Identifying inference attacks against healthcare data repositories

Jaideep Vaidya, PhD1, Basit Shafiq, PhD2, Xiaqian Jiang, PhD3, Lucila Ohno-Machado, MD, PhD3

1 Rutgers University, Newark, NJ, USA; 2 Lahore University of Management Sciences, Lahore, Pakistan; 3 University of California at San Diego, San Diego, CA, USA

Abstract

Health care data repositories play an important role in driving progress in medical research. Finding new pathways to discovery requires having adequate data and relevant analysis. However, it is critical to ensure the privacy and security of the stored data. In this paper, we identify a dangerous inference attack against naive suppression based approaches that are used to protect sensitive information. We base our attack on the querying system provided by the Healthcare Cost and Utilization Project, though it applies in general to any medical database providing a query capability. We also discuss potential solutions to this problem.

Introduction

Health Information Technology is critical to driving progress in medical research, and providing better healthcare while reducing its costs. Finding new pathways to discovery requires having adequate genomic, transcriptomic, clinical, behavioral, and social data, and analyzing them in ways relevant to health. The increasing digitization of health information makes it possible to perform such analysis and realize the intrinsic value of such data. However, without appropriate controls, digitization magnifies the risk to privacy, due to the ease of retrieval, analysis, and linkage.

Privacy and confidentiality are critical to healthcare. Many illnesses and treatments have inherited stigma, due to which patients might be reluctant to participate in treatments if their privacy concerns are not sufficiently addressed. On the other hand, improving privacy protection encourages people and organizations to share data and realize hidden insights. For example, orphan diseases can be treated more effectively when more observations from different regions in the world are shared and aggregated. Personalized medicine can be targeted to individuals more accurately if more patients similar to the person of interest are observed and analyzed.

One mechanism through which researchers generate a hypothesis is through exploration of the data. For this, it is necessary to query the data, and to compute different statistics from it. However, this opens the door to potential data disclosure. To alleviate this, medical data repositories often suppress data that may be used to uniquely identify individuals. However, this is often insufficient to protect confidentiality. Indeed, in this paper, we identify a new query inference attack that can breach existing safeguards and can be used to iteratively reveal personally identifiable information (PII). It is important to note that our attack does not require the matching of another dataset in order to produce breach of privacy, and can be fully realized through the exposed query interface.

To illustrate the attack and the problems it creates, we target the HCUPnet – a free, on-line query system based on data from the Healthcare Cost and Utilization Project (HCUP). HCUP, is a family of health care databases and related software tools and products developed through a Federal-State-Industry partnership and sponsored by the Agency for Healthcare Research and Quality (AHRQ), one of the 12 agencies with the US Department of Health and Human Services. HCUP data has been used for several medical studies1,2, and access to such data is critical for medical progress. One of the key tools produced by HCUP is the HCUPnet, a free, on-line query system that provides instant access to health statistics and information on hospital inpatient and emergency department utilization. Given that the HCUPnet allows querying of data to anyone with an internet connection, to protect privacy, even when publishing statistical data, HCUPnet suppresses values that can lead to individually identifiable information. For example, for state databases, all values that are based on less than 10 tuples are suppressed, and instead a * is published1. However, this only gives a false sense of privacy, without really protecting it. Query Inference attacks can be used to breach the privacy protection and learn significantly more information that should have been protected. In the next section, we

1 The formal notice reads: Values based on 10 or fewer discharges or fewer than 2 hospitals in the State statistics (SID) are suppressed to protect confidentiality of patients and are designated with an asterisk (\textasteriskcentered)}
describe this query inference attack with a specific example. Following that, we describe potential solutions to this problem, and then conclude the paper.

Query Inference Attack

In this section, we first go through what a typical query to HCUPnet looks like and what is the output that is provided by HCUPnet. Following that, we describe how to mount a query inference attack using the reported data. As described earlier, HCUPnet is a free, on-line query system that provides instant access to health statistics both at the national level and at the state level for participating states. While it is possible to get general information on all hospital stays, one can also search by information on specific diagnoses/conditions and surgeries/procedures. For example, it is possible to search for hospital stays resulting from cancer of ovary in New Jersey in the year 2010. Several outcomes and measures can be selected for which statistics are required such as number of discharges, mean length of hospital stay (days), mean hospital charges (dollars), mean hospital costs, and percent of patients who died in the hospital. It is also possible to further report the results based on patient and hospital characteristics such as patient age, gender, insurance coverage (Medicare, Medicaid, private, uninsured, other), location of patient’s residence (large central metro, suburbs, medium or small metro, and non-metro), race/ethnicity, hospital ownership (public, for-profit, not-for-profit), hospital teaching status (teaching vs. not), hospital location (metropolitan vs. non-metropolitan), and hospital bedsize (small vs. medium vs. large).

For example, Figure 1 shows the results of a query asking for the the number of patients discharged by New Jersey hospitals in the year 2009, whose principal diagnosis was ovarian cancer. While we only show the results tabulated by age-group and by ethnicity, HCUP also releases the other tabulations by gender, payer (insurance status), hospital location, ownership, teaching status (of the hospital), and bedsize. Note that there were only 735 such patients (all female). Note that all of the *s (highlighted in yellow) represent the suppressed information where the number of patients who would contribute to the values to be reported are less than 10.

| Race/ethnicity | Native American | Other | Missing |
|----------------|-----------------|-------|---------|
| Total # discharges | 18 | 19 | 22 |
| Age group | White | Black | Hispanic | Asian/Pacific Islander | Native American | Other | Missing |
| 1-17 | 535 | 82 | 56 | 18 | | 19 | 22 |
| 18-44 | 40 | 13 | * | * | * | * | * |
| 45-64 | 31 | 32 | * | * | * | * | * |
| 65-84 | 35 | 13 | * | * | * | * | * |
| 85+ | 29 | * | * | * | * | * | * |
| Total number of discharges | 735 | 535 | 82 | 58 | 18 | * | 19 | 22 |
| Mean Costs | 15,101 | 14,835 | 18,903 | 16,006 | 15,628 | * | 20,774 | 9,813 |

Figure 1: HCUPnet released data

However, based on the data released (even with suppressions) it is possible to infer more information. A malicious user can pose multiple queries and correlate the results to obtain bounds on the suppressed data. For example, from Figure 1 the query requestor knows that there are 735 total discharges. These are further subdivided as 535 White, 82 Black, 58 Hispanic, 18 Asian/Pacific Islander, 19 Other, 22 Missing. Based on this, the user can easily infer that the number of Native American discharges (which is suppressed) is $735 - (535 + 82 + 58 + 18 + 19 + 22) = 735 - 734 = 1$.

Now, consider the data about mean costs reported by HCUPnet for the same group of people split up by ethnicity.

| Race/Ethnicity | Total | White | Black | Hispanic | Asian/Pacific Islander | Native American | Other | Missing |
|----------------|-------|-------|-------|----------|------------------------|-----------------|-------|---------|
| Total number of discharges | 735 | 535 | 82 | 58 | 18 | * | 19 | 22 |
| Mean Costs | 15,101 | 14,835 | 18,903 | 16,006 | 15,628 | * | 20,774 | 9,813 |

Considering this additional information, it is trivial to figure out that the mean costs for the Native American (which are also suppressed) are $32,970 (simply by subtracting the mean costs \times the number of discharges for all other types from the overall costs).
As a matter of fact, significantly more information can be extracted even from a single query. Consider the query shown in Figure 1. It is possible to formulate several mixed integer programming problems to get more specific bounds on the remaining suppressed values. For example, in Figure 2, the variables $x_{11} - x_{16}, x_{21} - x_{25}, x_{31} - x_{33}, x_{41} - x_{44}, x_{51} - x_{56}$ represent the suppressed values (highlighted in yellow).

Each row provides a constraint – for example, since there are 70 patients in the age-group 18-44, out of which 40 are white, and 13 are black, the total number of remaining patients is $70 - 40 - 13 = 17$. Now, since $x_{21}, x_{22}, x_{23}, x_{24}, x_{25}$ represent the number of patients in the remaining ethnic groups aged 18-44, the following constraint can be formulated: $x_{21} + x_{22} + x_{23} + x_{24} + x_{25} = 17$. Similarly, each column also provides a constraint. For example, Figure 2 shows that there are 58 Hispanic patients, out of which 32 are aged 45-64, and 13 are aged 65-84 (implying that $58 - 32 - 13 = 13$ patients are in the remaining age-groups). Since $x_{12}, x_{21}, x_{52}$ represent the number of Hispanic patients aged 1-17, 18-44, 85+, respectively, the following constraint can be formulated: $x_{12} + x_{21} + x_{52} = 13$. Once all such constraints are formulated, the bounds on each variable can be deduced by solving 2 optimization problems, the first asking for the minimum value (for example, $\min x_{11}$) for the variable of interest ($x_{11}$, in this case) satisfying all constraints, while the second asks for the maximum value (e.g., $\max x_{11}$). Table 1 shows an example of the integer programming problem for determining the bounds of $x_{11}$ (the number of Black Patients discharged in the age group 1-17) based on the data in Figure 1. Since these problems are mixed linear integer programming problems, standard branch and bound techniques can be used to solve them.

### Table 1: Inferring the bounds of $x_{11}$

| Total discharges | White | Black | Hispanic | Asian/Pacific Islander | Native American | Other | Missing |
|------------------|-------|-------|----------|------------------------|-----------------|-------|---------|
| 735              | 535   | 82    | 56       | 13                     | 18              | 19    | 22      |
| Age group        |       |       |          |                        |                 |       |         |
| 1-17             | 3     | 1     | $x_{11}$ | $x_{12}$ [0-2]         | $x_{13}$ [0-1]  | $x_{14}$ [0] | $x_{15}$ [0-1] | $x_{16}$ [0-1] |
| 18-44            | 70    | 40    | 13       | $x_{21}$ [2]          | $x_{22}$ [2]    | $x_{23}$ [0] | $x_{24}$ [0-1] | $x_{25}$ [1-8] |
| 45-64            | 330   | 236   | 31       | 32                     | $x_{31}$ [10]   | $x_{32}$ [0] | 11       | $x_{33}$ [10] |
| 65-84            | 298   | 229   | 35       | 12                     | $x_{41}$ [2-8]  | $x_{42}$ [1] | $x_{43}$ [2-8] | $x_{44}$ [4-10] |
| 85+              | 34    | 29    | $x_{51}$ | $x_{52}$ [1-4]         | $x_{53}$ [0-1]  | $x_{54}$ [0] | $x_{55}$ [0-1] | $x_{56}$ [0-1] |

Figure 2: HCUPnet inferred data

We used lp_solve to solve these to get the bounds depicted in Figure 2. For example, the bound for $x_{13}$ was inferred to be [0-1], which means that there is at most 1 Asian/Pacific Islanders of age 1 – 17 with ovarian cancer. Using a separate query that tabulates the mean age of the patients with respect to the categorization by ethnicity, we were also...
able to infer that the age of the Native American woman was 75.

The damage through correlating additional queries can be enormous. For example, based on several additional queries, we are able to figure out that exactly 1 Native American woman diagnosed with ovarian cancer went to a privately owned, not for profit, teaching, hospital with more than 435 beds in 2009. Furthermore, the woman did not pay by private insurance, had a routine discharge, with a stay in the hospital of 33.5 days, with her home residence being in a county with 1 million plus residents (large fringe metro, suburbs), and her age was 75 years. As detailed earlier, we can even figure out that the woman’s hospital costs (which are also suppressed) are exactly $32,970. Note that Sweeney identified that more than 87% of the US population can be uniquely identified with their gender, date of birth, and zip-code. In this case, we have the gender, the age, and an indication of the zip-code based on the NJ state counties with more than 1 million residents. Clearly, the possibility for a significant breach of privacy exists for a wide swath of people.

Potential Solutions

It should be clear that naïve suppression does not prevent privacy breaches. Indeed the problem of query inference in statistical databases has been identified for over two decades. More sophisticated solutions are necessary to appropriately protect medical information in biomedical data repositories. One potential solution is through query auditing. Query auditing keeps track of the information revealed through queries. When done in an online fashion, submitted queries are checked in a continuous manner, to prevent inference disclosure. This is accomplished by combining responses to past queries with the response to the current query to determine if a breach occurs by responding to the new query. Given a privacy policy, it is now possible to determine if a new query can lead to inference of private information, and should be denied. As a matter of fact, query denials can also leak information. Techniques such as simulatable auditing have been proposed to account for this information leakage. However, they significantly lower the utility of data, by proactively denying a larger number of queries. Recently, we have proposed techniques for non-deniable auditing that are more efficient and provide more data utility with the same privacy cost. This is done by employing the classic primal and dual simplex algorithm on bounded variables (from optimization analysis) to constantly monitor the bounds of protected values. Query denials are addressed by taking into account the scope of the query’s real answer which can be obtained by cleverly constructing and solving a set of parametric linear programming (LP) problems. However, some of these approaches are restricted in the kinds of queries they can be applied to. Nevertheless, they can serve as a starting point for future research.

Another potential solution is through the use of more formal privacy protection models such as Differential Privacy. Differential privacy provides a formal and quantifiable privacy guarantee irrespective of an adversary’s background knowledge and available computational power and is actually a condition on the data release mechanism and not on the dataset. A randomized algorithm is differentially private if for any pair of neighboring inputs, the probability of generating the same output, is within a small multiple of each other. This means that for any two datasets which are close to one another, a differentially private algorithm will behave approximately the same on both data sets. This notion provides sufficient privacy protection for users regardless of the prior knowledge possessed by the adversaries. In the context of supporting queries such as those provided by HCUPnet, a potential solution would be to reveal differentially private counts as opposed to the true counts. This would make it infeasible to accurately pose the optimization queries described earlier since the contribution of each individual data item is masked. Indeed, Barak et al. propose methods to provide privacy, accuracy, and consistency when releasing contingency table data even when multiple tables are released. Dwork et al. also develop techniques for ensuring differential privacy under continual observation, which is necessary for any database that is not static, and will be updated over time. Techniques such as these will have to be adapted to the healthcare setting and deployed to ensure privacy.

Conclusion

In this paper we have identified an important class of inference problems that exist with the existing way healthcare data can be queried. We have shown a practical application of this attack technique against HCUPnet data, and have examined some potential solutions that can counter this. In the future, we plan to develop and test such sophisticated techniques to ensure privacy while still maintaining a high degree of utility to enable medical research.
References

1. Hellinger F.J. Hiv patients in the hcup database: a study of hospital utilization and costs. Inquiry, 41(1):95–105, 2004.

2. Jiang H.J., Ciccone K., Urlaub C.J., Boyd D., Meeks G. and Horton L. Adapting the hcup qis for hospital use: the experience in new york state. Joint Commission Journal on Quality and Patient Safety, 27(4):200–215, 2001.

3. Beale E.M.L. Branch and bound methods for mathematical programming systems. In Discrete Optimization II Proceedings of the Advanced Research Institute on Discrete Optimization and Systems Applications of the Systems Science, Panel of NATO and of the Discrete Optimization Symposium, volume 5 of Annals of Discrete Mathematics, pages 201 – 219. Elsevier, 1979. doi: 10.1016/S0167-5060(08)70351-0. URL http://www.sciencedirect.com/science/article/pii/S0167506008703510.

4. Sweeney Latanya. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, (5):557–570, 2002.

5. Adam Nabil R. and Wortmann John C. Security-control methods for statistical databases: A comparative study. ACM Computing Surveys, 21(4):515–556, December 1989. URL http://doi.acm.org/10.1145/76894.76895.

6. Malvestuto F. M., Mezzini M. and Moscarini M. Auditing sum-queries to make a statistical database secure. In ACM Transactions on Information and System Security (TISSEC), volume 9, pages 31–60, 2006.

7. Chowdhury S., Duncan G., Krishnan R., Roehrig S. and Mukherjee S. Disclosure detection in multivariate categorical databases: Auditing confidentiality protection through two new matrix operators. Management Sciences, 45:1710–1723, 1999.

8. Chin Francis Y. L. and Özsoyoglu Gultekin. Auditing and inference control in statistical databases. IEEE Transactions on Software Engineering, 8(6):574–582, 1982.

9. Kenthapadi Krishnaram, Mishra Nina and Nissim Kobbi. Simulatable auditing. In ACM Symposium on Principles of Database Systems, pages 118–127, 2005.

10. Lu Haibing, Li Yingjiu, Atluri Vijay and Vaidya Jaideep. An efficient online auditing approach to limit private data disclosure. In International Conference on Extending Database Technology, pages 636–647. ACM, 2009.

11. Dwork Cynthia. Differential privacy. In 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006), pages 1–12, Venice, Italy, July 9-16 2006. doi: 10.1007/11787006.

12. Dwork Cynthia. Differential privacy in new settings. In Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA ’10, pages 174–183, Philadelphia, PA, USA, 2010. Society for Industrial and Applied Mathematics. ISBN 978-0-8987198-6. URL http://dl.acm.org/citation.cfm?id=1873601.1873617.

13. Barak Boaz, Chaudhuri Kamalika, Dwork Cynthia, Kale Satyen, McSherry Frank and Talwar Kunal. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS ’07, pages 273–282, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-685-1. doi: http://doi.acm.org/10.1145/1265530.1265569. URL http://doi.acm.org/10.1145/1265530.1265569.

14. Dwork Cynthia, Naor Moni, Pitassi Toniann and Rothblum Guy N. Differential privacy under continual observation. In Proceedings of the 42nd ACM symposium on Theory of computing, STOC ’10, page 715. New York, New York, USA, 2010. ACM. ISBN 9781450300506. URL http://doi.acm.org/10.1145/1806689.1806787 http://portal.acm.org/citation.cfm?doid=1806689.1806787.