Usable Deidentification of Sensitive Patient Care Data

Michael McQuaid
School of Information
University of Michigan
Ann Arbor, Michigan
mcq@umich.edu

Kai Zheng
School of Public Health
University of Michigan
Ann Arbor, Michigan
kzheng@umich.edu

Nigel Melville
Ross School of Business
University of Michigan
Ann Arbor, Michigan
npmelv@umich.edu

Lee Green
School of Family Medicine
University of Michigan
Ann Arbor, Michigan
greenla@umich.edu

1. INTRODUCTION

We propose a usability study for a system, called *data versioning*, to assist data stewards in deidentifying sensitive patient data. Two vital features of deidentified data are somewhat negatively correlated: preservation of the computational capabilities of data and minimization of disclosure risk. Our design is meant to help data stewards considering both features at the same time. Our interface is designed to explore two main questions. *First*, we ask whether the use of an abstract information visualization, resembling a topographic map, can assist data stewards in thinking about choosing a combination of two variables, EPP (extent of privacy protection) and value. *Second*, we ask whether an automatically generated narrative description of the consequences, provided as the actions are specified, can assist data stewards in thinking about their choices. In both cases, the outcome we seek is a better match between the user's security-related goals and actions.

2. RELATED WORK

The healthcare community has increasingly recognized that secondary use of electronic health record (EHR) data provides great promise for enhancing quality assurance, research, and surveillance.[2] Clinicians' day-to-day interactions with EHRs generate vast quantities of clinical and administrative data revealing rich details of patient health conditions, treatment effectiveness, and influence of social, behavioral, and policy factors. Secondary analysis of EHR data can thus help create a "rapid-learning" healthcare system to accelerate the advance of the U.S. evidence base by filling major knowledge gaps about healthcare costs, the benefits and risks of drugs and procedures, geographic variations, environmental health influences, the health of special populations, and personalized medicine.[1] The first step toward enabling secondary use of EHR data is to establish a data federation and sharing mechanism that makes multidimensional patient records collected at multiple institutions accessible to the clinical, policy, and public health research community. Such data federation and sharing efforts, however, could be associated with escalated risks to patient privacy and confidentiality if protection measures are not adequate; or could on the other hand compromise the value of data for secondary use because of unnecessary overprotection.[3]

3. METHOD

We developed an interface exhibiting what we hypothesize to be the important features, shown in Figure 1. The interface is divided into two parts. The *upper panels* (4 total) allow the data steward to input choices before deidentifying data. The main feature to test in our study is the visualization in the upper left corner, assisting the steward in defining a combination of two variables that characterize any decision about sharing sensitive data while preserving its value: the extent of privacy protection (EPP) and the extent to which relationships in data are preserved. We can quantify these to some extent as the size of bins into which we can partition the data, given what we know. This is a very rough quantification and changes as data analysis capabilities improve, prompting the design of the visualization described in Section 4.

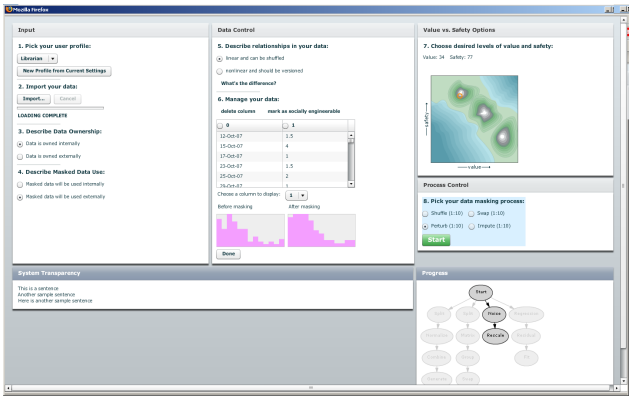

Figure 1: The prototype interface for stewards includes the following parts: upper right—conceptual map of risk and value; lower left—narrative explanation of the selections made in upper panels; lower right—graphical depiction of the selections made in upper panels, after the user presses the green start button.

The *lower panels* (2 total) provide narrative and graphical feedback about the choices made by the data steward. The narrative feedback consists of sentences whose appearance is triggered by choices made by the data steward. An

example sentence might be: *Values of* `age` *will be swapped so that all original values are preserved but swapped among subjects.* The steward can change the number and composition of these sentences by making different choices on the top panel. A separate interface (not shown here) allows a subject matter expert to create the mappings between actions and sentences without programming the interface.

4. STUDY DESIGN

Our study of data stewards involves two main features meant to better align a data steward's goals with her choices. *First*, we provide an information visualization, shown in the upper right corner of Figure 1. *Second*, we provide an interactive narrative description below the main control panel in Figure 1. We discuss these two features in turn.

The visualization shows different combinations of EPP (extent of privacy protection) and value. Our underlying algorithm can transform data according to many different combinations of these two variables, but not all combinations are desirable or attainable. Figure 2 shows two maps in which the undesirable region, low EPP and low value, is represented as the water in the lower left corner and the unattainable region, high EPP and high value, is represented as the water in the upper right corner.

The visualization is a metaphor and we hypothesize that people are so used to looking at actual topographic maps that we can transfer our understanding of them to a bivariate choice situation. What we would really like to do is to figure out the combination of the two variables that are most desirable to the data steward. The three mountain peaks represent identifiably different combinations. In a sense, we would like stewards to define themselves as fitting one of three categories but with some flexibility in case three turns out to not be the appropriate number of categories. These visualizations are proxies for five regions: undesirable, raw, deidentified, masked (maximizing value subject to a threshold EPP level), and masked (maximizing EPP subject to a threshold value level). The purpose of the topographic metaphor is to focus the steward's attention on identifying an appropriate region and to respond to that identification.

The two concept maps in Figure 2 differ in that (a) has shading by contour lines and more distinction of land from sea, and (b) has shading by region and less distinction of land from sea. We expect a data steward faced with (a) to be less likely to select water, but to be more indifferent between locations on land. We expect a data steward faced with (b) to be more likely to select water because of the diminished distinction between land and sea and the segmentation by region, making the sea region closest to land a more acceptable choice. We also expect a data steward faced with (b) to be less indifferent between locations on land and to select locations at higher altitudes and nearer to the target peaks representing three distinctly different mixes of EPP and value. We plan to investigate the contribution of the topographic metaphor by comparing to a map without it, as well as to a scheme with no map at all.

The interactive narrative below the control panel changes with choices made by the data steward as described above. The data steward may use the changing narrative to explore different combinations of choices. We expect that the data steward will articulate her goals in more detail after using the system. We expect this change to contrast with data stewards using the current standard, static flow charts.

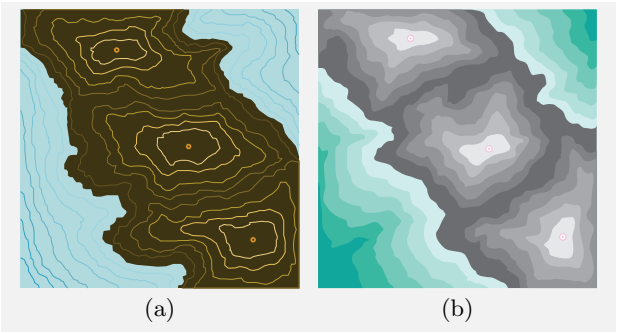

Figure 2: Alternative topo map designs, (a) with shading by contour lines and more distinction of land from sea, and (b) with shading by region and less distinction of land from sea.

In addition to the two main subjects of study, the interface in Figure 1 contains a third and fourth feature of interest. *Third*, the top center panel offers the data steward an easy interface for exploration of the data to be shared. In particular, histograms are offered, highlighting those columns suffering the disclosure risk inherent in small bins. We expect that data stewards using these histograms will make fewer mistakes in sharing columns with small bins than will data stewards without this facility. *Fourth*, the lower right panel contains a flow chart of the selected method. Because the system may require a long time to deidentify data while preserving statistical properties, this flow chart advises the data steward of progress and estimated time of completion.

5. FUTURE WORK

We are currently planning a user study to test the above expectations as hypotheses on a variety of data stewards already using advanced systems for data sharing. We expect this study to provide guidance about supporting flexibility for data stewards in the deidentification process. Our goal is to empower data stewards through a wider range of options enabled by visualization and narrative.

6. ACKNOWLEDGMENTS

We gratefully acknowledge Mark Goetz who programmed the user interface, and both Yelena Godina and Yesook Im who designed the conceptual maps used in the interface.

7. REFERENCES

[1] L. M. Etheredge. A rapid-learning health system. *Health affairs (Project Hope)*, 26(2):w107–18, Jan 2007.

[2] W. R. Hersh. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *The American journal of managed care*, 13(6 Part 1):277–8, Jun 2007.

[3] C. Safran, M. Bloomrosen, W. E. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang, D. E. Detmer, and Expert Panel. Toward a national framework for the secondary use of health data: an american medical informatics association white paper. *Journal of the American Medical Informatics Association : JAMIA*, 14(1):1–9, Jan 2007.