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

Gregory J. Matthews, PhD, Ofer Harel, PhD

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.

Key Words: 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

Acad Radiol 2013; 20:889–896

From the 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.). Received January 15, 2013; accepted April 3, 2013. **Address correspondence to:** O.H. e-mail: ofer.harel@uconn.edu

http://dx.doi.org/10.1016/j.acra.2013.04.011

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.