Published in final edited form as:

J Public Health Manag Pract. 2017 ; 23(4): e1–e4. doi:10.1097/PHH.0000000000000473.

A review of statistical disclosure control techniques employed by web-based data query systems

Gregory J. Matthews, Ph. D. [Assistant Professor],

Department of Mathematics and Statistics, Loyola University Chicago

Ofer Harel, Ph. D. [Professor], and

Department of Statistics, University of Connecticut

Robert H. Aseltine Jr., Ph.D. [Professor]

Oral Health and Diagnostic Sciences, University of Connecticut Health Center

Abstract

We systematically reviewed the statistical disclosure control techniques employed for releasing aggregate data in web-based data query systems listed in the National Association for Public Health Statistics and Information Systems (NAPHSIS). Each web-based data query system was examined to see whether 1) it employed any type of cell suppression; 2) it used secondary cell suppression; and 3) suppressed cell counts could be calculated. No more than 30 minutes was spent on each system. Of the 35 systems reviewed, no suppression was observed in over half ($n=18$); observed counts below the threshold were observed in 2 sites; and suppressed values were recoverable in 9 sites. 6 sites effectively suppressed small counts. This inquiry has revealed substantial weaknesses in the protective measures used in data query systems containing sensitive public health data. Many systems utilized no disclosure control whatsoever, and the vast majority of those that did deployed it inconsistently or inadequately.

Keywords

Privacy; Confidentiality; Vital Statistics; Public Health

Introduction

Interactive web-based data query systems are commonly used by state public health authorities to provide vital statistics and health surveillance data for use by researchers and policymakers. According to a recent survey of state coordinators for the Center for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS), a majority of US states currently host public health-relevant data query systems.¹ Such systems typically provide aggregate information to users in response to predefined, and in some cases customized, queries. Although query results are de-identified and presented only in aggregate form, care must be taken to prevent individuals from being re-identified. Re-identification may occur when individuals have rare conditions, attributes, or configurations of attributes or when ancillary information, such as publicly available voter lists or death registries, can be linked to files containing sensitive information. Best practices for statistical disclosure control are well established and widely referenced.²⁻⁵ Typically, privacy

protection is accomplished using primary and secondary, or complementary, suppression strategies, where data that do not meet certain quantitative thresholds -- e.g, cell sizes < 5, or marginal totals from which small cell counts can be derived -- are suppressed.⁶ (Primary suppression is the direct suppression of cells with small counts, whereas secondary suppression suppresses additional cells that do not have small counts themselves, but which need to be suppressed to protect the values in the primarily suppressed cells.)

Despite the well-established guidelines for privacy protection and statistical disclosure control related to WDQS,² there has not been a systematic review of what public health data systems are currently doing to protect the privacy of individuals whose information they maintain, nor has the adequacy of the privacy protections adopted in these systems been examined. In this brief, we present an overview of the privacy protections used by public health web-based data query systems. We focus on three related questions: 1) What approaches to privacy protection have states taken in presenting health data through web-based data query systems? 2) How well do these systems adhere to the strategies they have selected? 3) Are states' approaches to privacy protection for web-based data query systems adequate?

Methods

We reviewed 35 state-affiliated, public health-relevant web-based data query systems included in the inventory presented by the National Association of Public Health Information Systems (NAPHSIS).⁷ The statistics disclosure control techniques employed by all systems that allowed flexible generation of tabular results were documented. Following determination of the types of statistical disclosure methods employed by each site, a maximum of 30 minutes of examination per site was allowed to search for flaws in the implementation of disclosure control practices. Examples of flaws in the implementation disclosure control would include: observing cell counts below the stated suppression threshold, the ability to recover suppressed values using observed marginal totals, and disclosure by differencing using multiple queries⁸ (e.g., by obtaining query results for deaths among individuals age 35 or younger and deaths among age 34 or younger, which when differenced provides the number of deaths for those who are exactly 35 years old).

For each query system, we attempted to answer the following questions and documented results with screen shots of the query results when a problem was identified:

- Does this site employ cell suppression of any kind, and what is the threshold for suppression?
- If a site is employing primary cell suppression, are they also employing complementary cell suppression?
- If a site is employing primary cell suppression, can disclosure by differencing be employed to calculate the true value of a suppressed cell?

Results

Our summary of the statistical disclosure control practices employed by web-based data query systems is presented in Table 1. Whether and how each system addressed disclosure concerns varied considerably. Of the 35 systems evaluated, 18 made no effort to incorporate disclosure control techniques. Among those that claimed to incorporate such strategies, some form of suppression was most commonly employed, although the standards used -- minimum cell size tolerated, whether both primary and secondary cell suppression strategies were attempted -- varied dramatically. Two systems used primary suppression thresholds that were less stringent than those advocated in prominent guidelines² (i.e., less than 5). In 9 systems, suppressed cells were recoverable, either by examining marginal totals for rows or columns with suppressed cells or by “differencing” results obtained from multiple queries. In only 6 of the 35 systems reviewed were no suppression errors or flaws observed.

Discussion

This inquiry has revealed substantial weaknesses in the privacy protection measures used in state authorized web-based data query systems containing sensitive health and public health data. Many systems utilize no disclosure control whatsoever, and the vast majority of those that do deploy it inconsistently or inadequately, failing to meet the standards promulgated in the Center for Disease Control and Prevention's Disclosure Manual.⁶ Consequently, information from individuals whose attributes make them “outliers” is freely available; our review found that this includes highly sensitive information such as causes of death due to HIV or suicide. In some cases this resulted in the ability to recreate person- or record-level data, which carries far greater disclosure risk than aggregate data and should not be publicly available without a data use agreement and formal human subjects review. None of the systems reviewed in this study required a human subjects review or completion of a data use agreement to access data.

Results from a recent survey of state BRFSS coordinators¹ suggest that the risks associated with the disclosure control practices used in web-based data query systems are not appreciated. BRFSS coordinators whose states had deployed a query system reported that privacy protection was one of the least challenging aspects to presenting health data over the internet, falling well below such concerns as the cost of hardware and software and a lack of internal information technology support. We hope that the results of our analysis will prompt states to re-examine their approaches to privacy protection and statistical disclosure control in their public health data systems.

Public Health Implications

Access to population level health data is essential to public health planning and surveillance efforts. Although the aggregate data presented in state health- and public health-related web-based data query systems support these critical public health activities, the inadequacy of safeguards to prevent re-identification of ostensibly de-identified data presents serious risks to personal privacy and jeopardizes public support for such data access systems.

Acknowledgements

This project was partially supported by the State of Connecticut and 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

1. Ahuja, M. Strengthening Public Health through Web-Based Data Query Systems.; Doctoral Dissertations. 2015. p. 993<http://digitalcommons.uconn.edu/dissertations/993>
2. Hundepool A, Domingo-Ferrer J, Franconi L, et al. A CENTre of EXcellence for Statistical Disclosure Control Handbook on Statistical Disclosure Control Version 1.2. 2010
3. O'Keefe CM, Rubin D. Individual privacy versus public good: protecting confidentiality in health research. *Statistics in Medicine*. 2015; 34(23):3081–3103. [PubMed: 26045214]
4. Matthews G, Harel O. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*. 2011; 1:1–29.
5. Skinner, C. Statistical disclosure control for survey data.. In: Pfeiffermann, D., Rao, CR., editors. *Handbook of Statistics Vol. 29A: Sample Surveys: Design, Methods and Applications*. Elsevier; North Holland: 2009. p. 381-396.
6. NCHS Research Data Center. Disclosure manual: Preventing disclosure: Rules for researchers. 2012. <http://www.cdc.gov/rdc/Data/B4/DisclosureManual.pdf>
7. Web-based Data Query Systems (WDQS). National Association for Public Health Statistics and Information Systems (NAPHSIS) website. <https://naphsisweb.sharepoint.com/Pages/WebbasedDataQuerySystemsWDQS.aspx>
8. Fraser, B., Wooten, J. A proposed method for confidentialising tabular output to protect against differencing. Australian Bureau of Statistics, Data Access and Confidentiality Methodology Unit; 2005.

Table 1

This table contains the categorization of each of the 35 web-based query systems reviewed in this study.

State	No Suppression Observed	Observed counts below suppression threshold	Suppressed cells recoverable	No suppression errors
Alabama	X			
Arizona	X			
Arkansas (Dept. of Health)			X	
Arkansas (Cancer Registry)			X	
California (DPH)	X			
Colorado			X	
Connecticut	X			
Florida (CHARTS)	X			
Georgia (OASIS)	X			
Illinois				X
Kansas	X			
Massachusetts				X
Maine				X
Maryland	X			
Minnesota (DPH)	X			
Minnesota (MIDAS)	X			
Mississippi	X			
Missouri	X			
New Hampshire			X	
New Jersey	X			
New Mexico			X	
New York				X
New York City				X
North Carolina	X			
Oklahoma			X	
Oregon	X			
Pennsylvania		X		
Rhode Island	X			
South Carolina	X			
Tennessee			X	
Texas	X			
Utah			X	
Washington		X		
Wisconsin			X	
Wyoming				X
Totals	18	2	9	6