Published in final edited form as:

Acad Radiol. 2013 July ; 20(7): 889–896. doi:10.1016/j.acra.2013.04.011.

An Examination of Data Confidentiality and Disclosure Issues Related to Publication of Empirical ROC Curves

Gregory J. Matthews, PhD and Ofer Harel, PhD

Division of Biostatistics and Epidemiology, Department of Public Health, Amherst, MA (G.J.M.);
Department of Statistics, University of Connecticut, 215 Glenbrook Road U-4120, Storrs, CT
06269 (O.H.)

Abstract

Rationale and Objectives—Grant funding institutions often require organizations to share their collected data as widely as possible while safeguarding the privacy of individuals.

Summaries based on these data are often released. Here, the receiver operating characteristic (ROC) curve is explored for potential statistical disclosures in the presence of auxiliary data.

Materials and Methods—Formulas are introduced for calculating the missing data points from the full data set, given that a user has an empirical ROC curve and a subset of the data used to generate such a curve. Further, a discussion of the plausibility of this scenario is presented.

Results—Diagnostic test data were simulated and an ROC curve was produced. Using a subset of the true data and the points on the empirical ROC curve, an attempt was made to reproduce the missing parts of the data. Disease statuses were able to be determined exactly, whereas test scores were solved for up to their rank.

Conclusions—If an individual or organization possessed the points of an empirical ROC curve and a subset of the true data, the true data underlying the ROC curve can be reproduced relatively accurately. As a result, the release of summaries of data, including the ROC curve, must be given careful thought before their release from a statistical disclosure perspective.

Keywords

ROC analysis; statistical disclosure control; privacy; sensitivity; specificity; disclosure; missing data

Many agencies that fund medical and public health research require that data collectors take precautions to protect the privacy of the individuals whose data are being collected (1,2). However, many of these same agencies also require data collectors to provide a plan to disseminate these collected data while still maintaining privacy (3). The first step in maintaining privacy of individual level data—referred to as microdata—that will be released for research is to remove obvious identifiers (4) such as 18 identifiers outlined in the Health Insurance Portability and Accountability Act (5–7). These include information that could be easily used to identify an individual such as name, birth date, and social security number. However, simply removing these types of obvious identifiers is not enough to ensure individuals' privacy. An example of this can be found in previous work (8), where the author was able to take deidentified public health data that was released to the public and combine these data with publicly available voting records in order to identify individuals in

the released data. Therefore, although removing obvious identifiers is a necessary first step, it is certainly not sufficient to maintain the privacy of individuals.

RATIONALE AND OBJECTIVES

In general, there are a wide array of proposed methods for controlling statistical disclosure in microdata, for example, matrix masking (9) and synthetic data (10–13). Although these methods, to some degree, add a layer of privacy to the data that will potentially be released, quantifying just how much protection these methods provide is another challenge. If a measure of privacy was established, data-releasing institutions could simply meet this privacy threshold before releasing data. However, there are many possible ways that disclosures can take place, and therefore many different proposals for how to quantify privacy. Linkage-based measures of privacy in which a malicious data user is trying to identify a record in the data are presented elsewhere (14–18). Further proposals for assessing privacy can be found in the computer science literature (19–21). Measures of privacy based on inferential privacy include work on differential privacy (22) and its variants (23–27) as well as measures of privacy incorporating area under the receiver operating characteristic (ROC) curve (28,29).

Although there are some clear statistical disclosure issues with releasing microdata to the public, there are less obvious disclosure issues when other types of data are released, for instance, tabular data or summary statistics. Tabular data, often consisting of count data, pose many different privacy issues in terms of statistical disclosure. One common problem with tabular data occurs when small cell counts occur in a table. For instance, if a table cell contains a 1, the combination of attributes occurring in this cell is unique at least in the data, which could lead to an identification disclosure. As a result of this potential disclosure, small cell counts are often suppressed in released data tables. Summary statistics, such as the sample mean or estimated regression coefficients, may also pose the potential for statistical disclosure. Other possible data summaries may also be vulnerable to statistical disclosure in the presence of auxiliary information.

In this article, we focus on the ROC curve (30–32) and explore some potential statistical disclosure issues involved when a malicious data user has some subset of the true microdata. This is accomplished here through an attempt to learn private information about individuals' diagnostic test scores and disease status based on a simulated example. Under the assumption that a malicious data user has access to the true values of an empirical ROC curve and a subset of the data, this article examines what information can be learned about the subset of the data to which a malicious data user does not have access.

One of the main problems with controlling statistical disclosures and maintaining privacy in general is the possibility that a malicious data user may possess auxiliary data that he or she can use to learn a private attribute of an individual from released data that is meant to remain private. Specifically, for summary data, if an individual collects a large subset of the true data, even possibly all the observations except for one, that individual can potentially use that released summary statistic in conjunction with the auxiliary data to learn the value of the single datum that is missing. Although this may seem like an unrealistic example to some, this may be possible on a small scale when individuals disclose their data to another party. This could particularly be an issue with results being reported directly to patients (33,34). This exact scenario occurs, for example, in public health data exchanges in which data are aggregated from many sources. Public health officials may pool data from many different individual hospitals and perform analysis on the aggregated data and potentially publish results. This is increasingly easy to accomplish as an ever-increasing number of health care providers move toward electronic health records (EHR). Although a researcher

should not share a hospital's data with an unauthorized hospital, each hospital will have access to the raw data that they contributed to the research. If any of the hospitals are particularly large and contributed a substantial percentage of the data, they may be able to learn some information about patients at the other hospitals involved in the study. Even worse, if hospitals were to collude and combine their data, they may be able to potentially learn even more about the patients whose raw data they do not possess. Therefore, the need for greater awareness of statistical disclosure control is important, especially in a society increasingly reliant on data in a vast array of fields.

As a whole, statistical disclosure control is a broad topic and a full review is beyond the scope of this article. Several comprehensive reviews of the statistical disclosure control literature have been published (35,36).

MATERIALS AND METHODS

The false-positive rate (FPR) is defined as the probability that an individual not having disease is incorrectly classified as having the disease, and its empirical estimate is calculated as the number of false positives divided by the number of nondiseased individuals. Similarly, the true-positive rate (TPR) is defined as the probability that an individual having disease is correctly classified as having the disease, and its empirical estimate is calculated as the number of true positives divided by the number of diseased individuals. An individual is classified as having a disease if his or her test score is above some predefined cutoff, c . Otherwise, the individual is classified as not having the disease. The ROC curve considers all possible cutoffs for classification, and FPRs and TPRs are recorded for each cutoff. The ROC curve is created by plotting each pair of FPR and TPR calculated based on each of the cutoffs. By creating the ROC curve in this way, it will always begin at the origin at (0, 0) and extend to the point (1, 1). As the diagnostic accuracy of the test is increased, the curve will tend toward the upper left corner of the plot. Alternatively, those diagnostics tests that perform poorly will appear as an approximate 45° line from the origin to the point (1, 1).

In this article, we assume that a malicious data user is trying to learn the true disease statuses and test scores of the individuals in the study whose data were used to create the empirical ROC curve. Further, we are assuming that this user has the exact values of the empirical ROC curve (ie, based on the empirical true- and false-positive values) and a subset of the true data set. Given these two sets of information—the points on the ROC curve and a subset of the true data used to create the ROC curve—the question of interest here is how much can users learn about the raw data values in the full data set that they do not already have in their possession.

Plausibility

A common question in setting up this study was the question of whether this situation is at all plausible. How could a data user obtain a subset of the true data? We offer several realistic scenarios in which it is possible to obtain some or even a substantially large subset of the data.

- **Collusion:** Multiple individuals share the results of their radiology screening with each other to collect a subset of the true data. This is especially of concern when the number of observations in a study, n , is small.
- **Multiple data sources:** In a multisite study, data may be collected from many sources such as hospitals or other organizations. Each institution involved in a study would have access to their raw data, which composes a subset of the total pooled data for a study. This problem would be further compounded if institutions agreed to collude with one another. For example, a group of hospitals under the

same management could combine their radiology results for a larger study, and each individual hospital would have a subset of the full data.

- Updating data sources over time: If an individual or organization has a data set at some time point and then at a later time data points are added to the data set and the newer, larger data set is used to compute summary statistics, the individual or organization with the old data set has a subset of the full data.

These are just a few situations in which individuals could possess a subset of the full data set, which they can potentially use in conjunction with some released statistics based on the full data set to learn something about the remaining data points that are unknown to them. We should make it clear that the work presented in this article is entirely based on hypothetical scenarios, although we do believe that the potential exists for this type of disclosure to take place. In an era when medical centers are moving toward EHR, the dissemination of vast quantities of data can be moved and received very easily. This scenario, where a malicious data user gains access to a large subset of the total data through EHR, is of much more concern because a potentially very large subset of the total data can be obtained with ease. This is likely of much greater concern than a situation such as collusion between individuals because the collection of a large subset of a large set of radiology results through means other than electronic seems onerous.

Notation

Consider a data set D with n observations containing two variables for each individual: a diagnostic test score, t , and a true disease status, d . D^* is a subset of D with m observations removed so that there are $n - m$ observations in D^* . Let the vector of observed test scores in D^* be $t = (t_1 \dots t_{n-m})$ and the vector of observed disease statuses be $d = (d_1 \dots d_{n-m})$ where d_i is 1 when the i -th individual has the specified disease and is 0 otherwise. Let the unobserved test scores be $t^X = (t_1^X \dots t_m^X)$ and the unobserved disease statuses be $d^X = (d_1^X \dots d_m^X)$.

Using this notation, let $P = \sum_{i=1}^{n-m} d_i + \sum_{j=1}^m d_j^X$ and $N = n - \sum_{i=1}^{n-m} d_i - \sum_{j=1}^m d_j^X$ respectively, be the total number of diseased and nondiseased subjects in the full data set D . Similarly, define $P^* = \sum_{i=1}^{n-m} d_i$ and $N^* = n - m - \sum_{i=1}^{n-m} d_i$ to be the number of diseased and nondiseased subjects in the observed data, D^* , respectively.

This allows one to define FPR and TPR for D in terms of the number of false positives (FP), the number of true positives (TP), and a diagnostic cutoff (c):

$$FPR(c) = \frac{FP(c)}{N^* + \left(m - \sum_{j=1}^m d_j^X\right)} \text{ and } TPR(c) = \frac{TP(c)}{P^* + \sum_{j=1}^m d_j^X}$$

where

$$FP(c) = \sum_{i=1}^{n-m} (1 - d_i) I(t_i \geq c) + \sum_{j=1}^m (1 - d_j^X) I(t_j^X \geq c) \quad TP(c) = \sum_{i=1}^{n-m} d_i I(t_i \geq c) + \sum_{j=1}^m d_j^X I(t_j^X \geq c) \quad I(t_i > c) = \begin{cases} 1 & : t_i \geq c \\ 0 & : t_i < c \end{cases}$$

Using the true values of the empirical ROC curve, the true values for FPR and TPR are known for values of the cutoff equal to each of the true test scores from the full data. If a malicious data user has a subset of the true data, D^* , the unknown values are the unobserved test scores t^X and the unobserved disease statuses d^X . Combining the true test scores and

disease statuses from D^* and the true TPR and FPR values from the empirical ROC curve, the unknown values of t_j^X and d_j^X , $j = 1 \dots m$ could be found by setting up a system of equations with one equation for each true value of TPR and one equation for each true value of FPR based on the true points from the empirical ROC curve.

RESULTS

Example 1: One Missing Data Point

Suppose there is a small data set, D , with 10 observations in it as displayed in the two left columns of Table 1. Each observation in the full data consists of the true disease status and true test score for an individual. Further, a malicious data user may have access to some subset of the data, D^* , as in the two right columns of Table 1. The user would also likely be aware of the true number of subjects in the full data set because the true value of n is often reported in published studies. Here, a malicious data user who possesses the data in the two right columns of Table 1 has the goal of trying to learn the true value of t_1^X and d_1^X . This can be accomplished once this user gets access to the data in Table 2 that contain the points along the empirical ROC curve based on the full set of data, D . A plot of the empirical ROC curve for the data set can be seen in Figure 1.

As an example of how to set up the equations to solve for the values of d_1^X and t_1^X , consider a cutoff value of $c = 1.05$. This gives the following equation:

$TP(1.05) = \sum_{i=1}^9 d_i I(t_i \geq 1.05) + d_1^X I(t_1^X \geq 1.05)$ Computing the first summation involves counting the number of observations where $t_j \geq 1.05$ and d_j is 1. Based on the observed data from two right columns of Table 1, there are five observations in which t_j meets the requirement and of those three have disease status $d_j = 1$. This yields

$$TPR(1.05) = \frac{3 + d_1^X I(t_1^X \geq 1.05)}{4 + d_1^X}$$

because $P^* = 4$. This formula can be calculated for TPR and a similar formula calculated for FPR using each observed value of t_j as well as the unknown value of t_1^X as a cutoff.

Next, combining the data in Table 2 with the data in the two right columns of Table 1, a malicious data user can solve for the values of t_1^X and d_1^X . Here, this problem was solved using the R function “optimize” by performing onedimensional optimization and optimizing over the sum of the absolute differences between the true values of FPR and TPR and the values of TPR and FPR based on t_1^X and d_1^X . In this simple case, the optimize function can be run once assuming that $d_1^X = 0$ and once assuming that $d_1^X = 1$, each time solving for the single variable t_1^X . Whichever of the two runs of the optimize function achieves the smaller sum of absolute difference, that result will be kept yielding a plausible guess for t_1^X and d_1^X . In this case, the sum of the absolute deviations is minimized at 0 when t_1^X is near its true value and $d_1^X = 1$. Because the ROC curve depends only on ranks, many different values of t_1^X are solutions to minimizing the sum of these absolute deviations. The optimize function happens to return the value $t_1^X = 0.67$ in this case. This is close to the actual value of true missing test score, 0.61, but any value between 0.53 and 1.05 is also a possible value for the missing test score, t_1^X because the ROC is based only on ranks of the data. The range of possible solutions to the missing test score will depend on where it falls in the ordered list of test scores. For

instance, if the missing test score had been 2.71 in this data set, all one could say about the missing value was that it was larger than 1.32. As for the disease status, which is likely a much more sensitive piece of information, it can be found exactly in this scenario.

Of course, this example is simply illustrative and is likely unrealistic based both on the small number of observations in the full data set and the ability of an individual to obtain $n - 1$ observations of the full data set. Obtaining a data set D^* with $n - 1$ observations is something of a worst case scenario. The data snooper has almost all of the information in D , while there is still something private for them to learn. (Once they have the full data set, there is no longer anything left to keep private from them.)

Example 2: More Than One Missing Data Point ($m \geq 2$)

A more realistic scenario would include a larger data set and a larger proportion of observations missing from it. The following example starts with a data set with $n = 100$ observations, 50 diseased and 50 nondiseased. Test scores for the nondiseased group follow a standard normal distribution with mean 0 and variance 1. Test scores for the diseased subjects follow a normal distribution with mean 1 and variance 1. In this fashion, a data set was simulated with a resulting empirical ROC curve as shown in Figure 2.

Here, we choose $m = 10$, leaving 90 observations in the data set D^* . The question of interest is: What can be learned about the 10 missing observations by combining the points on the empirical ROC curve and the data set D^* ?

Rather than set up equations and solve for the unknown quantities as before, which may be cumbersome when $m \geq 2$, we chose to attempt to solve this problem with more than one missing observation using a Markov chain Monte Carlo (MCMC) technique (37) to solve for t^X and d^X . This procedure was implemented using R, and full details of the algorithm used are described in the Appendix.

Table 3 presents one solution to the missing data points found using the MCMC procedure. The data found in the two left-most columns, when added to the 90 known data points, will create an empirical ROC curve that is exactly the same as the empirical ROC curve created from the true data set. One can see that the guessed disease statuses are exactly the same as the true disease statuses, whereas the guessed values of the test scores are often very close, but certainly not exact. Some values, most notably in the bottom two rows of the left-most column, are substantially different from the true values of the test score. This is simply due to the random guessing of the search algorithm. The ROC curves only need to be matched based on the ordering of the diagnostic test scores. So, although these guesses are substantially different than the truth, it does reflect the information that the test score is on the lower end of test scores in the data set, and that is all that can be learned about that observation. Alternatively, test scores in the middle can be found with a relatively high degree of accuracy. This occurs when the missing value has to be between two very close values that are both known and have different disease statuses. This forces the guess of the test score to fall in a very narrow window, giving us more information about the true value of the test score. In the end, one can solve for disease statuses uniquely, whereas a solution to the test scores is not unique but correctly ordered.

After all of the missing disease statuses in D^* are determined, a malicious data user can then attempt to learn attributes of the individuals in the data set. This could happen, for instance, if all the missing disease statuses in D^* are found to be the same. A data user can then infer the exact value of disease status for all of the individuals in D that are not in D^* . However, even if not all of the missing disease statuses are the same, as in this example, the possibility of some type of disclosure still exists. For instance, say there is a particular disease that

occurs in 10% of the population. If the malicious data user learns that 90% of the individuals who are in D but not in D^* have that particular disease, some amount of disclosure has occurred. Simply knowing that every individual in a population has a 10% chance of having a disease is not a privacy concern; however, in this example, a data user has learned that the individuals in this data have a 90% chance of having a disease. Our knowledge of the probability that an individual has a 90% chance as opposed to a 10% chance of having a disease is not based on the fact that the individual belongs to the general population of interest nor is it based on knowing some demographic characteristic of the individual that affects the likelihood of the disease. The updated information about the probability of having a disease is based on simply knowing that the person belongs to the dataset.

CONCLUSIONS

With a subset of a true data set and the true values of an empirical ROC curve, the remaining missing disease statuses can be recreated and the missing test values can be found up to their ordered location within the test scores. This allows a malicious data user to learn data points that contain true disease statuses and the ordered location of the test score for individuals who are in the larger data set, but outside of the subset of data that he or she possesses. This allows a user to closely reproduce the microdata associated with the creation of the ROC curve that would most likely not be released in its raw form because of ethical or legal requirements.

Further, although it would certainly be considered a disclosure if an individual could be linked to an observation in the data set resulting in the learning of the true disease status, it seems like a difficult undertaking in reality—although certainly possible depending on the auxiliary data that a user possesses. More plausible as a potential disclosure would be a scenario in which a malicious user had a subset of the data that occurred in such a way that all or nearly all of the observations that were missing had the same disease status. Once the data user combined D^* with the ROC curve and learned that all of the observations that were missing were the same disease status, the user can infer the disease status of any subject that is in the full data set, but who is not in D^* . No linking or identification needs to take place in order for a private attribute to be learned. All that needs to be known about a subject is that he or she is in the data set D .

Some individuals may not particularly care if information about themselves is not kept private, but this does present the potential for a situation in which their decision to reject their right to privacy has the potential to affect the privacy of others. In this sense, privacy is no longer an individual choice, but must be considered collectively. Each time participants in a database reveal some detail about themselves, the privacy of every entry in the entire database could potentially suffer. In this sense, privacy is not necessarily something that is just a personal decision, but rather one that individuals must (or should) consider in terms of how their decision to disseminate their personal information publicly affects the privacy concerns of everyone else who has participated in the databases. Because ROC curves are often used to assess diagnostic accuracy in a radiology setting, we recommend that researchers working in this area be aware of some of the potential threats to privacy when working with and releasing ROC curves in their research.

Because an individual releasing his or her own private data can lead to the disclosure of information about other members of a database, data collectors (eg, researchers, radiologists) may want to discuss this issue with individuals at the time of data collection. Along with all of the consent forms and the guarantees of confidentiality presented by the datacollecting agency, data collectors may want to mention how the privacy of other individuals can be affected by the release of others' private information, as it is quite likely

that many study participants are unaware of how disclosure of their information may lead to disclosure of information pertaining to other individuals. For the sake of the privacy of others, this may make participants more reluctant to release their information. Further, individuals or organization that are publishing and releasing summary statistics or plots such as the ROC curve need to at least be aware of the potential for statistical disclosures to take place. These individuals and organizations can limit the potential disclosures by safeguarding their own raw data and making sure that only authorized users have access to all, or even a subset, of the data. Also, organizations should be aware, when releasing plots such as the ROC curve, if a separate organization has access to a subset of the full data. Therefore, agencies should keep good records of exactly who has access to different parts of the full data, which will allow informed decisions to be made as to which plots and summary statistics can be released without undue levels of risk of disclosures occurring.

Future and ongoing work in this area includes attempting the same type of disclosures with a ROC curve and a data set D^* , but rather than using the points on the empirical ROC curve, the points on a smooth ROC curve would be used. A technique similar to the one described here could possibly be used to create a disclosure scenario similar to the examples presented here. One other issue to overcome regarding the empirical ROC curve is how to collect the true data points from this curve. Here, it is assumed that a malicious data user would have a data set D^* and also the exact values of the points on the empirical ROC curve. Although it is certainly possible to possess a data set such as D^* , it is less likely that a malicious data user would have the exact values of the empirical ROC curve. A more plausible scenario is that a data user would have D^* and an image of the empirical ROC curve. Although this certainly makes the problem more difficult to solve, the hurdle is likely a small one because there are many publicly available software programs for extracting data from an image. After a malicious data user transforms an image of the empirical ROC curve into data points, the procedure presented here can once again be used to try to find the true values of the data that the malicious data user does not have. Finally, although many variables will affect what the malicious data user can learn about the unobserved portion of the data, two of the most important ones include the sample size of the total data, n , and the size of the observed data, m . In the future, we believe an examination of how differing values of m and n are related to the amount of information that can be learned about the unobserved data points would be useful. Results of such a study would aid in creating guidelines for practical use.

Acknowledgments

This project was partially supported by National Institute of Mental Health Award Number K01MH087219. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

REFERENCES

1. National Institutes of Health. [Accessed May 20, 2013] NIH data sharing policy and implementation guidance. Available at: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
2. National Science Foundation. [Accessed May 20, 2013] Chapter II - proposal preparation instructions. Available at: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp
3. National Institutes of Health. [Accessed May 20, 2013] Final NIH statement on sharing research data. Available at: <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>
4. Chertoff J, Pisano E, Gert B. Core curriculum: research ethics for radiology residents. *Acad Radiol*. 2009; 16:108–116. [PubMed: 19064218]
5. Health Insurance Portability and Accountability Act of 1996. [Accessed May 20, 2013] Pub.L. 104–191, 110 Stat. 1936. Available at: <http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/html/PLAW-104publ191.htm>

6. Fetzer D, West O. The HIPAA privacy rule and protected health information: implications in research involving DICOM image databases. *Acad Radiol*. 2008; 15:390–395. [PubMed: 18280936]
7. Avrin D. HIPAA privacy and DICOM anonymization for research. *Acad Radiol*. 2008; 15:273. [PubMed: 18280924]
8. Sweeney L. k-anonymity: a model for protecting privacy. *Int J Uncertainty Fuzziness Knowledge Based Syst*. 2002; 10:557–570.
9. Cox LH. Matrix masking methods for disclosure limitation in microdata. *Survey Methodol*. 1994; 6:165–169.
10. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. *J Official Stat*. 2003; 19:1–16.
11. Reiter JP. New approaches to data dissemination: a glimpse into the future (?). *Chance*. 2004; 17:11–15.
12. Reiter JP. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *J Roy Stat Soc Series A*. 2005; 168:185–205.
13. Matthews GJ, Harel O, Aseltine RH. Examining the robustness of fully synthetic data techniques for data with binary variables. *J Stat Comput Simulation*. 2010; 80:609–624.
14. Paass G. Disclosure risk and disclosure avoidance for microdata. *J Business Econ Stat*. 1988; 6:487–500.
15. Duncan G, Lambert D. The risk of disclosure for microdata. *J Business Econ Stat*. 1989; 7:207–217.
16. Bethlehem JG, Keller W, Pannekoek J. Disclosure control of microdata. *J Am Stat Assoc*. 1990; 85:38–45.
17. Marsh C, Skinner C, Arber S, et al. The case for samples of anonymized records from the 1991 census. *J Royal Stat Soc*. 1991; 154:305–340.
18. Skinner C, Marsh C, Openshaw S, et al. Disclosure control for census microdata. *J Official Stat*. 1994; 10:31–51.
19. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncertainty, Fuzziness and Knowledge Based Syst*. 2002; 10:571–588.
20. Machanavajjhala A, Kifer D, Gehrke J, et al. L-diversity: privacy beyond k-anonymity. *ACM Trans Knowl Discov Data*. 2007; 1:3. <http://doi.acm.org/10.1145/1217299.1217302>.
21. Li, N.; Li, T.; Venkatasubramanian, S. t-closeness: privacy beyond k-anonymity and l-diversity. *Data Engineering; IEEE 23rd International Conference; ICDE*. 2007. p. 106–115. 2007
22. Dwork, C. Differential privacy. *ICALP: Springer*; 2006. p. 1–12.
23. Machanavajjhala, A.; Kifer, D.; Abowd, J., et al. *International Conference on Data Engineering*. Cornell: Cornell University Computer Science Department; 2008. Privacy: theory meets practice on the map; p. 10
24. Dwork C, Nissam K. Privacy-preserving datamining on vertically partitioned databases. *Adv Cryptol*. 2004:528–544.
25. Blum, A.; Dwork, C.; McSherry, F., et al. Practical privacy: the sulq framework. *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*; 2006. p. 128–138.
26. Dwork, C.; Mcsherry, F.; Nissim, K., et al. Calibrating noise to sensitivity in private data analysis; *Proceedings of the 3rd Theory of Cryptography Conference*. Springer; 2006. p. 265–284.
27. Smith, A. [Accessed May 20, 2013] Efficient, differentially private point estimators. Available at: <http://arxiv.org/abs/0809.4794>
28. Matthews GJ, Harel O, Aseltine RH. Assessing database privacy using the area under the receiver-operator characteristic curve. *Health Serv Outcomes Res Methodol*. 2010; 10:1–15.
29. Matthews G, Harel O. Assessing the privacy of randomized vector-valued queries to a database using the area under the receiver operating characteristic curve. *Health Serv Outcomes Res Methodol*. 2012; 12:141–155. <http://dx.doi.org/10.1007/s10742-012-0093-y>.
30. Zou KH, Liu A, Bandos AI, et al. Statistical evaluation of diagnostic performance: topics in ROC analysis. *Chapman and Hall/CRC Biostatistics Series*. 2011

31. Pepe, MS. The statistical evaluation of medical tests for classification and prediction. Oxford University Press; 2003.
32. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8:283–298. [PubMed: 112681]
33. Jha S. Communicating results directly to patients: don't ignore the price tag of this added value. Acad Radiol. 2012; 19:643–645. [PubMed: 22578222]
34. Kuhlman M, Meyer M, Krupinski E. Direct reporting of results to patients: the future of radiology? Acad Radiol. 2012; 19:646–650. [PubMed: 22578223]
35. Skinner C, Pfeiffermann D, Rao CR. Statistical disclosure control for survey data. Handbook of statistics, vol. 29A: sample surveys: design, methods and applications. 2009:381–396.
36. Matthews GJ, Harel O. Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy. Stat Surv. 2011; 5:1–29.
37. Gelman A, Carlin J, Stern H, et al. Bayesian data analysis. Chapman Hall/CRC. 2003

APPENDIX

Description of the MCMC Procedure

1. Initial values for the vectors t_0^X and d_0^X are generated by sampling from the observed values contained in D^* .
2. D^* , t_0^X and d_0^X are combined to form a candidate complete data set and are used to calculate TPR and FPR for the ROC curve.
3. The true values of TPR and FPR are compared to the values of TPR and FPR based on the candidate complete data set. The sum of the absolute deviations (SAD) between the true and candidate TPRs and FPRs is calculated.
4. $t_{t+1}^X = t_t^X + Z$, where Z is an m dimensional multivariate normal distribution with mean 0, variance 1, and covariance 0 and d_{t+1}^X is generated by randomly replacing one randomly chosen element of d_t^X with either a 0 or a 1.
5. SAD_{t+1} is calculated using D^* , t_{t+1}^X , and d_{t+1}^X at time $t + 1$. If $SAD_{t+1} < SAD_t$, $t_{best}^X = t_{t+1}^X$, and $d_{best}^X = d_{t+1}^X$; otherwise, $t_{best}^X = t_t^X$ and $d_{best}^X = d_t^X$.
6. Steps 4 and 5 are repeated until $SAD = 0$.

Receiver-Operating Characteristic Curve

Figure 1.
Empirical receiver operating characteristic curve for example 1.

Figure 2.
Empirical receiver operating characteristic curve for example 2.

TABLE 1Full Data D in the Two Left Columns, and a Subset of Data, D^* , in the Two Right Columns

Test Score	Disease Status	Test Score	Disease Status
2.98	1	2.98	1
2.71	1	2.71	1
1.50	1	1.50	1
1.32	0	1.32	0
1.05	0	1.05	0
$t_1^X=0.61$	$d_1^X=1$		
0.53	0	0.53	0
0.51	1	0.51	1
0.20	0	0.20	0
1.85	0	1.85	0

TABLE 2

ROC Data

FPR	TPR
0	0
0	0.2
0	0.4
0	0.6
0.2	0.6
0.4	0.6
0.4	0.8
0.6	0.8
0.6	1.0
0.8	1.0
1	1.0

TABLE 3

Results

Solution		Truth	
t^x	d^x	t^x	d^x
2.099	1	2.121	1
1.798	0	1.782	0
1.391	0	1.346	0
0.760	1	0.729	1
0.496	1	0.505	1
-0.096	1	-0.088	1
-0.610	1	-0.607	1
-0.813	1	-0.772	1
-6.686	0	-1.593	0
-8.978	0	-2.260	0