Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data

Guidelines and Resources for Health Data Organizations

NAHDO-CDC Cooperative Agreement Project
CDC Assessment Initiative

December 2004

The National Association of Health Data Organizations (NAHDO)

Acknowledgements

The National Association of Health Data Organizations (NAHDO) developed this document with funding and support from the Centers for Disease Control and Prevention (CDC) Assessment Initiative. We are grateful for the leadership and guidance provided by the CDC Project Team, Patricia Schumacher and Alex Charleston.

The lead author of this document is Barbara Rudolph, Ph.D., NAHDO's Data and Research Senior Scientist Consultant. Dr. Rudolph integrated documents and research conducted by Michael Stoto, Ph.D., Harvard University; and Luis Paita, Ph.D., NAHDO's former Deputy Director. Guidance for this paper was provided by the NAHDO-NAPHSIS Expert Panel consisting of Garland Land, Missouri Department of Health and Human Services; Lois Haggard, Ph.D., Utah Department of Health; David Solet, Ph.D., King County Health Department; Gregory Crawford, Kansas Department of Health and Environment; and Daniel Goldman, M.D., M.P.H., Vital Net. We thank the Assessment Initiative states and the private and public state health data organization staff who selflessly provided input and shared their data dissemination practices with us.

The NAHDO-CDC Cooperative Agreement is a joint project promoting collaboration and partnership with the National Association of Public Health Statistics and Information Systems (NAPHSIS). As national associations representing statewide population-based health data bases and the agencies that maintain them, NAHDO and NAPHSIS, with support from CDC, are working to build data and technical capacity at the state and local levels.

Table of Contents

Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data

A. Introduction

Public health agencies are increasingly disseminating health statistics on the Internet. Researchers are increasingly using public health data sets for health services research, including longitudinal and cross-market studies. Broader use and dissemination of public health data sets serves a public good, highlights public health's important role in health and health improvement, and places additional value on public health data assets. Local public health and community based non-profit agencies also rely on easily accessible public health data. Yet, there is more risk of personal disclosure of sensitive information when it is displayed on the Internet; the Internet is an impersonal access tool that increases the velocity of interactions and as a result allows for rapid use and dissemination to others who may or may not have a good understanding of appropriate data use.

The confidentiality issues of greatest concern are discovering the identity of someone who is represented in a public health database and discovering that person's personal or medical characteristics through tabulated data. Depending on the nature of the database, the knowledge that someone is in it can itself be harmful. The likelihood of disclosure is higher when there are relatively few people with knowable demographic characteristics such as sex, age, and race in a small community.

This document addressing the management and technical disclosure controls for micro-data in public health web-based data dissemination systems is part of a guideline set—all aimed at assisting the public health data manager in designing or updating web-based information systems. These guideline sets include: Statistical Approaches for Small Numbers**:** Addressing Reliability and Disclosure Risk in Web-Based Data Dissemination, Security of Data for Web-based Data Dissemination Tools, and Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data; as a package, the guideline set will address: reduction in the risk of inappropriate disclosure of sensitive information, provision of reliable statistics, and increased security of the data.

B. Summary of approaches

Within this document there are two broad headings under which a variety of approaches exist. We have labeled the first "management and institutional controls;" the second we titled, "data modification and alteration methods." Both should be applied to web-based micro-data dissemination systems by public health agencies as they advance and expand their dissemination agendas. In a separate guideline we describe statistical methods for data modification and appropriate interpretation.

Multiple protective layers to assure anonymity and confidentiality should include the management and technical controls and data modification and alteration described below:

1. Data protection agreements: Required for the release of Limited Data Sets under HIPAA, these agreements are designed to both inform users and control use of the data.
2. Limited Data Set: A subset of the full data set is created for public use, dropping identifiable data elements.
3. On-line query system: Users are not allowed to download or obtain copies of raw data files. Instead data reside on the host machine. The users conduct their own analysis by submitting queries and obtaining calculated results. Downloads of the results in useable formats may or may not be permitted.
4. User authentication and Access Validation: Password protection to CD-ROM and public use files and for access to web query systems.
5. Education and Training of public use file users: The data-providing agency educates users about the disclosure risk of micro data, the types of analyses that are considered breeches of confidentiality, and the legal issues associated with disclosure.
6. Making preconstructed tables and pivot tables available. Pre-constructed tables allow review of results prior to release and prevention of specific queries that might lead to disclosure of personal identity and health information. Pivot tables, while still controlling the release, do allow for some alternative displays of the data for the end-user.
7. Anonymizing/de-identifying data. Anonymizing a micro data file by removing information such as names, addresses, policy numbers, etc.
8. Cross-tabulations and micro-aggregation: The display is fixed in terms of number of rows and columns and/or the data is aggregated to avoid disclosure.
9. Restriction of geographic detail. Rare events or events occurring in small geographic areas are removed or altered to avoid disclosure.
10. Recoding into intervals and rounding. Grouping values of data elements that are continuous or rounding up to a higher number.
11. Cell suppression. Removing data values below pre-determined cell sizes and rules regarding the display of margins.

In addition to the approaches listed above, there are now several new software packages that provide technical support for protecting public health data. The National Center for Health Statistics (NCHS) sponsored the development of disclosure limitation software for two-way tables by OptTek Systems, Inc.. The OptTek software includes the following functionality:

- cell suppression
- controlled rounding (minimum-distance controlled rounding)
- unbiased controlled rounding
- controlled rounding subject to subtotal constraints
- synthetic substitution (controlled tabular adjustment)

For more information on the NCHS tool contact Larry Cox at NCHS. The second tool was created by RTI International and it is called MASSC^{SM} and it focused on reducing disclosure risk for surveys where sampling methods have been used. For additional information on this tool, contact Dr. Michael Samuhel at samuhel@rti.org .

While this document is focused at maintaining control over information, we must remain cognizant of the need for information and therefore, not over-protect the information. As noted in the Introduction, the use and dissemination of public health data serves a public good.

We have artificially separated the statistical approaches and security of the data from the approaches listed above, but the data manager will likely want to incorporate those approaches as well. We have provided references as available for further review, and some examples from public health agencies across the country. We have also described the strengths and weaknesses of each approach.

C. Management and Institutional Controls

There are a variety of tools available for web-based data dissemination which can reduce disclosure risk. These tools vary from user-directed education and agreements to tools that alter the access to data on the system.

1. Data protection agreements

Public health agencies have been using data release agreements for years; these agreements may be quite restrictive. The Internet has changed the format of data protection agreements (DPA). In open query systems, site users are not required to submit their name and do not need to sign a paper agreement but rather may be required to read a web document and click on a button indicating they have read the document and agree to follow the rules outlined in that document. The effectiveness of this approach has not yet been tested to our knowledge. Not knowing the effectiveness suggests that this DPA is not an approach that could be used as a stand-alone; it must be used in combination with other approaches. In closed web-based systems, data protection agreements are generally required prior to password assignment and log-in.

References:

NAHDO Web-based Data Dissemination Systems Users Group Web cast, NAHDO-CDC Cooperative Agreement, June 2002
HIPAA Workshop, Sponsored by DHHS, conducted by NAHDO, December 2003

Agencies Using Approach

South Carolina’s Office of Research and Statistics

Massachusetts Department of Public Health

Strengths and Limitations of Approach:

The DPA establishes institutional control, is compliant with HIPAA Privacy provisions for a limited data set, and serves to educate public health staff and data users as to their due diligence and legal obligations to protect the data and properly use the data.

2. Limited data sets

A reduction in the number and type of data elements is a likely choice for most web-based systems. While this can provide a reduction in risk it also limits the types of questions that can be answered from the remaining data. It is one the methods suggested by HIPAA regulations for the release of health data by covered entities. Many public health agencies are exempt from this regulation; however, it is likely that most public health agencies are influenced by it, or operate under public health laws with similar requirements.

References:

HIPAA Privacy Rule http://www.hhs.gov/ocr/hipaa/

Agencies Using Approach:

AHRQ's HCUP system for hospital discharge data
http://www.ahrq.gov/data/hcup/
Numerous state agencies

Strengths and Limitations of Approach:

A set back from this approach is the loss of specificity in some data elements and the fact that by eliminating confidential data elements for linking data across time or institution are not available for longitudinal studies, or for episode of care analyses. An advantage is that a limited data set streamlines the data acquisition process by not requiring IRB approval and yet still supports most statistical studies.

3. On-Line query systems limits

Web systems use data modification and alteration methods and rely on limited datasets to ensure protection of native files. It is very important for system developers to create a new dataset that is separately housed from the native file to prevent file corruption and access by unauthorized users. Query systems can also be designed to return limited tables and pivot tables, also limiting the risk of identification. Systems can also be configured to limit access to micro-data through logon access and upfront data user agreements.

Agencies Using Approach:
Florida: http://www.floridacharts.com/charts/chart.aspx
Kansas: kic.kdhe.state.ks.us/kic/
Missouri: http://www.health.state.mo.us/GLRequest/MICAdef.html
Pennsylvania: http://www.phc4.org/Default.htm
Tennessee: http://oit.utk.edu/helpdesk/
Utah: http://www. http://ibis.health.utah.gov/
Washington: Vista/PHw - Washington State Center for Health Statistics
(www.doh.wa.gov/OS/Vista/homepage.htm)
https://fortress.wa.gov/doh/epiqms/
Wisconsin: http://www.dhfs.state.wi.us/healthcareinfo/qsmain.htm

Strengths and Limitations of Approach:

A clear strength is the systems ability to help the novice user query the data base and generate custom-made tables "on the fly" in seconds, without requiring any programming or statistical skills on the part of the user. Suppression of numbers for rural areas and subgroup characteristics results in loss of information to user. Queries may not support detailed analyses, but rather serve as a preliminary study tool.

4. User authentication and access validation

It is possible to implement password protection to CD-ROM and public use files and for access to web query systems. Other less restrictive alternatives include simply requiring registration of the user for each use. Access to the system may be restricted to only those who have completed a data use agreement.

Agencies Using Approach:

Missouri Department of Health and Senior Services
Missouri Information for Community Assessment, (MICA)
Contact: Garland Land, Director
Center for Health Information Management and Evaluation
Missouri Department of Health and Senior Services
P.O. Box 570
Jefferson City , Mo. 65102
573-751-6272
landg@dhss.mo.us

Utah Department of Health, Indicator Based Information System (IBIS)
Contact: Lois Haggard, Ph.D., Utah Department of Health
loishaggard@utah.gov

Strengths and Limitations of Approach:

This is relatively inexpensive and technically it is not difficult to implement. This technique provides added protection to data access and supports tracking of users. Administration costs include set up and maintenance of a logon/identification process and potentially monitoring of use.

5. Education and training of public use file users

Some web-based systems are complex enough to recommend that there is appropriate training of users. This will limit the number of users of the system—unless the training mechanism is also a web-based system. For example, the user could be required to pass a short test taken from training material on the website. This is a technique used by a number of large universities in regard to human subjects' provisions. Many Federal agencies have "user training" for database users. Medicare has established training centers for users of Medicare claims data. New methods such as web based training sessions can be done using technology for "WEBINARS".

Agencies Using Approach:

Washington State Department of Health, Data Users Conferences
National Center for Health Statistics, NHIS data file user training
AHRQ, HCUP users training http://www.ahrq.gov/data/hcup/

Strengths and Limitations of Approach:

In-person training requires a substantial investment by the data providing agency. Less costly methods can be implemented via web seminars. Those who attend training may get better access to data with less modifications/alterations, thereby permitting improved analyses. Training requirements also place limitations on the number of individuals who can be certified to use the data, especially if in-person methods are used. Few individuals can invest the time and expense to travel to attend training sessions.

6. Pre-constructed tables and Pivot Tables

Some query systems are constructed to produce only those tables that have been pre-designed by the data agency. Others allow the user to implement the pivot functionality.

Both of these approaches limit the types of queries and the output from those queries, protecting the data from misuse. These forms of output would require that a significant number of queries would be run before one could potentially put together all the underlying data on an individual within the data, and to learn something new about that individual or to be able to identify the individual within the data.

D. Data Modification and Alteration

These are various technical strategies to protect public health data and include methods that can modify or alter the data file, reducing the probability that individuals can be uniquely identified in some way.

Data modification and alteration methods can be relatively complex, although the most complex methods are usually associated with the need for better statistical reliability (this is discussed further in the "Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk" document. Their application to public health data sets may require significant new programming of many web-based systems and analytic/source file programs. These techniques will significantly alter the information available in the individual micro-level records and could reduce the utility of the data sets to some of their primary customers. This, however, may be a better option compared with alternatives such as data aggregation and suppression of cells containing small numbers. While it may be possible to provide more details given the use of these new methods, it may come at the cost of statistical versus "real" data. Each method has strengths and weaknesses (See Strengths and Limitations of Approach).

Described below are methods which are based only on alterations to the existing data—not included are techniques that swap in data from other geographic locations or synthetic data from statistical modeling approaches—these approaches are found in the "Statistical Approaches for Small Numbers.." document.

1. Anonymizing/de-identifying data: Anonymizing a micro data file by individuals is the most common method of data modification.
2. Cross-tabulations and micro-aggregation: Data are presented in tabular format (individual data are not released). For continuous variables in the data, means, variances, and covariances may be released.
3. Restricting geographic detail: For rare events resulting in small numbers, geographic details may not be made available.
4. Recoding into intervals and rounding: Grouping values of data elements that are continuous (e.g., date of birth recoded into age categories), resulting in ordinal variables with discrete values.
5. Cell suppression : Removing data values from the cell based on pre-determined cells sizes and rules regarding display of margins…

References:

Risk of Disclosing Individually-Identifiable Information from Public Use Hospital Patient Discharge Data Files, Braday. H., Duffy, L., Powell, A., UC Data Archive and Technical Assistance, UC Berkeley, March 2002.

Discharge Data: Assuring Confidentiality While Providing Timely and Meaningful Information—is it possible? Rudolph, B., University of Wisconsin, 2003.

1. Anonymizing/de-identifying data files

State agencies often use various encryption algorithms for unique patient identifiers, transforming the identifier into a stable unique number. This number and its' linkage to the individual are separated and stored apart from each other in locked files, bank vaults, other secure arrangements. This is also a requirement of HIPAA privacy regulations.

References:

UC Data Archive & Technical Assistance, February 2002.
http://odwin.ucsd.edu/idata/

Agencies Using Approach:

At least 17 state health data agencies use an encrypted ID.

AHRQ HCUP data system http://www.ahrq.gov/data/hcup/

Strengths and Limitations of Approach:

This method by itself may not adequately control disclosure risk since other characteristics included on the file could be used to associate or "construct" an individual's identity with a record on a micro data file. For example, probabilistic matching techniques using variable such as age, gender, Zip code, date of hospitalization discharge, etc., can be used to link individuals to events when there are public records of the event (motor vehicle accident, other highly unusual circumstances), or knowledge of the individual. The benefits associated with an encrypted ID include being able to link across healthcare events allowing alternative forms of analysis such as analysis of episodes of care for chronic conditions.

2. Cross-tabulations and micro-aggregation

Nearly all systems produce aggregated cross-tabulations in order to reduce the risk of identity. For example, individuals can be aggregated according to age bands and gender for each of the cross-tab columns (or rows) depending on the question asked. This makes it very difficult to identify the individuals within those cells, as long as there is a large enough population and adequate cell sizes. Depending on the size of the underlying population and the statistic being used "adequacy of cell size can be as few as 3 or as high as 25-30 cases."

Agencies Using this Approach:

Nearly all state data organizations use aggregations to address small cells and data reliability in their web-based systems.

Strengths and Limitations of Approach:

Aggregation creates an acceptable cell size for statistical purposes. However, in order to achieve this result, specific information may be lost in the aggregation process. Aggregation should be carefully applied given the projected specific purposes for the web-based data system. It can result in serious loss of information for analysis, yet identity could be inferred from tables if no other techniques are applied by using multiple tables.

3. Restricting geographic detail

An example of this type of restriction of geographic data elements is when: in-state zip and out-of-state zip codes with less than 30 discharges in a calendar year are coded at the county or state level respectively. This reduces the probability of identifying an individual based on their location within a zip code. This type of reduction in information can be problematic however, for those seeking information on rural areas and the access to healthcare.

Agencies using this Approach:

Utah Office of Healthcare Statistics, Utah Department of Health, Center for Health Data, Utah Department of Health, Salt Lake City, UT 84114-2101, Phone: 801-538-9191 or contact: loishaggard@utah.gov

Wisconsin Bureau of Health Information, Department of Health and Family Services http://www.dhfs.state.wi.us/healthcareinfo/qsmain.htm

Strengths and Limitations of Approach:

This approach creates a greater pool of individuals within a geo-area, allowing statistical tests to be used. Community level information may not be available for planning at that level.

4. Limiting the number of data elements in a micro file

For example, this may be used for special handling of sensitive diagnoses: age, sex, and zip code are encrypted if the discharge involves Major Diagnosis Code (MDC) "25-Human Immunodeficiency Virus Infection" or Diagnosis Related Groups (DRG) "433, 521-523 - Alcohol/Drug Abuse or Dependence." This assures that individuals with HIV cannot be identified from use of the file.

Strengths and Limitations of Approach:

This limits the probability of uniquely identifying an individual, while preserving useful information for health assessment, health planning, and utilization studies. There may be a loss of specificity for some research and public health applications.

5. Recoding into intervals and rounding

For example, date of birth may be mapped into 5 year age categories; or individuals over 80 years of age are grouped together, while younger ages may be in 5 year categories. This is done to prevent disclosure of individuals in categories where there are only a few individuals. Rounding might be used for age, or for variables such as family income.

Agencies Using Approach:

Numerous state agencies including:

Utah Office of Healthcare Statistics, Utah Department of Health, Center for Health Data, Utah Department of Health, Salt Lake City, UT 84114-2101, Phone: 801-538-9191 or contact: loishaggard@utah.gov

Wisconsin Bureau of Health Information, Department of Health and Family Services, http://dhfs.wisconsin.gov/stats/queries.htm

Strengths and Limitations of Approach:

Many state agencies recode dates (birth, admission, discharge) for web-based data dissemination systems, adding a protective layer to the data by reducing risk of re-identification. However, researchers may need exact dates for linking or specific analyses, or the submitting healthcare provider may need information for quality assurance activities. With proper permission and encryption codes, the identifiable information can generally be reconstructed for linkages, etc.

6. Cell suppression methods

De-identified health information displayed in tables, whether web-based or document-based, can still result in re-identification when cell sizes are small. The primary means for protecting confidentiality in web-based data dissemination systems, as in more traditional dissemination systems, is the suppression of "small" cells, plus complementary cells, in tables. This approach often results in a substantial loss of information and utility. Alternative approaches include "perturbation" methods such as "data swapping" and "controlled rounding" that can limit disclosure risk while maximizing information available to the user. These approaches are described in the "Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk" document.

E. Rule Flaws & Strenths

In this section, we will describe several approaches used in public health agencies for suppression algorithms, the reader should note that the Missouri/Garland Land approach has been supported by the National Center for Health Statistics. Arguably, each approach "rule" has both flaws and strengths.

1. "The Numerator Rule"

The numerator rule is designed to prevent the release of information when there are fewer than *x* individuals in a given category. Complementary categories (cells in the same row or column of a small cell) must also be suppressed to avoid discovery of the number of cases by subtraction. For instance, suppose there were 10 AIDS deaths among men in a small community. Reporting that 9 of the decedents were White men is tantamount to saying that 1 was Black. With complementary suppression data quickly become unusable.

The best rationale for "*numerator-based*" data suppression is confidentiality protection, not statistical reliability. Suppression rules generally work well in protecting identity but may not prevent someone trying to uncover certain characteristics. Because marginal counts or complementary cells, including ones with large numbers, must also be suppressed in order to prevent calculation of the non-reported cell, the information lost can be substantial. [See Mike Stoto's paper, page 32.] There are algorithms to minimize the number of complementary cells that must be suppressed, but they do not guarantee non-identifiably (Federal Committee on Statistical Methodology, 1994).

References:

Doyle P, Lane JI, Theeuwes JJM, and Zayatz LM, eds., 2001. *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*. Amsterdam: Elsevier Science BV.

Duncan GT, 2001. Confidentiality and statistical disclosure limitation. In *International Encylopedia of the Social and Behavioral Sciences* (cited in Duncan et al., 2001).

Federal Committee on Statistical Methodology, 1994. Statistical Policy Working Paper 22 – Report on Statistical Disclosure Limitation Methodology. Washington: Statistical Policy Office, Office of Management and Budget.

Fienberg SE, Makov UE, Steele RJ, 1998. Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* 14: 485-502.

Stoto, Michael. Statistical Issues in Interactive Web-based Public Health Data Dissemination Systems, RAND Health, September 19, 2002.

Agencies/Systems Using Approach:

Utah Department of Health Indicator-Based Information System (IBIS).
(Threshold N=5)
Contact: Lois Haggard, Utah Department of Health
loishaggard@utah.gov

Vista/PHw - Washington State Center for Health Statistics
(Threshold N = 5)
Contact: David Solet, Washington Center for Health Statistics.
(www.doh.wa.gov/OS/Vista/homepage.htm)

VitalNet
(User-specified threshold)
Contact: Daniel Goldman, System Developer, Expert Health Data Programming
(www.ehdp.com/vitalnet/)

2. "Numerator-based cell suppression variations"

In this variation of the "*numerator-based suppression*" rule, not only are all statistical cells with one to five subjects suppressed, but there is additional suppression of all statistical cells that would allow for the calculation of any other cells with values of 1-4. [6] While the suppression helps in protecting individual identity it also results in loss of information.

Reference:

Cohen, Bruce B., 2001. *Guidelines for the release of aggregate statistical data: Massachusetts perspective on issues and options.* Presentation at the Assessment Initiative/NAPHSIS Conference, September12, 2001.
(Bruce.Cohen@state.ma.us)

Agencies/Systems Using Approach:

Massachusetts Department of Public Health
See Confidentiality Policy and Procedures

3. The "Denominator Rule"

The "*Denominator Rule*" is designed to prevent the display of information when the population under consideration is less than a certain size, such as 100,000 population. The assumption made is that there are a limited number of persons with any given set of characteristics in a small population, therefore by extending the population covered to a larger size, one can protect the identity of individuals. Overtime a number of state agencies have used as a minimum population 30 cases/events. When the denominator is less than 30, the cell is suppressed. No attention is then paid to the actual cell size. This can pose problems for statistically testing—the reliability of the result may be questioned.

Additionally, confusion may exist about what DENOMINATOR means:

- Is it the number in the POPULATION?; or
- Is it the number of EVENTS (e.g. deaths of any cause)?; or
- Is it the number if deaths of any cause in a certain age group or geographic area?

One disadvantage of using the denominator rule alone is that there's a potential for a table with adequate numbers in all its cells to be suppressed. One possible solution is to restrict the application of this rule to rare events (which means also small numbers in the numerator) or extremely skewed distributions.

References:

Cohen, BB, 2001. *Guidelines for the release of aggregate statistical data: Massachusetts perspective on issues and options.* Presentation at the Assessment Initiative/NAPHSIS Conference, September12, 2001.

Agencies/Systems Using Approach:

Indicator-Based Information System (IBIS).
(health.utah.gov/ibis-ph)
Utah surveys using this method:
YRBS: National, unweighted denominator < 50 cases then:
--include 95% confidence interval when reporting percentage/means
--suppress estimate and footnote (estimate based on <50 cases and is unstable)
--group variable to increase number (e.g. combine grades)
State/Local: Estimate suppressed when denominator < 100 cases
Contact: Lois Haggard, Utah Department of Health
loishaggard@utah.gov

4. "Numerator and Event Denominator Rule"

Only the margins of a table are displayed if any table cell subtracted from the number of total events in the same data file for the same characteristics yields a small number (e.g. less than 10).

For example, a cell with one Black female aged 25-34 AIDS death would be published if there were 15 Black female aged 25-34 total deaths. The assumption is that it may be possible to identify the diagnosis of a person if there are fewer than 10 people with the same demographics characteristics and who had the same event (death, in this case, or perhaps birth or hospitalization).

In addition, if less than two row or column totals are greater than five then all the row or column totals are suppressed. This additional rule prohibits determining a suppressed table if the margin totals are small.

This rule protects against release of data when there is a small difference between the number of events in a cell of a table and the total events related to the cell. The rule however, does allow for some small numbers to be displayed when there is a large difference between the events displayed and the total events related to the cell.

This rule requires an interactive determination of the total cell counts of the file to compare with the proposed table. This is a major disadvantage if one is attempting to build an open query systems. Algorithms might be designed to address this, but the cost of doing so would be high.

Reference:

Land, G. Dec. 2001. *Confidential data release rules.* Presentation to WDDS Users Network.

Agencies/Systems Using Approach:

Missouri Information for Community Assessment (MICA).
Missouri Department of Health and Senior Services
Contact person: Garland Land. Director
Center for Health Information Management and Evaluation
Missouri Department of Health and Senior Services
P.O. Box 570
Jefferson City , Mo. 65102
573-751-6272
landg@dhss.mo.us
(www.dhss.mo.gov/MICA/nojava.html)

Criteria Used in Utah:
≥ 100 persons are in population of interest
≥ 20 cases in the numerator are in population of interest
Utah Department of Health
Contact: Lois Haggard, Utah Department of Health
loishaggard@utah.gov

5. Numerator/Denominator-Based Suppression

Cell sizes based on a combination of denominator[3] (population from which the health events arise) and numerator[4] (health event) are suppressed in accordance with the table shown below.[5] Aggregate data with denominator and numerator values greater than those indicated in the table may be considered sufficiently de-identified so as not to constitute confidential information, and may be disclosed. This method has a similar weakness in regard to the need to apply the method and then determine whether there is any additional privacy risk, this precludes interactive non-restricted query systems. These systems, however, could likely be designed with restricted display of output, and other technical measures (micro-aggregation) to control for release of additional information about the individual.

DENOMINATOR (D)	NUMERATOR (N)	STANDARD
10-29	1-4	Suppress numerator and any other cells[6] that would allow for the calculation of any other cells with values of 1-4
10-29	5-29	Suppress any cells that would allow for the calculation of any other cells[6] with values of 1-4
0-8	0-9	Suppress numerator
=N	=D	Suppress numerator unless privacy risk is minimal

<u>Reference:</u>

Cohen, Bruce B., 2001. *Guidelines for the release of aggregate statistical data: Massachusetts perspective on issues and options.* Presentation at the Assessment Initiative/NAPHSIS Conference, September12, 2001.
(Bruce.Cohen@ma.state.us)

<u>Agencies/Systems Using Approach:</u>

Massachusetts Department of Public Health
See Confidentiality Policy and Procedures

6. <u>Alternative Suppression Standards</u>

In the Missouri Department of Health and Senior Services, any Center may develop an alternative aggregate data release standard if it decides not to follow any of the standards above, provided that the standard is at least as restrictive as the above stated standards (see "*Numerator and Event Denominator Rule*" Missouri); and any alternative standard is documented by the Center and approved by the Privacy Officer prior to implementation. This approach provides flexibility for special circumstances and merged databases.

In many public health agencies, suppression standards are based on the specific database, mandates via funding organizations, history of the data release, preferences of specific data stewards. This can cause problems when databases are merged to answer specific questions, for example, if a cancer registry has a "denominator" rule of 1000 and a hospital discharge system has a "numerator" suppression rule of <5 in a cell, both databases could "charge inappropriate release of information" when a merged file is created for web-based data dissemination. The solution is as stated above, prior approval by a Privacy Officer who can mediate the two alternative rules.

References

Land, G. Dec. 2001. *Confidential data release rules.* Presentation to WDDS Users Network. landg@dhss.mo.us

Agencies/Systems Using Approach:

Missouri Department of Health and Senior Services
Contact: Garland Land Director
Center for Health Information Management and Evaluation
Missouri Department of Health and Senior Services
P.O. Box 570
Jefferson City , Mo. 65102
573-751-6272
landg@dhss.mo.us

F. Summary

This guideline addresses the various options for reducing disclosure risk for public health data in web-based data dissemination systems. The combination of methods is up to the user given the context of their environment, data system constituents and mandates, type of user web access, and assessment of risk of disclosure. In the attached Appendix there is a decision-tool for assessing risk of disclosure.

Guideline Use
We hope users of this document will notify the National Association of Health Data Organizations with additions or corrections. Please send an email to: Barbara Rudolph, Ph.D., Senior Scientist for Research and Data. brudolph@nahdo.org.