Stat
The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.100X/sta.0000

Examining statistical disclosure issues involving digital images of ROC curves

Gregory J. Matthews[a]*and Ofer Harel[b]

Received 31 July 2015;

It has been established that knowing the true values of empirical ROC curve (i.e. false positive and true positive rate pairs for all thresholds) along with a subset of the full data set consisting of $n-1$ observations can cause unwanted disclosures. Here, we explore a similar problem with two main extensions. First, rather than knowledge of the true values of the empirical ROC curve, we start only with an image of the empirical ROC curve. Second, rather than considering only subsets of $n-1$, we look at several differently sized subsets. Given this information (i.e. empirical ROC image and a subset of the full data set), we experimentally act as a data snooper and explore what can be learned about unobserved portions of the full data set.

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

1. Introduction

Statistical disclosure is defined as the disclosure of some private piece of information that was not previously known and was learned through the release of some data set or a statistic based on that data. In general, there are many different types of disclosures that can occur, and, for each type, there are a variety of ways in which they can occur. Two common types of statistical disclosures include identification and attribute disclosures. Identification disclosure is characterized by discovering the identity associated with a record in the released data set (e.g. learning that observation number 15 belongs to John Smith). While simply learning the identity of an individual or organization of a record in a data set does not by itself constitute a problem in terms of privacy, the concern is that this type of disclosure can lead to another type of disclosure that is often more problematic: attribute disclosure. This is characterized by learning of some private piece of information about an individual or organization in a data set (e.g. John Smith has cancer). It is not difficult to imagine how identity disclosures can easily lead to attribute disclosures.

[a] Department of Mathematics and Statistics, Loyola University Chicago, 1032 W. Sheridan Road, Chicago, IL 60660
[b]Department of Statistics, Room 323, Philip E. Austin Building, University of Connecticut, 215 Glenbrook Rd. U-4120, Storrs, CT 06269-4120
***Email: gmatthews1@luc.edu**

One famous example of an identification disclosure leading to an attribute disclosure is described in Sweeney (2002). Supposedly anonymized health insurance data was released to the public, however, certain quasi-identifiers such as age, birthdate, and gender were not removed from the data set. This allowed a so called linkage attack where auxiliary data in the form of publicly available voting records was used to link a medical record to an individual causing an identification disclosure. From there, it was only a small step toward learning private pieces of medical information about the identified individuals creating an attribute disclosure.

In order to prevent the type of disclosures described in Sweeney (2002), many method of statistical disclosure control (SDC) have been proposed for microdata (i.e. data on the individual level). Examples of SDC techniques for micro data include matrix masking (Cox, 1994) and synthetic data (Raghunathan et al., 2003; Reiter, 2004, 2005; Matthews et al., 2010), to name two of the multitude of possibilities. For a complete review of the topic of statistical disclosure control see manuscripts such as Skinner et al. (1994), Matthews & Harel (2011), and O'Keefe & Rubin (2015).

While there are many issues involving disclosure when releasing sensitive microdata, it may be less obvious that there is a potential for statistical disclosure to occur via the release of summary statistics or tabular data. A common problem with the release of tabular data is cells with small counts. In the extreme case, when a cell in a table has a count of 1, the combination of attributes associated with that given cell are unique in the data set. This creates a vulnerability for a malicious data user to attempt to create an identity disclosure and which could possibly lead to an attribute disclosure. The release of summary statistics when used in conjunction with auxiliary data can also create the potential for unwanted statistical disclosures to take place. For example, consider the sample mean of a data set based on n data points. In an extreme case, if some data user possessed $n-1$ of the data points and then learned the sample mean based on all of the n data points, that data user can learn the exact value of the remaining unknown data point that they do not possess. While knowing $n-1$ out of n total data points is, in many cases unlikely, it is realistically conceivable that an individual or an organization would have some subset of the full data set, which could allow someone in possession of such auxiliary data to learn attributes about the set of people whose data they do not possess.

While there are many scenarios where a data user could come to possess some subset of the a full data set, some of the more realistic situations include collusion, horizontal partitioning (Karr et al., 2004), and updating data over time. Collusion would occur if some subset of the observational units in a data set shared their information with each other allowing all of them to collectively possess a subset of the data. Horizontal partitioning could occur if a full data set consists of data from multiple data sources (e.g. hospitals, municipalities, states, etc.). When a data set is constructed in this manner, each individual entity necessarily possesses a subset of the true data. Finally, an organization may have a data set at some specific point in time, but they may not be allowed access to an updated version of that same data set at a later date with possibly more observations. In this scenario, this organization could try to learn about the remaining observations in the larger, more recent data set by combining the subset of the true data that they possess and analyses summaries based on the full data set.

The sample mean is a simple example of a statistic that can potentially cause statistical disclosures, other statistics have the potential to cause disclosures through their release. In this manuscript, we focus our attention on the receiver-operating characteristic (ROC) curve (Pepe, 2003). The ROC curve is a graphical evaluation of the performance of a binary classifier or a diagnostic test and consists of a plot of the collection of false positive rates on the x-axis versus the true positive rates on the y-axis for all possible test cut-offs.

Matthews & Harel (2013) presented an example of the aforementioned type of disclosure where they demonstrated that if a data user possessed the true points on the empirical ROC curve and a subset consisting of the complete data set missing only one data point (i.e. data subset of size n-1), by combining these two sources, the complete data can be approximately reproduced. Specifically, the true disease statuses can be reproduced exactly and the test scores can be recovered up to their rank. This practically means that if the subset of the data possessed by a malicious data user is of size *n*-1, then knowing the rank of a test score may allow someone to obtain a very small window of possible true test scores for the single observation that is not observed. While Matthews & Harel (2013) presented simple scenarios with a small amount of missing data (i.e. 1 missing observation), the intention of this current manuscript is to extend that work in two ways: (1) Examining how the size of the possessed subset of data affects the users ability to accurately reproduce the full data set, and (2) exploring what happens when the actual values of the empirical ROC curve are replaced with only an *image* of the ROC curve as it is usually published in scientific journals. These are important issues to study, so that researchers can make informed decisions in terms of the potential for statistical disclosures to take place when images of empirical ROC curves are published. (It should be noted that we are certainly not advocating for empirical ROC curve to never be published; merely exploring what types of potential statistical disclosure concerns one should have prior to publication.)

In Matthews & Harel (2013), a simulated data user was trying to recreate the full raw data using a large subset of the full data (i.e. all observations except for one) and a summary based on the complete data (i.e. numeric values of the empirical ROC curve). In this manuscript, a simulated data user is again trying to recreate the complete data set based on a subset of the full data (e.g half of the full data) and a summary of the data based on the full data. However, in this case the summary based on the complete data is not a numeric summary, but rather an electronic image of an empirical ROC curve. Using information collected by converting an image to numeric data and a subset of the complete data, a malicious data user could attempt to learn about the data points to which they do not have access. This process of converting from graph to data may have been a difficult, if not impossible task even in the recent past, but with a simple Google search anyone with an internet connection can download one of the many available programs to convert an image to data.

The remainder of this manuscript continues with a details of the experiment in Section 2 and a description of the steps a malicious data user would take to create a disclosure in Section 3. Section 4 presents the results and the manuscript concludes in Section 5 with a discussion.

2. Description of the experiment

Researchers in many fields use ROC analysis as an evaluative tool to assess the performance of binary classifiers. Often authors will publish the empirical ROC curves that result from their research such as Pepe et al. (2001); Shaw et al. (2003); Sven N. Reske & Perner (2006); Geng et al. (2014) to name just a few. As many of these ROC curves often involve sensitive data (e.g. cancer status), it is possible that there may be some potential statistical disclosure issues with the release of these images. In this manuscript, this question is explored under a certain set of assumptions about the auxiliary information that a data snooper may have.

Specifically, the main question we are interested in exploring here is: given an image of an ROC curve and a subset of the data used to generate the ROC curve, what can be learned about the remaining subset of the full data that a user does not possess. In order to create this situation, an image of an ROC curve from Faraggi et al. (2003) was used. In the aforementioned manuscript, they present the ROC curve shown in figure 1 (Figure 5(a) in the original paper). Via a request to the authors, we obtained the original data used to create this image, which consisted of disease statuses and test scores. The data contains 80 total data points where 40 of the observations were “healthy" and the 40 remaining observations were classified as “MI" (corresponding to myocardial infarction), and the test scores range in values from 0.1047 to 9.3710 with a mean of 3.0670. The mean test scores within each of the groups was 1.846 and 4.288 for those with status “healthy" and “MI", respectively. A box plot of the distribution of the test scores from the two groups is displayed in figure 4.

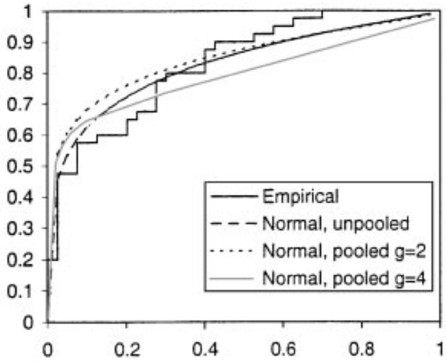

Figure 1. Figure 5(a) from Faraggi et. al. (2003)

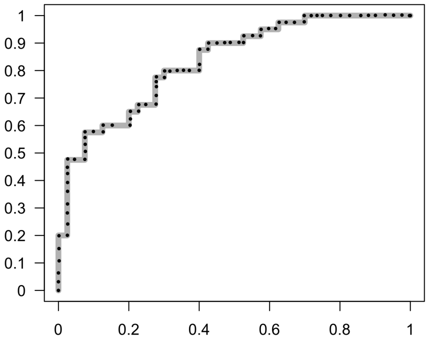

Figure 2. The gray line corresponds to the true empirical ROC curve reproduced from figure 1 while the black points represent the points extracted by DataThief

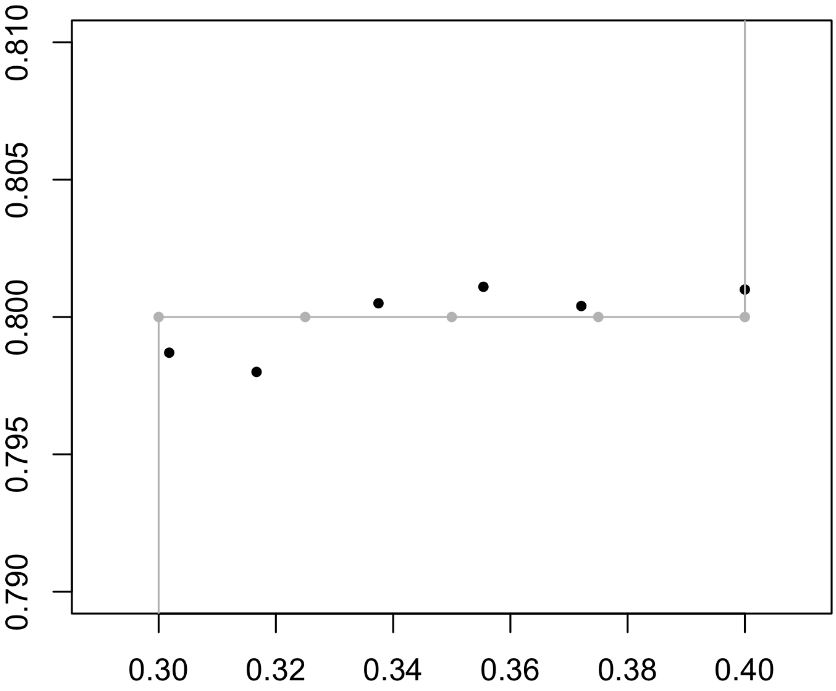

Figure 3. This figure is the same as figure 2 focused on a smaller window to highlight the differences between the true empirical ROC curve and the points extracted by DataThief. As before, the gray line corresponds to the true empirical ROC curve reproduced from figure 1 while the black points represent the points extracted by DataThief

A sample of size k was randomly selected from the n=80 total data points and this subset was then treated as data that a malicious data user possessed in addition to the image of the empirical ROC curve. Using only this information, along with the fact that n=80, we attempted to mimic the behavior of a malicious data user who was trying to

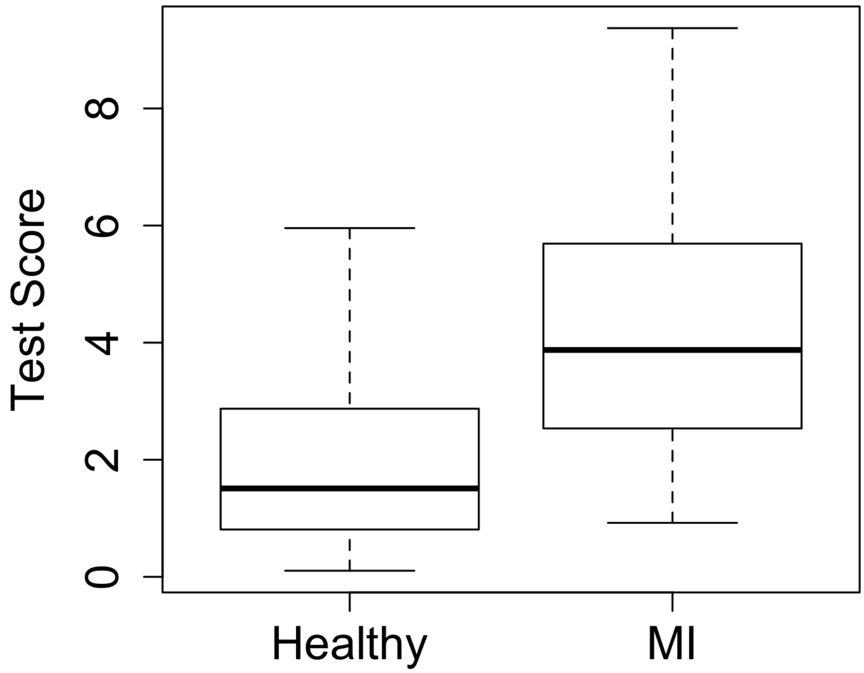

Figure 4. Distribution of test scores in the raw data

reproduce the full, original data set as closely as possible.

In summary, the experimental malicious data is attempting to reproduce the entire original data set with n=80 and is assumed to have the following information:

1. An electronic image of figure 5(a) from Faraggi et al. (2003) (Figure 1).
2. A subset of k observations of the true data.
3. The knowledge that the true data set contains n=80 data points from Faraggi et al. (2003).

3. Acting as a data snooper

We are assuming that the malicious data user has a subset of size k of the full data, knowledge of the size of the full data set (n=80) and the image of the true empirical ROC curve. Our main question, again, is using this information, and only this information, what can be learned about the remaining n-k observations that are in the true data that the malicious data user does not have access to? To test this, we acted as a malicious data user and attempted to recreate the portion of the data to which we did not have access to. This was achieved in two basic steps:

1. Extract data points of the empirical ROC curve from the electronic image
2. Use an MCMC procedure to solve for the n-k values from the full data set that we do not have access to

3.1. Extracting data points

The first step in acting as a data user seeking to reproduce the data is to extract the numeric values of the ROC curve from the image. Data points were extracted using a program called DataThief (Tummers, 2006). DataThief is freely available shareware written in Java that runs on many platforms including Unix, Windows, and Mac. DataThief works by loading the image a user wants to extract data from into the DataThief interface and then tracing points along the line or curve of interest. Once the user specifies the scale of the data, DataThief will output a data table containing (x,y) coordinates of the curve based on the points that were traced along it. Figure 2 shows the true empirical ROC from Faraggi et al. (2003) as a line and the numeric data points extracted by DataThief are shown as black points along the curve and Figure 3 is a more detailed version of figure 2 over a smaller window, which highlights the deviations of the true and extracted points. As mentioned, one will notice that these points do not necessarily fall exactly on the true ROC curve as they were traced by hand using a mouse in about 5 minutes. If one wanted to, they could spend more or less time tracing the curve to extract more or less data points from the curve, and thus more or less detail.

3.2. MCMC Procedure

Once the data points along the empirical ROC curve were extracted (i.e. false positive and true positive rates), an MCMC procedure (Gelman et al., 2003) similar to the procedure used in Matthews & Harel (2013) was used to recover plausible values of the missing raw data points (i.e. test scores and disease statuses). Details of the procedure can be found in the appendix, but the basic idea is to replace the values of test score and disease status of the missing data points with plausible values, construct an empirical ROC curve based on this data, and compare the difference between true ROC curve based on data from DataThief to the empirical ROC curve based on the true data combined with the plausible values of the missing data. The difference between these two ROC curves is recorded and a new set of plausible values for the missing data points based on a random draw from a normal distribution for test scores and a random sample from a Bernoulli distribution for disease statuses. If the new difference is smaller than the previous difference, the new set of plausible values is accepted, if not, then the old guess is kept. This procedure was repeated until the difference between the empirical ROC curve from DataThief and empirical ROC curve based on the known data combined with a set of plausible values was minimized. In Matthews & Harel (2013), the true values of the empirical ROC curve were known exactly, and the MCMC procedure was stopped only when the difference between the guessed empirical ROC curve and the true empirical ROC curve were 0. Here, however, the points on the empirical ROC curve are only approximate and some error is introduced when extracting the data using DataThief.

4. Results

The size of the subset of the data known to the data user, k, was taken to be 10, 40, and 70 in this experiment. For each value of k, one subset of data was drawn from the full data set and five MCMC simulations to reproduce the $n-k$ unknown data points were performed. In all of the simulations for all three values of k, the total number of diseased observations was able to be recovered exactly. Further, the disease status could be recovered accurately up

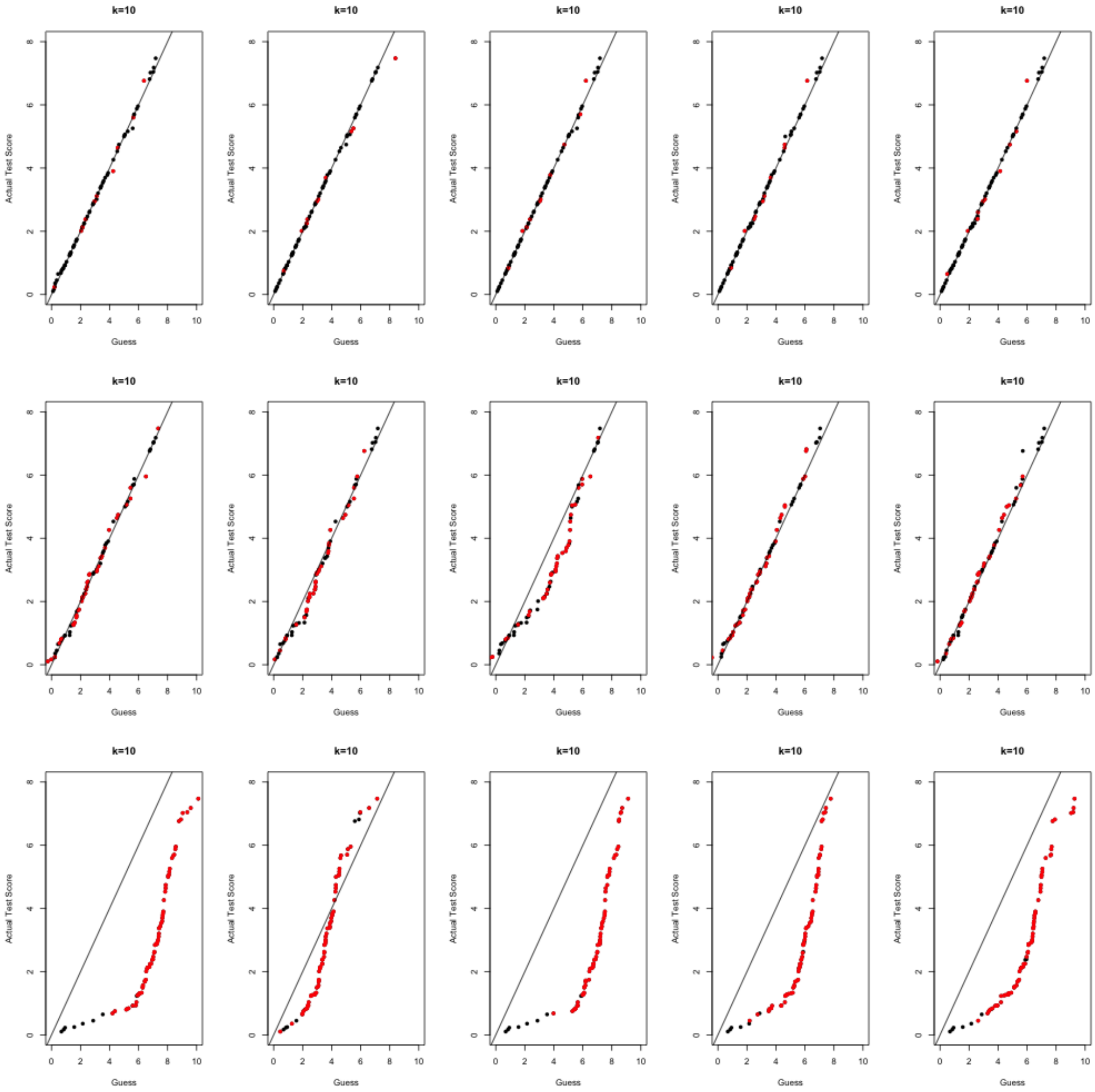

Figure 5. Predicted values of test scores versus actual values of test scores. Black points are known and red points are inferred.

to the ordering of the test scores, which would allow someone to make a statement about the disease status of the individual with the third largest test score, for example. The actual value of the test score can be recovered up to the rank exactly, but the actual values of the test can still be recovered with increasing accuracy as k increases. The result of each simulation presents one possible set of test scores that is possible given the ranking of the test scores and the restrictions imposed by the known values of some of the test scores. The five simulations that are presented here under each scenario are by no means exhaustive and are merely meant to illustrate a few of the many possible correct solutions. Results of these plausible sets of test scores are summarized by the mean squared error of the difference between the final guessed values of test scores compared to the known true values of test scores corresponding to missing values in the known data set are shown in table 1. When only ten true observations are know to the intruder,

k	1	2	3	4	5	average MSE
70	0.0284	0.0971	0.0364	0.0467	0.0743	0.0566
40	0.0369	0.3596	0.8656	0.1038	0.0357	0.2803
10	15.6538	0.9891	14.7052	7.6971	8.1942	9.4479

Table 1. Mean Squared Error of the test values

or 12.5% of the full data, the guessed values of the test scores have a fairly large MSE with an average of 9.4479. In one simulation, the MSE was less than 1, indicating a relatively accurate recovery of the true test scores compared to other simulations. As expected, for larger values of k, the MSE is smaller. For instance, when k is 40, or 50% of the full data, the average MSE of 5 simulations is 0.2803, and in three out of the five simulations the MSE is actually less than 0.11. With k=70, or 87.5% of the full data, and only 10 missing data points, the MSE of the true test scores versus the missing test scores ranges from 0.0284 to 0.0971 with an average across the five simulations of 0.0566. While it is impossible to recover the exact values of the test scores, as they can only be recovered up to their ranks, the precision with which test scores can be recovered becomes greater as the auxiliary data set increases in size imposing more restrictions on possible solutions.

To better visualize the accuracy of these simulations Figure 5 displays the results of each of the five simulations for each of the three different values of k=10, 40, and 70. Each plot compares the predicted test scores in the data set on the x-axis to the actual test scores on the y-axis. Both sets of predicted and actual test scores were sorted from smallest to largest, and the ordered points were plotted against each other. Black points represent the known test scores and the red points are the guessed values of the data that are not in the subset of the data. Each plot further displays the 45 degree line, which represents perfect prediction (i.e. The predicted value is exactly the same as the actual value). One can see that when k is small (i.e. k=10) prediction is much less accurate than the other values of k. Alternatively, when k is 70 the missing test scores can can be predicted with a high degree of accuracy indicated by the line of black and red points both falling very close the the 45 degree line. When k is 40, 50% of the data is missing, but missing values of test score can often be reproduced closely. While there is certainly less accuracy that when k is 70, it is still surprising how accurately test scores can be recovered when an individual knows only 50% of the true test scores and disease statuses. Finally, when k=70, the remaining $n-k$=10 observations can be recovered with even more accuracy and more consistency when k=40. Most of the simulations with k=40 show that the reproduced data is similar to the true data. However, in the third simulation when k=40, there are some problems with the reproduced data set. When k=70, not only are the unknown values reproduced more accurately, but there is also more consistency across simulations.

5. Discussion

While the release of summary statistics likely poses less of a risk of statistical disclosure than the release of microdata, this type of data cannot be released publicly without concern, especially when a data user has some subset of the full data. It is easy to imagine how the release of a sample mean based on n observations along with some subset of the true observations can lead to an individual learning about the actual values used to calculate the sample mean. It is less obvious but still a potential problem with other summary statistics such as regression coefficients. Matthews & Harel (2013) demonstrated that if a user knows the values on the empirical ROC curve and a subset of the true data, one can learn some information about the raw values of disease and test score.

In this manuscript, rather than starting with a summary statistic and a subset of the data, we replace the numeric summary statistics with an actual image of a statistical graph and attempt to learn about the raw values that generated such an image. Data was extracted from the image using a publicly available software application, and then test scores and disease statuses were attempted to be discovered using the same procedure explained in Matthews & Harel (2013). The difference here is that rather than using the exact values of the empirical ROC curve, only approximate values, which were extracted from an image of an empirical ROC, are used here. (Publication of an image of the empirical ROC curve is very common in the statistical literature.) In terms of learning the true disease statuses and test scores, there is very little difference between using the true values of the empirical ROC curve and approximate values of the ROC curve obtained by using DataThief. Therefore, we believe that the results presented here are similar to results that would be obtained if the true values of the empirical ROC curve were known.

It is not surprising that when a large portion of the data is known, that missing data points can be recovered with a relatively high degree of accuracy; Likewise, when only a small portion of the data is known, the recovery of the missing data is much more difficult. While some information can still be recovered, the accuracy of the predicted test scores is much lower. What many may find surprising is the degree of accuracy with which missing test scores can be recovered when a full 50% of the data is missing.

One of the practical implications of what we have demonstrated is that the privacy of individuals is not solely in the hands of the data releasing organizations. Each individual in the data set controls some small piece of the overall confidentiality of the data set and thus some small piece of each individuals privacy. Thus we argue it is the ethical responsibility of the individuals' in the data set to maintain the privacy of their individual information, even if they have no desire to keep their data private, for the sake of the other members of the data set who do, in fact, wish to have their data remain private. To summarize our point, disclosure control not only rests in the hands of data releasers, but also with the individuals in the data set.

Finally, one of the big assumptions of this work is that a data user possesses a subset of the true data. This is a big assumption that may be unrealistic in certain settings. However, there are also many scenarios in which the possession of a subset of the true data is plausible. Matthews & Harel (2013) mention three scenarios in which subsets of the true data can occur: collusion, multiple data sources, and updating data over time. Collusion occurs when multiple individuals share their data to create a subset of the true data, which is of particular concern in the size of the data, n, is small. Another potential situation where users possess a subset of the larger data set occurs when data is aggregated across organizations. In that case each organization possesses some subset of the full, larger data set. Finally, when data is updated over time, a data user may possess all of the data up to a specific time point t, but not the remaining newest data points, giving them a subset of the full data set. It is also possible that several of these examples could be combined leading to organizations combining their longitudinal data.

Appendix

Description of the MCMC procedure

1. Initial values for the vectors $\mathbf{t}_0^X$ and $\mathbf{d}_0^X$ are generated by sampling from the observed values contained in $D^\star$.
2. $D^\star$, $\mathbf{t}_0^X$, and $\mathbf{d}_0^X$ 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. $\mathbf{t}_{t+1}^X = \mathbf{t}_t^X + \mathbf{Z}$ where $\mathbf{Z}$ is an m dimensional multivariate normal distribution with mean 0, variance 1, and covariance 0 and $\mathbf{d}_{t+1}^X$ is generated by randomly replacing one randomly chosen element of $\mathbf{d}_t^X$ with either a

0 or a 1.

5. SAD_{t+1} is calculated using $D^{\star}$, $\mathbf{t}^{\mathrm{X}}_{\mathrm{t+1}}$, and $\mathbf{d}^{\mathrm{X}}_{\mathrm{t+1}}$ at time $t+1$. If $SAD_{t+1} < SAD_t$, $\mathbf{t}^{\mathrm{X}}_{\mathrm{best}} = \mathbf{t}^{\mathrm{X}}_{\mathrm{t+1}}$ and $\mathbf{d}^{\mathrm{X}}_{\mathrm{best}} = \mathbf{d}^{\mathrm{X}}_{\mathrm{t+1}}$, otherwise, $\mathbf{t}^{\mathrm{X}}_{\mathrm{best}} = \mathbf{t}^{\mathrm{X}}_{\mathrm{t}}$ and $\mathbf{d}^{\mathrm{X}}_{\mathrm{best}} = \mathbf{d}^{\mathrm{X}}_{\mathrm{t}}$.
6. Steps 4 and 5 are repeated until SAD is minimized. (This can determined visually by comparing the the the best guess points overlaid on the DataThief points.)

Acknowledgements

This project was partially supported by Award Number K01MH087219 from the National Institute of Mental Health. 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

Cox, LH (1994), 'Matrix masking methods for disclosure limitation in microdata,' *Survey Methodology*, **6**, pp. 165–169.

Faraggi, D, Reiser, B & Schisterman, EF (2003), 'ROC curve analysis for biomarkers based on pooled assessments,' *Statistics in Medicine*, **22**, pp. 2515–2527.

Gelman, A, Carlin, J, Stern, H & Rubin, D (2003), *Bayesian Data Analysis*, Chapman & Hall/CRC.

Geng, Y, Lu, W & Zhang, HH (2014), 'A model-free machine learning method for risk classification and survival probability prediction,' *Stat*, **3**(1), pp. 337–350, doi:10.1002/sta4.67.

Karr, AF, Karr, AF, Lin, X, Lin, X, Sanil, AP, Sanil, AP, Reiter, JP & Reiter, JP (2004), 'Secure regression on distributed databases,' *J. Computational and Graphical Statist*, **14**, pp. 263–279.

Matthews, G & Harel, O (2013), 'An examination of data confidentiality and disclosure issues related to publication of empirical ROC curves,' *Academic radiology*, **7**(20), pp. 889–896.

Matthews, GJ & Harel, O (2011), 'Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy,' *Statistics Surveys*, (5), pp. 1–29.

Matthews, GJ, Harel, O & Aseltine, RH (2010), 'Examining the robustness of fully synthetic data techniques for data with binary variables.' *Journal of Statistical Computation and Simulation*, **80**(6), pp. 609–624.

O'Keefe, CM & Rubin, DB (2015), 'Individual privacy versus public good: Protecting confidentiality in health research,' *Statistics in Medicine*, pp. n/a–n/a, doi:10.1002/sim.6543.

Pepe, MS (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press.

Pepe, MS, Etzioni, R, Feng, Z, Potter, JD, Thompson, ML, Thornquist, M, Winget, M & Yasui, Y (2001), 'Phases of biomarker development for early detection of cancer,' *Journal of the National Cancer Institute*, **93**(14), pp. 1054–1061.

Raghunathan, TE, Reiter, JP & Rubin, DB (2003), 'Multiple imputation for statistical disclosure limitation,' *Journal of Official Statistics*, **19**(1), pp. 1–16.

Reiter, JP (2004), 'New approaches to data dissemination: A glimpse into the future (?),' *Chance*, **17**(3), pp. 11–15.

Reiter, JP (2005), 'Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study,' *Journal of the Royal Statistical Society, Series A: Statistics in Society*, **168**(1), pp. 185–205.

Shaw, LJ, Raggi, P, Schisterman, E, Berman, DS & Callister, TQ (2003), 'Prognostic value of cardiac risk factors and coronary artery calcium screening for all-cause mortality,' *Radiology*, **228**(3), pp. 826–833, doi: 10.1148/radiol.2283021006, pMID: 12869688.

Skinner, C, Marsh, C, Openshaw, S & Wymer, C (1994), 'Disclosure control for census microdata,' *Journal of Official Statistics*, **10**, pp. 31–51.

Sven N. Reske, BNHWGFFDKPMGG, Norbert M. Blumstein & Perner, S (2006), 'Imaging prostate cancer with ^{11}c-choline PET/CT,' *Journal of Nuclear Medicine*, **47**(8), pp. 1249–1254.

Sweeney, L (2002), 'k-anonymity: A model for protecting privacy,' *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, **10**(5), pp. 557–570.

Tummers, B (2006), 'DataThief III,' http://datathief.org/.