing Data and Analytics Resources Securely in a Biomedical Research Grid Environment." Journal of the American Medical Informatics Association 15.3 (2008): 363-373.

Lee, D. "Differential Privacy: A User Case-Study." 19-20 October 2007. Carnegie Mellon Center for Computational Thinking MindSwap on Privacy.

<<http://www.cs.cmu.edu/~CompThink/mindswaps/index.html>>.

LeFevre, K, D DeWitt and R Ramakrishnan. "Multidimensional k-Anonymity." IEEE ICDE. Atlanta: IEEE, 2006.

Li, J, et al. "Current Developments of k-Anonymous Data Releasing." 2006.

Li, N, T Li and S Venkatasubramanian. "t-Closeness: Privacy beyond k-anonymity and l-diversity." IEEE 23rd International Conference on Data Engineering (ICDE 2007). Istanbul, Turkey: IEEE Computer Society, 2007. 106-115.

Lieberman, DA, et al. "of colon polyps detected by colonoscopy screening in asymptomatic black and white patients." Journal of the American Medical Association 300.12 (2008): 1417-1422.

Machanavajhala, A, et al. "l-diversity: Privacy beyond k-anonymity." TKDD 1.1 (2007): 1-21.

Malin, B. "An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future." Journal of the American Medical Informatics Association 12 (2005): 28-34.

Malin, B and L Sweeney. "How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems." Journal of Biomedical Informatics 37.3 (2004): 179-192.

Markle Foundation. "Model Privacy Policies and Procedures for Health Information Exchange." April 2006. Connecting for Health.

<http://www.connectingforhealth.org/commonframework/docs/P2_Model_PrivPol.pdf>.

Markle Foundation. "Policy Notice to Consumers." June 2008. Connecting for Health.

<<http://www.connectingforhealth.org/phti/docs/CP2.pdf>>.

McSherry, F and K Talwar. "Mechanism design via differential privacy." Foundations of Computer Science (FOCS). IEEE Computer Society, 2007. 94-103.

McSherry, F. "Privacy Integrated Queries." TBD (2009): TBD.

MSN. "Microsoft Online Privacy Notice." 2007. <<http://privacy.microsoft.com/>>.

Nissenbaum, H. "Privacy as a contextual integrity." Washington Law Review 79.1 (2004): 119-158.

Nissim, K, S Raskhodnikova and A Smith. "Smooth sensitivity and sampling in private data analysis." Symposium on Theory Computing (STOC). Association for Computing Machinery (ACM), 2007. 75-84.

Øhrn, A and L Ohno-Machado. "Using Boolean reasoning to anonymize databases." Artificial Intelligence in Medicine 15 (1999): 235-254.

OHSU Healthcare System (Administrative Policy Manual). "Permitted Uses & Disclosures of Protected Health Information (ADM 04.22)." 14 April 2004.
<<http://www.ohsu.edu/cc/hipaa/docs/04-22.shtml>>.

Ohno-Machado, L, S Vinterbo and S Drieiseitl. "Effects of Data Anonymization by Cell Suppression on Descriptive Statistics and Predictive Modeling Performance." Journal of the American Medical Informatics Association 9.6 (Supplemental 2002) (2002): S115-S119.

Safran, C, et al. "Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper." Journal of the American Medical Informatics Association 14.1 (2007): 1-9.

Shuchi, C, et al. "On the Utility of Privacy-Preserving Histograms." The 21st Conference on Uncertainty in Artificial Intelligence (UAI 2005). Edinburgh: Association for Uncertainty in Artificial Intelligence, 2005.

Su, TA and G Ozsoyoglu. "Controlling fd and mvd inferences in multilevel relational database systems." IEEE Transactions on Knowledge and Data Engineering 3 (1991): 474-485.

Sweeney, L. "Achieving k-anonymity privacy protection using generalization and suppression." International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10.5 (2002): 571-588.

Sweeney, L. "Guaranteeing anonymity when sharing medical data, the datafly system." Proceedings, Journal of the American Medical Informatics Association. Washington, DC: Hanley & Belfus, Inc, 1997.

Sweeney, L. "k-anonymity: a model for protecting privacy." International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10.5 (2002): 557-570.

United States Department of Health and Human Services. Medical Privacy - National Standards to Protect the Privacy of Personal Health Information. 2008. <<http://www.hhs.gov/ocr/hipaa/>>.

US Department of Health and Human Services (Office for Civil Rights). "HIPAA Administrative Simplification [Regulated Text]." 16 February 2006.
<<http://www.hhs.gov/ocr/AdminSimpRegText.pdf>>.

Vaccarino, V, TR Holford and HM Krumholz. "Pulse pressure and risk for myocardial infarction and heart failure in the elderly." Journal of the American College of Cardiology 36 (2000): 130-138.

Vijayan, J. "HIPAA audit: The 42 questions HHS might ask." 19 June 2007. Computer World: Security.
<<http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9025253>>.

Wellner, B, et al. "Retargetable Approaches to De-identification in Medical Records." Journal of the American Medical Informatics Association 14.5 (2007): 564-573.

Woolson, RF. Statistical Methods for the Analysis of Biomedical Data. John Wiley & Sons, Inc., 1987.

Xiao, X and Y Tao. "M-invariance: towards privacy preserving re-publication of dynamic data sets." SIGMOD Conference. ACM, 2007. 689-700.

Zerhouni, EA. "A Vision for Transforming Medicine in the 21st Century." 2006. National Institutes of Health. <<http://www.nih.gov/about/director/slides/vision.pdf>>.

Appendix

Appendix 1: Differential Privacy concept

The concepts of differential privacy are based on the research published by Cynthia Dwork and Frank McSherry (Chawla, Dwork and McSherry, On Privacy-Preserving Histograms; Blum, Dwork and McSherry; McSherry, Privacy Integrated Queries). Described here are both the sanitization concept and the privacy mechanism theorem associated with it (Lee).

Sanitization Concept

To protect the individuals comprising the data, individuals are masked within the data by creating a "sanitization point" between the user interface and the data (Figure 12). The green shape on the left in this figure represents the database while the right-most circle represents the answer or result set provided to the user interface. The middle section is the interactive sanitizer, defined as \mathcal{K} , which introduces random noise to produce uncertainty, hence privacy.

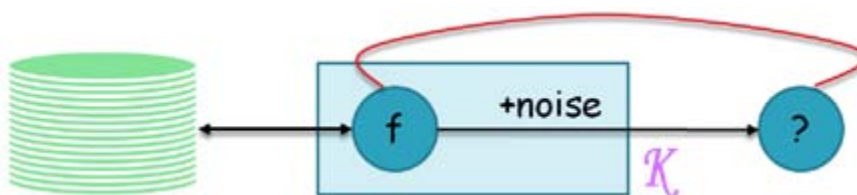


Figure 12: Sanitization Concept

Note the magnitude of the noise is given by the theorem:

If many queries f_1, f_2, \dots are to be made, noise proportional to $\sum_i \Delta f_i$ suffices. For many sequences, we can often use less noise than $\sum_i \Delta f_i$. Note that $\Delta \text{Histogram} = 1$, independent of number of cells

The sanitizer requires the creation of a carefully detailed algorithm and will be discussed in more detail in the differential privacy section below. A safe answer on the amount of noise to apply is that of a standard deviation equal to the total number of queries. If the number of queries is not known, then the standard deviation can be proportional to the square of the queries asked so far. As for the noise itself, it should be newly seeded each time a query is applied.

By doing this, this algorithm will be able to address all attacks. Consequently, for each person, the increase in probability of the individual being attacked (or anyone else for that matter) due to the contribution of their data is nominal. The example given is foiled for two reasons: a) the addition of noise will (formally) complicate the polynomial reconstruction and b) the number of queries is limited by the degree of privacy guaranteed, and n is generally going to be way too many queries.

Mathematically, at this sanitization point, the differential privacy algorithm will apply a little bit of noise within each cell of the histogram. Descriptively, if one's result set from the database is a report with a set of rows and columns, for each value within each row and column cell is a small bit of error added to the original number. Provided that the sanitization point can limit the number of questions being asked and/or add more noise as more questions are being asked, then the differential privacy algorithm can guarantee the privacy of the individuals that make up these aggregates.

Differential Privacy

Our privacy mechanism, K , gives ϵ -differential privacy for all transcripts t , all databases DB , and all data items (rows in the DB) Me , the ratio

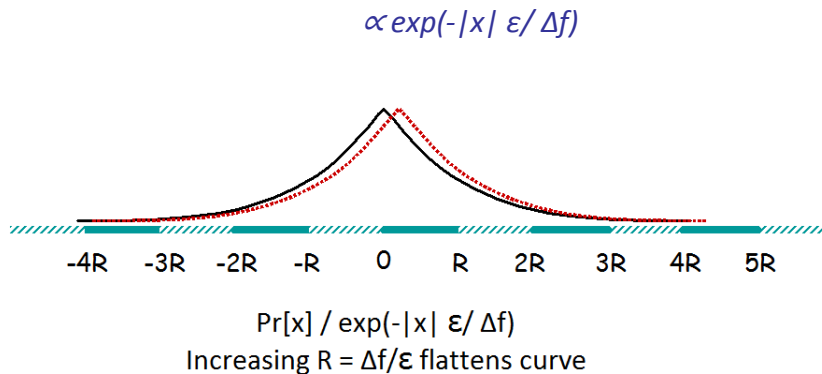
Nothing more about an individual can be learned when her information is in the dB than when it is not in the dB .

$$\frac{\Pr[\mathcal{K}(DB - Me) = t]}{\Pr[\mathcal{K}(DB + Me) = t]} = e^\epsilon$$

And to achieve differential privacy....

$$\Delta f = \max_{DB, Me} |f(DB+Me) - f(DB-Me)|$$

The differential privacy theorem is a simple mathematical formula that describes the fact that nothing more can be learned about an individual when her information is in the DB ($DB+Me$) than when it is not in the DB ($DB-Me$). This is important from the context of joining one set of data that is in the DB (e.g., the masked medical data) with another set of data that is not in the DB (e.g., Cambridge, MA voter list). If there is no noticeable difference between $f(DB-Me)$ and $f(DB+Me)$ then there is no perceptible risk by joining the two data sets together. Therefore, to achieve this differential privacy, we will need to add scaled symmetric noise



Note that possible responses, R , is defined as $R = \Delta f / \epsilon$. The black line represents the $f(DB-Me)$ while the red line represents the $f(DB+Me)$. This means that increasing the value of R will flatten the curve; the flatter the curve the more privacy is provided. The flatter curve means that no response is much more likely in one case in comparison to other.

The question that is incurred then is what difference must noise obscure or how much can $f(DB+Me)$ exceed $f(DB-Me)$. This goes back to the above noted formula of differential privacy:

$$\Delta f = \max_{DB, Me} |f(DB+Me) - f(DB-Me)|$$

where the noise depends on the sensitivity of the system.

Figure 13 shows the real value is $f(DB-Me) = 104$ while the added noise value of $f(DB+Me) = 105.2$. The statistical difference between these two values seems insignificant, but this depends on the sensitivity of one's statistics. If you have highly sensitive data where a 1.2 difference results in a \$1.2 billion sale price, then any change to the data is unacceptable. While if your data is relatively insensitive (and most data analyses are), this slight error of difference is well within an acceptable error margin. A key presumption of this design is that one will need to calibrate the noise to ϵ and the number of queries. More explicitly to this latter point, as the number of queries increase, the more noise will need to be added to ensure that the noise algorithm itself can counter an attack. The vast majority of analytical queries do not require a substantial increase in noise due to the relatively low number of questions asked. As seen in Figure 14, noise is independent of the database size so privacy is insured but accuracy varies; the larger the database the higher the accuracy.

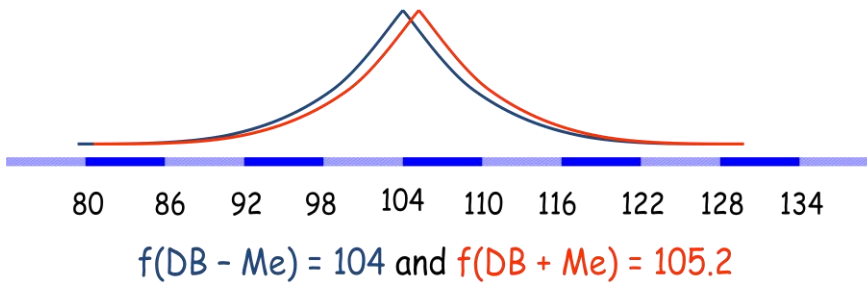


Figure 13: Example of noise and sensitivity

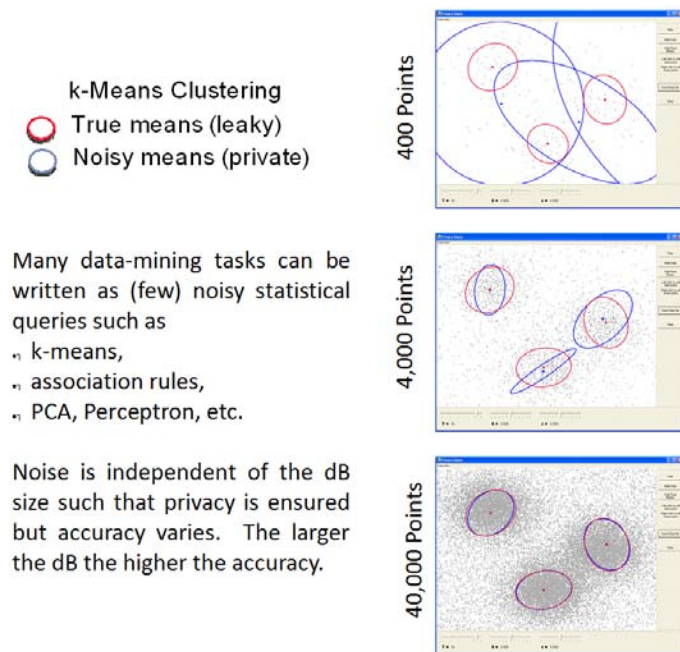


Figure 14: Differential privacy can be applied to many different statistical techniques

Generating the Noise

In many types of reporting solutions, we are primarily concerned with counts and summations. Specifically, these reporting systems involve unique visitor counts, event transaction counts, and page view summations. As we are dealing with discrete data, instead of continuous data, a basic differential privacy histogram algorithm to implement is:

$$prg(R, Seed\ for\ A) = (n_1, n_2, \dots, n_n)$$

Where

<i>prg()</i>	This is a pseudo-random number generator (RNG) such as <i>SRAND</i>
R	This is a magnitude of the noise to be applied to the system. The larger the number, the more error is added. More error means more privacy but could skew the meaning of the numbers if the value is too high.
<i>Seed for A</i>	This is the seed that is used generate an array of numbers where A is the original query.
<i>n₁, n₂, ...</i>	An array of numbers that will alter the counts, in effect adding noise.

The routine, *prg()*, is based on the property that for any interval $[x, z]$ of the density function

$$p(x) \cong e^{\left(\frac{|x|}{R}\right)}$$

we can analytically find the point y such that $p([x, y]) = p([y, z])$. A single random bit can tell us on which side of y our sample should be drawn, and we can recursively apply the technique to the appropriate subinterval (i.e.: either one of $[x, y]$ or $[y, z]$).

To generate the noise, the RNG will create a stream of numbers, for example:

0	0	1	1	1	...	1	0	0	0	0	1
---	---	---	---	---	-----	---	---	---	---	---	---

The resulting exponential distribution translation of this stream is:

-	.	2	+	1	...	+	6
---	---	---	---	---	-----	---	---	---	---	---	---	---

The generated noise will then be applied to the result set. The generation of noise uses the above RNG (there is no such thing as a true random number generator in mathematics) to create the stream of 0s and 1s. We then take the stream of numbers and translate it into the noise. For example, if you look at the right most set of numbers (starting after the ...) within the array, you'll notice 1, then 5 0s, then 1 again. The first 1 denotes a positive value (vs. negative), the five 0s represent a count of 5, and the final 1 represents the full stop. Hence, these seven values represent a positive 5 count + 1 count with full stop equaling +6.

Effect on Data

This stream of numbers is then applied directly to the results. Recall, that the stream of noise values was -2, +1, ..., +6 and the original values are that of A = 36, B = 22, ..., N = 102 as noted in Figure 15.

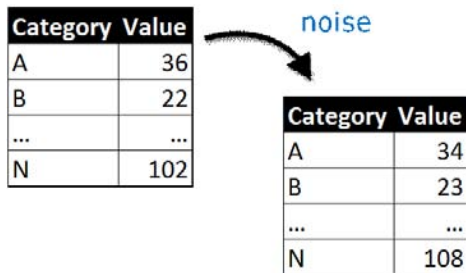


Figure 15: Effect of noise

The noise added to the system provides a new result of $A = 34, B = 23, \dots, N = 108$. By adding this noise to the system, it is not possible for one to drill down to a value of 1, which would expose a single individual (e.g., $A = 1$). After all, if you are able to drill down to one individual, even through what is perceived as non-identifiable attributes (e.g., birth date, zip code, and gender); it becomes possible to identify this individual. But the additional noise will never allow you to get down to a single individual of any attributes (or combination thereof). If the numeric answer to your question represents a single individual (e.g., $A = 1$), one is not sure if this truly represents an individual or if this is due to the noise applied. Note the definition of privacy is much more complex mathematically than the ability to drill down to one person; this view provides a simple analogy.

An issue of concern is if we apply too much noise (e.g., +1000) then we risk changing the meaning of the data completely. If too little noise (e.g., +0, +1, -0, -1, etc.) is applied we risk not protecting the data very well. Hence, we followed the safe rule that the magnitude of the noise applied to the RNG is the standard deviation of the total number of queries. In our case, the total number of queries per user ranged from 90 to 110 queries. Therefore, the standard deviation was 10 and the magnitude applied to the RNG, R , is 10. We limited the number of questions asked of the system otherwise it may have become possible to break the noise pattern. Since the noise generated by the RNG requires some seed value (i.e., a starting point), if you ask enough questions it becomes mathematically possible to determine how the RNG is generating its values. Once this is determined, it becomes possible to reverse engineer the system and determine the original set of values. Hence the importance of either limiting the number of questions asked (e.g., a hundred questions) or adding more noise to the system as more questions are being asked thus preventing an attack of this nature.

The effect on the customer data by the application of noise is reflected in the tables below. In Figure 16, the top and bottom tables are two of the same queries executed a few seconds of each other with the exponential noise applied.

Country	Unknown	Very Low	Low	Moderate	High
afghanistan	121563	11280	3851	3984	18109
albania	557376	70421	30896	30291	117331
algeria	444663	50614	14927	14949	47312
american samoa	36962	3370	1148	1130	5610
andorra	30567	4143	1540	1515	7763
angola	71292	4661	1837	1908	7076
anguilla	9003	979	415	490	2479
antarctica	26339	2551	835	910	4377
antigua and barbuda	20539	2389	1030	1047	5416
argentina	13099828	1299322	498624	490461	1865684
armenia	48843	4212	1446	1485	5617
aruba	52746	5955	2345	2331	12384

Country	Unknown	Very Low	Low	Moderate	High
afghanistan	121561	11279	3853	3983	18108
albania	557369	70419	30895	30292	117330
algeria	444668	50614	14927	14947	47312
american samoa	36962	3375	1148	1128	5615
andorra	30567	4146	1540	1516	7763
angola	71290	4658	1836	1909	7074
anguilla	9002	980	415	491	2475
antarctica	26340	2548	835	910	4377
antigua and barbuda	20544	2391	1029	1047	5414
argentina	13099829	1299324	498624	490459	1865683
armenia	48840	4213	1441	1486	5619
aruba	52744	5953	2346	2327	12386

Figure 16: Effect on data with the application of exponential noise for the same query; notice that {Afghanistan, Unknown} is 121563 for one execution and 121561 for the second time the same question is asked.

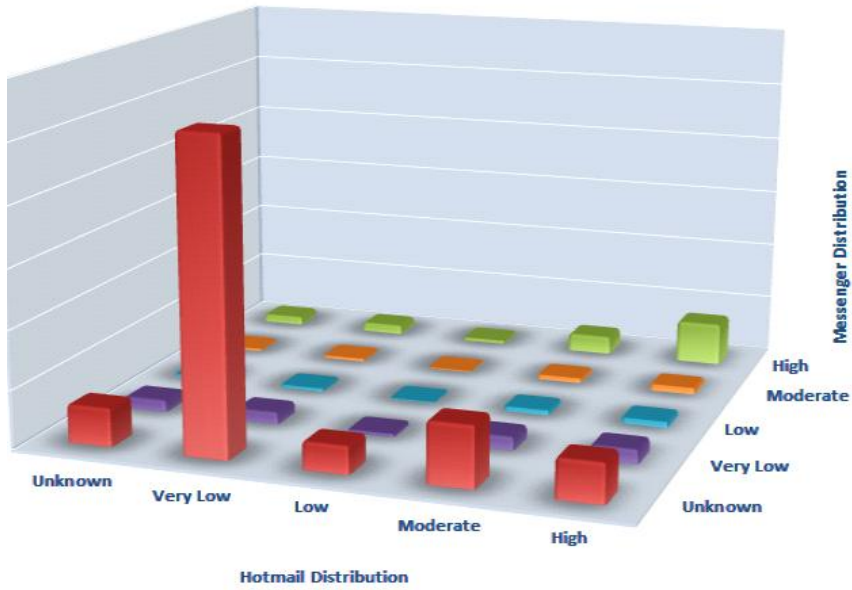


Figure 17: Graphical Effect on data

For example, reviewing the country Afghanistan, the “Unknown” value is 121563 in one case and 121561 in another. Because of this random exponentially distributed noise, we do not know what the “real” value is. The visual effect to this data is minimal, as depicted in Figure 17.

Appendix 2: Application of Differential Privacy Case Study

A case study was previously performed where these algorithms were applied to Microsoft Network (MSN) reporting data. MSN is one of the largest web portals with more than 2 billion page views a day offering advertiser-paid services including Search, Messenger, Live Mail, among many other services. This case study used MSN visitor data where the privacy preserving noise algorithms were applied to the data to determine reporting efficacy. MSN has 550 million Passport users (as of February 2006) and contains visitor self-reported data such as gender, birth date, occupation, country, and zip code. The web traffic data associated with MSN also has additional attributes including IP address, pages viewed, page view duration, browser, and operating system. MSN uses this data to provide customizable experiences of their users and to better understand how visitors are using the various services. At the same time, there are various major privacy issues including identity theft, fraud, and/or bad press (e.g., AOL released search engine queries that ended up revealing their users). If user expectations are not satisfied, customers will no longer trust the services provided. As data is accumulated, it becomes easier to segment the population and potentially identify individual users without directly using personally identifiable information. Hence, there was an interest within MSN to use the MSR research to determine the latter's applicability for all reporting.

To test the applicability of these privacy-preserving techniques within a reporting environment, two different analytics groups analyzed MSN visitor data. The "Sampled Users Web Analytics" group is a set of users who wanted to understand what MSN web sites people were using (e.g. the MSN home page, Moneycentral, MSN Autos, etc.). The "Customer Churn analytics" group is a set of users who were trying to better understand how customers are using the features of MSN (e.g. Messenger, Search, Hotmail, etc.). A considerable amount of time was spent with


both groups to determine the reporting features and the metrics they wanted. After determining this, an OLAP database built on Microsoft© SQL Server Analysis Services 2005 was created so these groups could have a multi-dimensional view of their data with fast response times. Once the groups were happy with the reports, the differential privacy algorithm was introduced in the form of a privacy preserving histogram (PPH). This allowed the groups to view both original reports and PPH-applied reports and to express any issues perceived in comparing the two report types.

[Sampled Users Web Analytics Group](#)

The “Sample Users Web Analytics Group” wanted to understand what visitors were using on the MSN.com and Windows Live web sites. The concept of PPH was introduced to this group after building a new reporting solution on an existing web analytics solution specific to their business requirements. In addition, an Analysis Services database was built for this new reporting solution to allow the group the ability to view the data from multiple dimensions – something that they had desired for quite some time.

Unfortunately, the resulting feedback was negative; the basic problem was that this group could not accept any amount of error in their data. For example, an initial query provided the number of visitors per country, where the total US visitors was 202 (Figure 18). A subsequent query against the data which broke out the same US visitors by gender would result in a total that was 203. While it could be understood that the added noise to the system could explain these differences, this group insisted that these numbers had to match at all times. Note, it did not matter that these numbers were not used for financial reconciliation and was only to be used for analysis purposes.

Country	Visitors
United States	202
Canada	31



Country	Gender	Visitors
United States	Female	128
	Male	75
	Total	203
Canada	Female	15
	Male	15
	Total	30

Figure 18: Totaling errors

From further discussions with this group, it appeared that they had been utilizing reporting systems that had perceived accuracy issues. It is this perception that resulted in the groups' negative feedback to additional noise being added to the resulting data.

Customer Churn Analysis Group

To provide the reports that this group desired, we built a brand new reporting solution. The customers were familiar with this data because it was built on an existing targeted marketing system. Similar to the "Sampled Users Web Analytics Group," a SQL Server 2005 database was created to initially filter and transform the data. Upon processing completion, an OLAP cube and a custom web UI was built on this data to provide the reporting interface. This new reporting system allowed the analyst to understand how MSN services (Messenger, Mail, Search, Spaces, etc.) were being used. Multiple iterations were also performed with this customer group insuring they had received the data they desired. A key difference between the groups is while the first group was able to use legacy reporting systems; the second group did not have access to any reporting. Within a few weeks of their initial request this group was able

to interact, validate, and provide feedback on a working reporting solution. Most importantly, they were able to analyze and re-analyze the data to determine the precision and accuracy of the data.

Once exponential noise was introduced into their churn reports (with the additional noise being seen in the result set), this group was satisfied with the results. After some use, they seemed to forget that a privacy algorithm was applied – even though the numbers in the reports were changing (due to the additional noise) and the statement “Privacy Preserving Histogram Applied” was applied directly into the reports. Based on conversations and surveys, it was very apparent that the collaborative effort in creating these reports led to the customer trusting the data – this is a key difference in comparison to the first group. Because of this trust, the small amount of error introduced into the system to ensure customer privacy was well within a tolerable error margin. It also helped that this group is a direct marketing group, so they were more familiar with the concept of customer privacy.

Appendix 3: C# Code Example

Below is a C# code example using the PINQ.dll; it reads the tab-delimited CoriData.txt data and provides the differential privacy applied answer.

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.IO;

using PINQ;

namespace PINQ_Skeleton
{
    class Program
    {
        // Reads a text file via the StreamReader
        public static IEnumerable<string> ReadFile(string filename)
        {
            StreamReader file = new StreamReader(filename);
            while (!file.EndOfStream)
                yield return file.ReadLine();

            file.Close();
        }

        // Logs data and denotes privacy increments
        public class PINQAgentLogger : PINQAgent
        {
            double total;

            public override bool apply(double epsilon)
            {
                total += epsilon;
                Console.WriteLine("*** privacy change **\tincrement: " + epsilon + "\t new total: " + total);

                return true;
            }
        }

        // Main class
        static void Main(string[] args)
        {
            // Create Querable connection to file
            var source = ReadFile(@"..\..\..\CoriData.txt").AsQueryable();
        }
    }
}
```



```

// LINQ sample to query the raw data; i.e. not PING applied
//   Filtered by: routine/avg risk screening
//   Grouped by: Outcome, Ethnicity
//   NOTE: this code below should be commented out; this provides an example of how to use LINQ
var RawQuery = source.Select(x => x.Split('\t'))
    .Select(x => new
    {
        Indication = x[0],
        Ethnicity = x[1],
        AgeCategory = x[2],
        AgeOver60 = x[3],
        Gender = x[4],
        SiteType = x[5],
        Outcome = System.Convert.ToInt16(x[6])
    })
    .Where(x => x.Indication.Equals("Routine/Avg Risk Screening"));

// Load Group Query
var GroupQuery = RawQuery
    .GroupBy(w => new { w.Ethnicity, w.Outcome })
    .Select(g => new { Grouping = g.Key, NumOfPatients = g.Count() });

// Output
Console.WriteLine("Original Raw Data");
Console.WriteLine("Ethnicity/Outcome, Number of Patients");
int QueryTotal;
QueryTotal = RawQuery.Count();
foreach (var obj in GroupQuery)
{
    Console.WriteLine("{0}, {1}", obj.Grouping, obj.NumOfPatients);
}
Console.WriteLine("Total, {0}\n\n", QueryTotal.ToString());

//
// PING sample to query the raw data with PING applied
//   Filtered by: routine/avg risk screening
//   Grouped by: Outcome, Ethnicity
//
// Provides PING-applied total count
var RawQuery = source.Select(x => x.Split('\t'))
    .Select(x => new
    {
        Indication = x[0],
        Ethnicity = x[1],
        AgeCategory = x[2],
        AgeOver60 = x[3],
        Gender = x[4],
        SiteType = x[5],
        Outcome = System.Convert.ToInt16(x[6])
    })
    .Where(x => x.Indication.Equals("Routine/Avg Risk Screening"));

```

```

        })
        // Routine/Avg Risk Screening
        // -----
        .Where(x => x.Indication.Equals("Routine/Avg Risk Screening")
            & x.Ethnicity.Equals("White non-Hispanic") & x.Outcome == 1);
        // .Where(x => x.Indication.Equals("Routine/Avg Risk Screening")
        //     & x.Ethnicity.Equals("White non-Hispanic") );
        // .Where(x => x.Indication.Equals("Routine/Avg Risk Screening")
        //     & x.Ethnicity.Equals("Black non-Hispanic") );
        // .Where(x => x.Indication.Equals("Routine/Avg Risk Screening")
        //     & x.Ethnicity.Equals("Black non-Hispanic") & x.Outcome == 1);

int RawQueryCount = RawQuery.Count();

// Open new connection to the file
var psource = new PINQueryable<string>(ReadFile(@"..\..\..\CoriData.txt").AsQueryable(), null);

// Load protected data
var pRawQuery = psource.PINQSelect(x => x.Split('\t'))
    .PINQSelect(x => new {
        Indication = x[0],
        Ethnicity = x[1],
        AgeCategory = x[2],
        AgeOver60 = x[3],
        Gender = x[4],
        SiteType = x[5],
        Outcome = System.Convert.ToInt16(x[6])
    })
    // -----
    // Routine/Avg Risk Screening
    // -----
    .PINQWhere(x => x.Indication.Equals("Routine/Avg Risk Screening")
        & x.Ethnicity.Equals("White non-Hispanic") & x.Outcome == 1);
    // .PINQWhere(x => x.Indication.Equals("Routine/Avg Risk Screening")
    //     & x.Ethnicity.Equals("White non-Hispanic") );
    // .PINQWhere(x => x.Indication.Equals("Routine/Avg Risk Screening")
    //     & x.Ethnicity.Equals("Black non-Hispanic") );
    // .PINQWhere(x => x.Indication.Equals("Routine/Avg Risk Screening")
    //     & x.Ethnicity.Equals("Black non-Hispanic") & x.Outcome == 1);

//Console.WriteLine("argument: {0}", args[0]);
double epsilon;
epsilon = System.Convert.ToDouble(args[0]);
for (int j = 0; j < 30; j++)
{
    Console.WriteLine("{0}, {1}, {2}", epsilon, RawQueryCount, pRawQuery.NoisyCount(epsilon));
}
}
}
}

```

To simplify the code, the code was re-compiled for the different categories of:

- Routine / Avg Risk Screening, White Non-Hispanic, Outcome = 1 (Advanced Neoplasia)
- Routine / Avg Risk Screening, White Non-Hispanic, Outcome = 0 (Not Advanced Neoplasia)
- Routine / Avg Risk Screening, Black Non-Hispanic, Outcome = 1 (Advanced Neoplasia)
- Routine / Avg Risk Screening, Black Non-Hispanic, Outcome = 0 (Not Advanced Neoplasia)

Upon compiling the code, execution of the executable from the command line was in the form of:

```
"PINQ Skeleton.exe" 0.001
```

where the first parameter is the desired epsilon value. Note within the code sample that the final for loop traverses thirty times, each time executing the `.NoisyCount()` method. The input epsilon value from the command line is the main parameter of the `.NoisyCount` method which then provides the answer with differential privacy applied at the provided epsilon value. Sample output of this execution can be seen below.

```
0.001, 18, 44.344049528914
0.001, 18, 1881.86559796183
0.001, 18, -315.785133749554
0.001, 18, 2106.74146919961
0.001, 18, 651.228582546968
0.001, 18, -37.073718748642
0.001, 18, -268.026476956296
0.001, 18, -445.898305492377
0.001, 18, 363.859343247873
0.001, 18, 812.100705873035
0.001, 18, -52.9857873009428
0.001, 18, 2300.44337566087
0.001, 18, 1126.80854873549
```

An additional note, if you use the `PINQAgentLogger ()` function for the `PINQueryable`

```
var protectedQuery = new PINQueryable<string>(source, new PINQAgentLogger());
```

then PINQ will provide the amount of information that is revealed. For example, when the logger notes that the privacy change increment is 111.11, this means that 111.11 total privacy units have been revealed. The lower the number of privacy units, the less information is revealed and the higher the privacy.