12-2016

Privacy Protection and Aggregate Health Data: A Review of Tabular Cell Suppression Methods (Not) Employed in Public Health Data Systems

Gregory J. Matthews
Loyola University Chicago, gmatthews1@luc.edu

Ofer Harel
University of Connecticut - Storrs

Robert H. Aseltine Jr.
University of Connecticut

Recommended Citation

Matthews, Gregory J.; Harel, Ofer; and Aseltine, Robert H. Jr.. Privacy Protection and Aggregate Health Data: A Review of Tabular Cell Suppression Methods (Not) Employed in Public Health Data Systems. Health Services and Outcomes Research Methodology, 16, 4: 258-270, 2016. Retrieved from Loyola eCommons, Mathematics and Statistics: Faculty Publications and Other Works, http://dx.doi.org/10.1007/s10742-016-0162-8

Noname manuscript No.
(will be inserted by the editor)

Privacy protection and aggregate health data: A review of tabular cell suppression methods (not) employed in public health data systems

Gregory J. Matthews · Ofer Harel · Robert H. Aseltine, Jr.

Received: / Accepted:

Gregory J. Matthews
Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL
E-mail: gmatthews1@luc.edu
.
Ofer Harel
Department of Statistics, University of Connecticut, Storrs, CT
E-mail: ofer.harel@uconn.edu
.
Robert H. Aseltine, Jr.
Division of Behavioral Sciences and Community Health
Institute for Public Health Research
University of Connecticut Health Center, East Hartford, CT
E-mail: aseltine@uchc.edu

Abstract Public health research often relies on individuals' confidential medical data. Therefore, data collecting entities, such as states, seek to disseminate this medical data as widely as possible while still maintaining the privacy of the individual for legal and ethical reasons. One common way in which this medical data is released is through the use of Web-based Data Query Systems (WDQS). In this article, we examined WDQS listed in the National Association for Public Health Statistics and Information Systems (NAPHSIS) specifically reviewing them for how they prevent statistical disclosure in queries that produce a tabular response. One of the most common methods to combat this type of disclosure is through the use of suppression, that is, if a cell count in a table is below a certain threshhold, the true value is suppressed. This technique does work to prevent the direct disclosure of small cell counts, however, primary suppression by itself is not always enough to preserve privacy in tabular data. Here, we present several real examples of tabular response queries that employ suppression, but we are able to infer the values of the suppressed cells, including cells with 1 counts, which could be linked to auxiliary data sources and thus has the possibility to create an identity disclosure. We seek to stimulate awareness of the potential for disclosure of information that individuals may wish to keep private through an online query system. This research is undertaken in the hope that privacy concerns can be dealt with preemptively rather than only after a major disclosure has taken place. In the wake of a such an event, a major concern is that state and local officials would react to this by permanently shutting down these sites and cutting off a valuable source of research data.

Keywords privacy, confidentiality, health policy, public health, data sharing

1 Introduction

Interactive web-based data query systems (WDQS) 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 compendium compiled by the National Association of Public Health Information Systems (NAPHSIS), state-supported WDQS relevant to public health are currently hosted in a majority of states in the United States (NAPHSIS, 2016). Such systems typically provide aggregate information to users in response to specific, 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. This can potentially occur when individuals have rare conditions or 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.

Standards for protecting the privacy of individuals captured in WDQS are well-established. For instance, the "Handbook on Statistical Disclosure Control" (HSDC) (Hundepool et al, 2006) describes many methods for protecting against statistical disclosures in tabular data and Shlomo et al (2015) compare different methods of statistical disclosure control based on disclosure risk and data utility measures specifically for flexibility generated tables. The HSDC classifies methods for protecting tabular data into three broad categories: pre-tabular, table redesign, and post-tabular.

Pre-tabular methods involve modifying the actual microdata prior to the table creation step. Methods in this category include, for example, data swapping (Fienberg and McIntyre, 2005, Dalenius and Reiss, 1982), post-randomization method (PRAM) (Gouweleeuw et al, 1998), and sampling (Skinner, 2009). Data swapping involves swapping individuals' records in the microdata in such

a way that the marginal counts of variables are maintained. After data swapping, a contingency table can be created based on the swapped data. The resulting table has an added layer of protection by adding uncertainty to the output. Even if a cell in the table contains a small number, there is uncertainty as to whether the number is the actual count or not. Therefore, it is difficult to accurately identify an individual whose data is used in the table. PRAM aims to prevent statistical disclosures by perturbing each record in a data set, which practically amounts to adding random noise to each record when the data are categorical. Once this method is applied to the microdata, as with data swapping, tables can then be created based on the perturbed data, which now have an added layer of protection. Finally, sampling is another example of a pre-tabular method for protecting tabular data. Sampling simply involves selecting a subset of the data that has been collected, and then creating output tables based on the sample of the data rather than the entire data itself. This adds uncertainty to the table in that even if a small number appears in a cell, one cannot be sure that the number that is being observed is the true number from the full data set. All of these methods work, to some degree, based on the concept of uncertainty. The data consumer cannot be sure that the numbers they are observing in the cells are exactly correct, which prevents them from learning any information about an individual with certainty. At the same time, they are able to learn about the population that they are interested in studying based on the tables as much of the original structure of the data is maintained.

A common example of table redesign methods is reduction of detail. Reduction of detail involves combining cells together based on collapsing some of the variable categories together to make one new category. For instance, rather than release a table with counts that are sorted by age in groups of five

years, it is safer from a statistical disclosure point of view to use ten or twenty year age groups resulting in fewer cells in the table.

The HSDC mentions three main methods of post-tabular disclosure control, namely, cell suppression (Cox, 1980, 1984, Cox et al, 1987), cell perturbation (Willenborg and de Waal, 2001), and rounding (Cox, 1984, Cox et al, 1987, Cox, 1987). Cell perturbation adds some amount of noise to each cell to add some uncertainty to the true value of the cell. Rounding is used to hide true values of cells by releasing a number in the cell that is relatively close to the truth. For instance, rather than releasing the actual counts of cells, an organization that is releasing tabular data may round all of the cells to the nearest 5 or 10. This guarantees that no cell with a 1, which is the most vulnerable value from a statistical disclosure point of view, is released.

Finally, cell suppression is a technique that searches for cell counts that are below some predefined threshold and hides the true count in the cell by reporting the true value as missing. This technique guarantees that small cell counts, which are vulnerable to an attack by a malicious data user, will never be known exactly. The National Center for Health Statistics (NCHS) Research Data Center Disclosure Manual does not allow its researchers to publish tables with cell count less than 5; these values must be suppressed (NCHS Research Data Center, 2012, Page 15). Further, they recommend that categories be re-categorized so that cell suppression is not needed at all. However, the application of cell suppression is not as straight forward as it may seem if one wishes to provide meaningful disclosure control. Simply suppressing the true values of cells that all fall below a threshold – these are called primary suppressions – does not guarantee that disclosures will not take place. For instance, if only one cell is primarily suppressed in a row, but the row total is provided in the table, the suppressed value can be easily derived. Therefore, for the sake of disclosure control, other cells that may not fall below the suppression thresh-

old may also need to be suppressed. These types of suppression are referred to as secondary, or complimentary, suppressions. The NCHS Disclosure Manual also makes it clear that complimentary suppressions are necessary to prevent disclosures through cell subtraction. As a result, they suggest using cell aggregation rather than suppression (NCHS Research Data Center, 2012, Page 15).

Cell suppression, which is a very common method of disclosure control in many of the NAPHSIS systems, is especially problematic in the specific setting where the user has the ability to create many different tables via many different queries. HSDC states, about post-tabular methods generally, "post-tabular methods suffer the problem that each table must be individually protected, and it is necessary to ensure that the new protected table cannot be compared against any other existing outputs in such a way, which may undo the protection that has been applied" (Hundepool et al, 2006, Page 175). Further, they go on to state a problem specifically about cell suppression, "Disclosive zeroes need to be suppressed and this method does not protect against disclosure by differencing. This is a serious problem if more than one table is produced from the same data source (e.g. flexible table generation)" (Hundepool et al, 2006, Page 176). Both of these statements warn that these methods, generally, post-tabular and, specifically, cell suppression, simply do not work in many instances when the user is able to create many different tables based on the same data set. This is exactly the case with a web-based data query system (WDQS).

In spite of well-established guidelines for privacy protection related to WDQS, many systems run by states still perform inadequately with respect to privacy when data are released in tabular form. Matthews et al (In Press) reviewed 35 state-affiliated WDQS and found that one of the most common measures of statistical disclosure control employed by these systems was sup-

pression of small cell counts in tabular queries. This means that one of the more common methods of disclosure control that was observed in our review of the NAPHSIS systems is one of the methods that is least suited to protect privacy in exactly these types of systems (Hundepool et al, 2006, Pages 175–176). In fact, Matthews et al (In Press) found that in 9 of the sites reviewed, true cell values were able to be recovered even with suppression being used for small cell counts. Further, 18 of the reviewed sites did not appear to be suppressing any cells counts, and only in 6 sites were no immediate problems discovered with cell suppression.

As a result, in this manuscript, we focus on the details of suppression techniques employed by WDQS and illustrate problems with using suppression as a method of statistical disclosure control for flexibly generated tables using actual examples from various systems. It is our hope that through this manuscript institutions will become more aware of the potentially privacy issues when releasing data in tabular format and will take necessary steps to ensure that privacy for individuals in maintained.

The remainder of this article discussed the methods that were used in evaluating different query systems in section 2. Section 3 presents actual observed examples of ways in which cell suppression is being employed in some online query systems. Finally, section 4 concludes with a discussion of the impact of these findings.

2 Methods

The National Association for Public Health Statistics and Information Systems' (NAPHSIS) website provides a catalog of Web-based Data Query Systems (WDQS) (NAPHSIS, 2016) by state that provide a wide array of vital statistics. In the process of performing a full review of the types of disclosure

control techniques in place by these various systems, it became clear that cell suppression was one of the most popular methods for controlling statistical disclosure in spite of the limitations of the method specifically in this setting. Additionally, many of these query systems that were performing cell suppression were implementing the technique in a manner that would often allow an astute user to infer the exact values of the suppressed cell counts rendering cell suppression nearly meaningless in terms of protecting sensitive pieces of data.

Therefore, as a test, in cases where a WDQS was employing cell suppression, attempts were made by the authors to infer the suppressed values of the cells. To demonstrate that one need not employ sophisticated and/or time consuming techniques to recover suppressed values, the time spent trying to recover suppressed values was limited to no more than 30 minutes for each WDQS. This demonstrates that these disclosures are not the result of complex analyses, but, rather, are based on simple attack methods that are well established in the statistical disclosure literature.

Examples of queries (or possible a set of queries) where suppressed values can be recovered were captured with screen shots of the table (or tables). These screen shots act merely as a record and are not presented fully in this manuscript to offer ambiguity in which query system produced the table in which the suppressed cells were able to be recovered. Here we simply present generic tables, devoid of any identifiers, each demonstrating an example of how suppressed cell values can be recovered using different techniques. While these tables are presented as generic, we stress the fact that all of these are actual examples that were observed by at least one WDQS in the NAPHSIS database.

3 Results

Among those WDQS that claimed to incorporate SDC strategies, some form of suppression was most commonly employed, although the standards used (e.g., minimum cell size tolerated, whether both primary and secondary cell suppression strategies were attempted) varied dramatically. Some WDQS's suppressed cell counts only below 10 whereas other systems suppressed cells only for values below 3. In practice virtually all the systems reviewed did not employ SDC techniques that were effective in preventing users from obtaining results that placed individuals at risk of a statistical disclosure. In the following sections we will illustrate the shortcomings of many public health WDQS using actual examples. Again, to protect the privacy of the individuals whose personal information may be compromised, we present the results of actual queries while hiding the identity of the query systems from which they were obtained.

3.1 Examples of disclosure

In general, there are a wide array of issues involved in statistical disclosure control that are beyond the scope of this manuscript and for a full overview see O'Keefe and Rubin (2015), Matthews and Harel (2011), Skinner (2009), or Willenborg and de Waal (2001). One main issue of statistical disclosure control is the study of techniques for modifying microdata in such a way that it can be released to researchers or the public in such a manner that the data is both useful and sensitive to the privacy of the individuals involved. There are many examples of de-identified microdata being disseminated to the public that was able to be re-identified such as Sweeney (2002), Narayanan and Shmatikov (2008), and Barbaro and Zeller (2006). For guidelines for de-identifying personal health information microdata see El Emam and Fineberg (2009), for

instance. Here, however, we are only interested in disclosures occurring from data released in tabular format as output from the WDQSs examined here. There are many different types of potential disclosures when releasing data in this manner and section 5.2 of Hundepool et al (2006) describe these, which include identity disclosure, individual and group attribute disclosure, and disclosure by differencing. Further, there are many types of risks based on the motives of an adversarial data user. El Emam (2010) describes what they refer to as prosecutor risk, journalist risk, and marketer risk. These different types of risks described vary based on the background information of an adversary and their motives. Prosecutor risk, for instance, involves an adversary who is trying to re-identify a particular individual about whom some background information is know and the individual is known to be in the data set of interest. Journalist risk involves trying to target a known individual, but it is not known whether the individual is actually contained in the data with certainty, and marketer risk is when an adversary attempts to re-identify as many of the observations in the released data as possible.

One of the most common types of disclosure, identity disclosure, occurs when an individual can be identified in the released data. This would occur, for example, if a table appeared with a cell count of one indicating that the set of characteristics for that particular cell were unique. This situation presents the possibility that the individual whose record accounts for the one in a table cell can be linked to an individual's identity. This type of event, locating a cell count of 1, in and of itself, does not create a disclosure, however, it often leads to other situations in which other types of disclosures take place and, therefore, is to be avoided.

Another type of disclosure, attribute disclosure, occurs when new information is discovered about an individual from the released data. This could occur in the current setting if, upon discovering a cell with a 1 count in it based en-

tirely on demographic characteristics, a subsequent query was made inquiring about some attribute of this individual. This would lead to a situation where a data user would have a combination of demographic characteristics that uniquely defined an individual in the data as well as information about some attribute of this individual. If this type of disclosure occurs in conjunction with an identity disclosure, a link can then be made between the identity of an individual and a learned attribute, such as disease status or cause of death.

While cell counts of 1 are a problem in terms of statistical disclosure control in tables, only getting rid of 1's will not solve the problem entirely. Another type of disclosure, group attribute disclosure, occurs when all individuals in group have the same attribute. This could occur if a user identifies a cell with a small cell count, such as 2. If, upon subsequent querying, the user learns that all members of this sub-group have the same disease, for example, a private attribute has been learned about each of these individuals. Notice, that an identity disclosure does not need to take place for this to occur. No record in the data has been identified as belonging to a specific individual, but, rather, we have learned something all individuals in the sub-group.

One last type of disclosure relevant to this situation occurs form the use differencing. Hundepool et al (2006) describes three types of potential disclosures by differencing which are geographical differencing (Duke-Williams and Rees, 1998), linking, and differencing of sub-populations. In the context of the query systems we evaluated, differencing of sub-populations is very simple approach for subverting suppression of small cell counts by releasing organizations. For example, say an individual was looking for men who died in a certain county at the age of 33. For a small county, this number may be very small and, if suppression were being used, would likely be suppressed. However, rather than querying men who died at exactly the age of 33, provided the WDQS allows this type of query, one could query the number of men in a county who died

Table 1 Query Result. All Deaths. Year: 2000, Age: Under 15

	Male	Female	Total
Black	5	**1**	6
White	50	38	88
Total	55	39	94

who were 32 or younger followed by a query of men who died in the county who were 33 or younger. These two queries are likely to return results that are not suppressed because the counts are likely to be sufficiently large. However, the difference between these two queries is exactly the suppressed value of the original query. Fraser and Wooten (2005) propose a method to protect tabular data against disclosure by differencing.

3.1.1 No Suppression

As an example of a potential disclosure in the presence of no employed methods of disclosure control, consider the following set of two queries. In the first query, the request is for a table based on all deaths of individuals under the age of 15 occurring in the year 2000 separated by race on the rows and gender in the columns. An example of a query appears in table 1. From this table, a user learns that based on year of death, age group, race, and gender that a unique combination can be created. The one in the upper right cell of this table indicates that only one black female under the age of 15 died in this state. Therefore, a subsequent query based on data of all deaths in the year 2000 of individuals under the age of 15 whose gender was female and race was black is based on only one data point. Therefore, if we based a query on this data and stratify by some attribute that we are interested in learning, say cause of death, without any disclosure control, the exact cause of death can be learned. Table 2 is based on the one data point that was identified from

Table 2 Query Result. All Deaths. Year: 2000, Age: Under 15, Race: Black, Gender: Female

Cause of Death	Count
Heart Disease	0
Cancer	0
HIV/AIDS	1
Homicide	0
Suicide	0

a previous query and is now stratified by cause of death. This second query allows us to learn that the cause of death was HIV/AIDS.

In summary, what a data user has learned is that within a given state only one black female under the age of 15 died in the year 2000, and that the cause of death for this individual was HIV/AIDS. Granted, no definite disclosure has yet occurred. However, it is not difficult to imagine a way of identifying exactly who this individual is especially in the digital age. Further, assuming that this individual could be identified, they are not the only person who's privacy have been potentially compromised. Specifically, in this case a user has learned that a child has died of HIV/AIDS. This likely indicates that the mother and/or father of this child is carrying the virus, which they then passed on to their child. This is an especially important aspect of this example to note that private attributes of individuals' who are not even in this database are potentially at risk here.

3.1.2 Suppression

Encouragingly, many states at the very least acknowledge the issue of potential statistical disclosure and have taken steps to control this. One of the most commonly observed methods of disclosure control among the state systems examined here was cell suppression. However, simple implementation of this method is not enough to control disclosures in many situations. For example,

assume that the suppression rule for a query system is to return all cells with a cell count less than or equal to 3 as missing values. This is done in an attempt to prevent small cell counts, especially 1's, from appearing in table cells, which could lead to identity and attribute disclosure. However, consider the following query based on data from a particular county of all men between the ages of 20-24 who died in 2008 separated by race in the columns. The results of this query are presented in table 3. The returned table contains three cells: one for black, one for white, and one for total. As a result of a small cell count in the cell corresponding to a race of black, the true value is suppressed in compliance with disclosure control protocols set forth by this table. Rather than returning the true value of the cell, an "X", for example, is placed in the table where the actual value would have appeared. However, even with suppression correctly applied to this table, it should be obvious, without even submitting another query, that the true value that is suppressed is a 1. The cause of this problem here is that only primary suppression (Cox, 1980) is applied to the table, whereas, from a disclose perspective, complementary suppression would aid in limiting disclosures here. Complementary suppression (Cox, 1980) involves suppressing cells that by themselves do not meet the requirement for suppression, but can be used to aid in discovering the true values of the primarily suppressed cells.

Returning to the original table, now that a 1 has been identified in a suppressed cell, a subsequent query can be used to attempt to create an attribute disclosure about this individual. This query is based on the same data as the original query with additional subletting to include only those individuals whose race is black. Based on the first query, it is known that the total number of data points that this query is based on is 1. Therefore, any resulting table can only have a 1 in it, even if the value is suppressed. The results of a query of this kind can be found in table 4. From this table, we can infer that the cause

Table 3 County: X, Age: 20-24, Sex: Male, Year: 2008

Race	Count
Black	X
White	9
Total	10

of death for this individual was cancer. In review, a data user has learned, based only on these two queries, that there was one black male who died in a given county in the year 2008 who was between the ages of 20-24 and that the cause of death was cancer. Further, the data user has learned all of this information using only two queries even in the presence of cell suppression to control against disclosures. This is a particularly troubling example because the operators of this query system are aware that statistical disclosure poses a threat to individual privacy, which can be inferred by their attempts to address this risk, but the disclosure controls are being implemented in such a way as to be practically doing almost nothing.

Table 5 presents the results of a query of all females aged 25-44 in a certain county by year and race. Cell suppression is implemented here as intended, however, because row and column totals are not suppressed the values that are suppressed can be filled in exactly with nothing more than subtraction. In both of these examples, suppression is implemented as intended, but, in both cases, the actual benefits from a disclosure control perspective are essentially nonexistent. Further, these tables, by themselves, do not protect the values of the suppressed cells, and, therefore, these tables could not even be released in static form. Here, however, a user has the ability to flexibly create other tables, they don't even need to use that extra flexibility to learn the true values of suppressed and supposedly protected cell values.

Table 4 County: X, Age: 20-24, Sex: Male,Year: 2008, Race: Black

Race	Count
All Causes	X
Cancer	X

Table 5 County: X, Age: 20-24, Sex: Male,Year

Year	Black	White	Other	Total
1990	82	6	X	89
1991	65	X	X	73
Total	147	10	X	162

3.1.3 Subverting Suppression with multiple queries

The previous examples involving suppression demonstrated instances where the true values of suppressed cells can be learned directly from one query. However, in other situations it may only take one or a few additional queries to infer the true values of suppressed cells. Consider, for example, that a data user discovers from a previous query that there was only one white male in a given county who died in the year 2008 who was between the ages of 25-44 at the time of his death. The user can then submit a query based on only this subset of attributes and separate the rows by cause of death. By restricting the data to certain causes of death, the user can eventually get to a result as in table 6. Therefore, the user knows the cause of death is either an accident or a suicide. While the cause of death is not known for sure at this point, it could be argued that this is already a disclosure as the user has learned that the cause of death was not natural.

However, with only one additional query, restricting cause of death to only accidents, the exact cause of death can be learned. Whereas the true value in the accidents cell was previously suppressed, in table 7 it is no longer suppressed, and it is revealed that the true value is zero. Now that the user knows

that there were zero accident deaths, the exact cause of death here is reveled to have been suicide. Again, in summary the user has a very specific and unique demographic description of this individual and, through multiple queries, was able to subvert suppression and learn the exact cause of death. Imagine if a health provider ever release a raw data record with demographic information that is this specific along with cause of death. This would probably not be tolerated.

One aspect of this example that should be noted is that the true value of a suppressed cell was learned because a cell with a zero in it was not suppressed. Often it is assumed that there is no danger in releasing the true value of a cell if that value is a zero. However, knowing that a cell is zero decreases the uncertainty that a data user has about the true values of suppressed cells. Cell values of zero can not simply be ignored from a disclosure perspective. They are an important part of reducing disclosures in tabular count data.

Next, consider a table as in figure 1 based on data of all African-American men who died in 2009 in a given state broken out by county and age group. Due to suppression rules, small cell counts are being suppressed. It also appears that, along with primary suppression, secondary suppression is also being applied here as indicated in the state's explanation about suppression found below the returned table. It is encouraging to see that some states recognize that primary suppression by itself is simply not enough to prevent disclosures, and this state has taken the next step in privacy by using secondary suppression. As a result, this table, considered alone, appears to be adequately protecting the true values of the suppressed cells, though it is unclear why some cells have unsuppressed ones appearing in the table. Assuming that the cells with one counts are determined to not pose a risk of disclosure, the cells that are suppressed can be easily learned by submitting only one more well created query. In this instance, a user can simply repeat the exact same orig-

Table 6 County: X, Age: 25-44, Sex: Male,Year: 2008, Race: White, Cause of death: Accidents or Suicide

Cause of Death	Count
Accidents	X
Suicide	X
Total	1

Fig. 1 An example of an implementation of complementary suppression that still allows suppressed values to be recovered.

African-American men who died in 2009

	Age																	
County	**0**	**14**	**18**	**21-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65-69**	**70-74**	**75-79**	**80-84**	**85+**	**Total**
a	1	1	1	1	3	3	1	4	3	4	5	4	9	6	9	4	4	**63**
b															**			**
cd												1				**		**
ef									1				1	3	**			**
g	1								1		1	3	1	2	1	**		**
h			**		2					1					**	**	**	**7**
i							1			1	1	1	1		1		**	**
j										**				**	**			**3**
k								1			**	**	**				**	**6**
l	1												1	1	1	**		**
m												**						**
n											**							**
o										1		2		1			4	**8**
p						1						1				**	**	**5**
q												**						**
r	**			1								1			**			**4**
s															**			**
t												**						**
u													**			**	**	**3**
Total	**	1	**	2	5	4	2	5	5	**	9	18	16	**	19	15	13	**142**

inal query and just remove the breakdown by age. This returns a table with only one column, where each row is the total count for a county. As a result of this query, the suppressed row totals from the original table, which were missing due to secondary suppression, can be filled in. This allows a user to fill in every suppressed cell value in the table so that, practically speaking, any protection originally provided by cell suppression is essentially voided by performing only one additional query and arithmetic.

Table 7 County: X, Age: 25-44, Sex: Male,Year: 2008, Race: White, Cause of Death: Accidents

Cause of Death	Count
Accidents	0

4 Disclosure by differencing

In addition to the aforementioned problems with suppression, some query systems failed to address a potential attack via differencing of tables (Australian Bureau of Statistics, 2005). This type of disclosure could occur when a user poses a query with very specific row or column specifications. This would cause many of the values in the returned cells would be small and thus sensitive. For example, suppose a user wanted to know how many instances of a disease there were by year for a specific age (e.g. 35). If these are small counties these values will likely be very small and therefore suppressed. Table 8 show a possible query result for a situation like this and suppresses all of the counts because they are too small (Columns 1 and 2). However, rather than querying how many 35 year olds have a specific condition from a certain county, a user could pose two additional queries that are slightly different and learn the exact values of the suppressed cells in the original query. For example, a user could ask how many people 35 years of younger have the condition and then how many people 34 years or younger have the condition. By posing the queries this way, in this state, the results will often be large enough to avoid most system's suppression rules for small cells. While this type of disclosure is certainly a problem even for static tables, it is an especially troublesome problem when many queries are allowed and tables can be flexibly generated.

Table 8 Differencing: Race: White, Gender: F, County: A, Age: 35

Year	Age=35	Year	Age≤34	Year	Age ≤ 35
1991	X	1991	23	1991	24
1992	X	1992	25	1992	26
1993	X	1993	22	1993	24
1994	X	1994	22	1994	23
1995	X	1995	19	1995	21
1996	X	1996	17	1996	18
Total	8	Total	128	Total	136

Each set of two columns in table 8 displays the results of each of three individual queries. The first two columns show the results when querying by race, gender, county, and specific age, in this case 35. The individual values are suppressed due to small cell counts. However, by changing the query criteria of age to less than or equal to 34, no results in the resulting query, columns 3 and 4, are suppressed due to small cell counts. A data snooper could then query again changing the age criteria to less than or equal to 35 resulting in the table displayed in columns 5 and 6, again with nothing suppressed because the counts are large enough. As a result of the manner in which these queries were devised, the difference between the result of the third query and the second query will result in exactly the suppressed values of the first query. This is an example of disclosure by difference that was actually observed in a WDQS, which, essentially renders suppression meaningless in this context. This is yet another example where the large amount of flexibility a user has in posing the query, while extremely useful, presents many potential privacy problems.

5 Conclusion

The results of this inquiry into the privacy protection measures utilized in state authorized WDQS are cause for serious concern. While many systems recognized the need to employ statistical disclosure control techniques, the

most common method employed was cell suppression, which is particularly poorly suited for this specific setting (i.e flexible table generation). Further, many systems that employed cell suppression deployed it inadequately. We would encourage organizations that release tabular data to follow the guidelines put forth in the Handbook for Statistical Disclosure Control (Hundepool et al, 2006).

As a result sensitive information from individuals whose attributes made them easily identifiable are freely available to those who choose to seek that information. Potential privacy issues related to causes of death were astonishing, ranging from HIV to cancer. Most distressing is that this information was so easily obtained; In no case did any of the disclosures presented above require more than 30 minutes of work.

Aside from the obvious invasion of privacy that such disclosures may entail, the pernicious (malicious?) use of such data can easily be used to cause tangible harm to individuals. Unscrupulous employers and insurers, currently under mandates by the ACA to not deny coverage for pre-existing conditions and with lifetime coverage limits eliminated, might be well rewarded for vetting potential employees and insured infectious diseases such as HIV that might be transmitted to loved ones or cancers that might have a genetic basis.

Given the explosion of de-identified patient data that will be available with the emergence of Health Information Exchanges (HIEs), it is critical that we develop query systems that will not compromise privacy. In addition to the public health importance of these data, many HIEs will rely on the dissemination of ostensibly de-identified data as part of their business plan. As we have seen, even the provision of aggregate data poses privacy risks, and the release of record-level information, so critical to efforts to provide population assessments, improve treatment outcomes, and facilitate quality improvement, carries even greater risks.

Data use agreements are a critical tool in safeguarding privacy, but consider this: would a hospital or healthcare provider allow open access to a medical records office so long as all visitors signed a pledge that they would not look at anything they weren't supposed to? Of course not, yet without restrictions on access to what is provided in WDQS this is essentially what a data use agreement does.

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.

Compliance with Ethical Standards

Conflict of Interest

None.

Human and Animal Rights

This article does not contain any studies with human participants or animals performed by any of the authors.

References

Australian Bureau of Statistics (2005) A proposed method for confidentialising tabular output to protect against differencing. Joint UNECE/Eurostat work session on statistical data confidentiality

Barbaro M, Zeller T Jr (2006) A face is exposed for AOL searcher No. 4417749. *New York Times*, 9 August 2006. Available: http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=0 Accessed 9 March 2016

Cox LH (1980) Suppression methodology and statistical disclosure control. Journal of the American Statistical Association 75:377–385

Cox LH (1984) Disclosure control methods for frequency count data. Tech. rep., U.S. Bureau of the Census

Cox LH (1987) A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association 82:520–524

Cox LH, Fagan JT, Greenberg B, Hemmig R (1987) Disclosure avoidance techniques for tabular data. Tech. rep., U.S. Bureau of the Census

Dalenius T, Reiss SP (1982) Data-swapping: A technique for disclosure control. Journal of Statistical Planning and Inference 6:73–85

Duke-Williams O, Rees P (1998) Can census offices publish statistics for more than one small area geography? an analysis of the differencing problem in statistical disclosure. International Journal of Geographical Information Science 12(6):579–605

El Emam K (2010) Risk-based de-identification of health data. The IEEE Computer and Reliability Societies

El Emam K, Fineberg A (2009) An overview of techniques for de-identifying personal health data. Access to Information and Privacy Division of Health Canada

Fienberg SE, McIntyre J (2005) Data swapping: Variations on a theme. Tech. rep., National Institute of Statistical Sciences, Research Triangle Park, NC

Fraser B, Wooten J (2005) A proposed method for confidentialising tabular output to protect against differencing. Joint UNECE/Eurostat work session on statistical data confidentiality

Gouweleeuw J, P Kooiman LW, de Wolf PP (1998) Post randomisation for statistical disclosure control: Theory and implementation. Journal of Official Statistics 14(4):463–478

Hundepool A, Domingo-ferrer J, Franconi L, Giessing S, Lenz R, Longhurst J, Nordholt ES, Seri G, paul De Wolf P (2006) A CENtre of EXcellence for Statistical Disclosure Control Handbook on Statistical Disclosure Control Version 1.01

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

Matthews GJ, Harel O, Aseltine, Jr RH (In Press) A review of statistical disclosure control techniques employed by web-based data query systems. Journal of Public Health Management and Practice

NAPHSIS (2016) NAPHSIS web-based data query systems (WDQS) webpage. https://naphsis-web.sharepoint.com/Pages/WebbasedDataQuerySystemsWDQS.aspx, accessed 9 March 2016

Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. Proceedings of the 2008 IEEE Symposium on Security and Privacy pp 111–125

NCHS Research Data Center (2012) Disclosure manual: Preventing disclosure: Rules for researchers. http://wwwcdcgov/rdc/Data/B4/DisclosureManualpdf p 15

O'Keefe CM, Rubin DB (2015) Individual privacy versus public good: protecting confidentiality in health research. Statistics in Medicine 34(23):3081–3103, DOI 10.1002/sim.6543, URL http://dx.doi.org/10.1002/sim.6543, sim.6543

Shlomo N, Antal L, Elliot M (2015) Measuring disclosure risk and data utility for flexible table generators. Journal of Official Statistics 31(2):305–324

Skinner C (2009) Statistical disclosure control for survey data. In: Pfeffermann, D and Rao, C.R. eds. Handbook of Statistics Vol. 29A: Sample Surveys: Design, Methods and Applications, pp 381–396

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

Willenborg L, de Waal T (2001) Elements of Statistical Disclosure Control. Springer-Verlag