Guidelines for Reporting Small Numbers to Protect Confidentiality

Version 2
March 3, 2015

Contact:

Juanita Heimann
Director's Office
Oregon State Public Health Division
971-673-1267
Juanita.A.Heimann@state.or.us

Note: All data presented in hypothetical examples in this report are fictional

Table of Contents

Executive Summary

Background and Scope

The Oregon Public Health Division (OPHD) has a responsibility to report data in such a way that protects private information about individuals from disclosure. Even when reporting aggregated data that contains no personal identifiers there is some risk of inadvertantly breaching confidentiality if the population on which the data are based is small. Although the specific concerns vary by data set, it is useful to have some basic division-wide guidelines for reporting data based on small numbers.

This report has a limited focus on issues related to the distribution of data collected and/or maintained by OPHD. Specifically, the report relates to:

- Aggregated data
- Full count and client data
- Data of a potentially confidential nature
- Concerns related to inadvertent breach of confidentiality through release of data based on small numbers

Key Concepts

Confidentiality refers to the concept that private information obtained from an individual will be kept secure and not disclosed without the authorization of that individual. Aggregated – or, summarized – data are generally presented as a frequency or rate of some health event within an underlying population defined by shared characteristics such as age and gender. Because aggregated data contain no explicit identifiers such as name or address, the circumstances in which the release of such data results in disclosure of confidential information are exceedingly rare.

But there is a risk of creating the *perception* that someone can be identified through linkage of statistical data released by OPHD with other sources of information about the underlying population. An information source might be a publicly available database with names and addresses such as Voter Registration records. Or, it may just be personal knowledge about neighbors in a small community. This kind of information linkage cannot be completely prevented but we can minimize the possibility of it happening.

Information linkage is most likely to occur when the combination of attributes that describe confidential data subjects in aggregated data (such as age, sex, race or county of residence) is so specific that there are only a few people in the underlying population who match that description. The closer the number of possible matches comes to one, the more certain the perception that a definitive identification can be made.

An exact measurement of disclosure risk is not possible because it depends partially on a 3rd party's perception that an exact match can be made using any number of external sources of information. But, the critical factor in determining the risk of disclosure from release of any particular data point is the size of the underlying population from which that data point was derived. From a data analysis perspective, this is generally thought of as the "denominator". The only circumstance in which individual data subjects can *definitively* be identified from the release of an aggregated data point is when they are 100% of an identifiable underlying population. In all other cases, some threshold of acceptable disclosure risk must be determined.

Methods for Protecting Confidentiality

There are a number of methods for controlling the disclosure of private information in aggregated data. Only two – further aggregation and cell suppression – are discussed in this report because they are the least compromising to the integrity of the data.

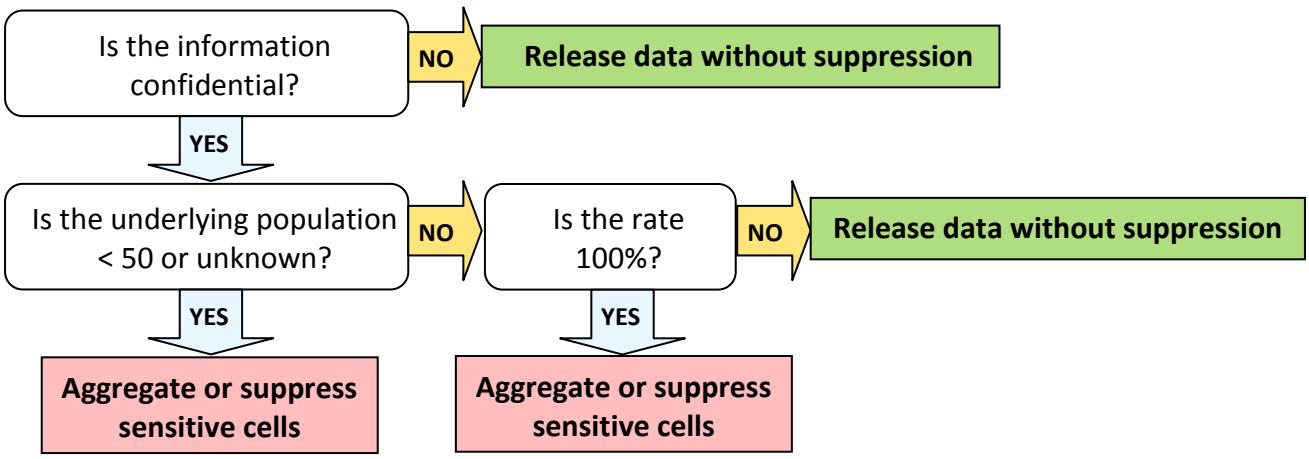
Geographic areas, time periods, race and age groups or other stratified groups may be rolled up into more general categories to create larger subgroups that increase the size of the population. When further aggregation does not result in sufficiently large numbers, the value of data table, chart or map cells that are deemed to be sensitive are not released.

When the only recipient of the data is a local or tribal health department within Oregon and they are requesting data for their own jurisdiction, discretion may be used to report small numbers data without applying any disclosure control methods such as suppression.

General Recommendations to OPHD Programs

1: Use these guidelines as a basic minimum standard
2: Have a written policy specific to data collected by each program
3: Use denominators as the primary threshold values to identify sensitive cells
4: Events with 100% rates should not be reported
5: Aggregate first if possible then suppress sensitive cells
6: Apply complementary suppression to tables with marginal totals
7: Apply rules for reporting data to protect reliability after confidentiality concerns are addressed

Data Release Decision Tree



Introduction

Background

The Oregon State Public Health Division (OPHD) has a responsibility to provide timely data on trends and patterns in the health of the state's population to our partners and the public. OPHD also has a legal and ethical responsibility to report data in such a way that protects private information about individuals from disclosure. Although these two objectives can be at odds with one another when data based on a small population is concerned, it is important to strike a balance that will protect individuals, preserve public trust, and permit the release of useful statistical information.

This dilemma impacts every OPHD program that manages data sets or provides data analysis and reports. Although the specific concerns vary by data set, it is useful to have some basic division-wide guidelines for reporting data based on small numbers. These guidelines will not only assist those who analyze and report data. They will provide transparency for data users.

Scope of Report

This report has a limited focus on issues related to the distribution of data collected and/or maintained by OPHD. Specifically, the report relates to:

- Aggregated data
- Full count and client data (with one exception)
- Data of a potentially confidential nature
- Concerns related to inadvertent breach of confidentiality through release of data based on small numbers

Furthermore, it is assumed that data are being reported in a manner accessible to the general public. When the only recipient of the data is a local or tribal health department within Oregon and they are requesting data for their own jurisdiction, discretion may be used to report small numbers data without applying any disclosure control methods such as suppression.

How to Use this Report

The recommended rules in this report are intended as guidelines rather than official OPHD policy. Each OPHD program which acts as a data steward is encouraged to develop its own written policy regarding data security and small number reporting using the recommendations in this report as minimum standards. Programs may determine that they require more stringent rules or specific exceptions to these rules because other public health considerations prevail (e.g., sentinel case reporting). Within data systems, certain elements may be considered confidential information while others are not. The discussion of underlying concepts in this report should be of assistance in making those decisions.

Key Concepts

Aggregated vs. Record-level Data

The greatest concerns about confidentiality relate to the collection, storage and sharing of record-level data containing Protected Health Information (PHI)[1]. The Safe Harbor method of data de-identification described by the Health Insurance Portability and Accountability Act (HIPAA) involves redacting 18 specific types of fields from a dataset. But this method of protecting record-level data prior to release is not well suited to aggregated data.

Aggregated – or, summarized – data lacks explicit identifiers such as names, phone numbers, and home addresses. And it is presented as a frequency of cases (e.g., number of people with Diabetes) or rate (e.g., Chlamydia incidence rate) within an underlying population rather than as a list of individuals and their characteristics. But, PHI that has been aggregated is still vulnerable to threats to confidentiality.

If applied to aggregated data, the Safe Harbor method would not allow OPHD to release data by county or any other geographic area smaller than the state. Furthermore, the Safe Harbor method is not based on a determination of the level of risk to confidentiality in a specific dataset. These Guidelines describe an alternative to the Safe Harbor method.

Full Count/Client Data vs. Survey Data

OPHD collects and manages many different kinds of data that represent individuals and their health experiences.

- A full count dataset theoretically contains every instance of a qualified event (e.g., death, birth, cancer or communicable disease diagnosis, immunization record, health care encounter).

- A client dataset contains health or encounter records of all the individuals who participate in or receive services from a particular health or social service agency (e.g., WIC).

- A survey dataset contains information gathered from asking questions of randomly selected respondents. For complex surveys such as the Behavioral Risk Factors Surveillance System (BRFSS), these respondents are weighted so that estimates can be made to infer the behavior of an entire population (such as all adult Oregonians).

Only the first two – full count and client data – raise concerns about breaching confidentiality. Individuals in population-based surveys represent the health status, risk factors, and opinions of a large number of persons with similar demographic characteristics. Therefore, their information is weighted to the number of people they represent. Because of this, releasing information on small numbers of individuals does not constitute a breach of confidentiality as long as there's no way to identify who was and wasn't surveyed.

[1] Health information, including demographic information collected from an individual, that (1) is created or received by a health care provider, health plan, employer, or health care clearinghouse; and (2) relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual; and (a) that identifies the individual; or (b) with respect to which there is a reasonable basis to believe the information can be used to identify the individual; and (3) is transmitted by electronic media, maintained in electronic media, or transmitted or maintained in any other form or medium.

One possible exception may occur with surveys such as the Oregon Healthy Teens (OHT) Survey which uses a sampling methodology that may make it possible to identify all survey participants. Although schools within a school district are sampled, often the students in those schools are not. Theoretically, all students in the selected schools participate. When OHT data are reported for geographic areas in which the participating schools are or can be identified, the data should be treated as if it were full count data for the purposes of determining risk to confidentiality.

What is Confidentiality?

Confidentiality refers to the concept that private information obtained from an individual will be kept secure and not disclosed without the authorization of that individual. Death and birth events are not in themselves confidential. But, all of the information pertaining to births or deaths is confidential. Confidential information includes cause of death, disease status, medical diagnoses, clinical procedures, health behaviors, etc.

Another concern related to the release of data based on small numbers is the statistical reliability of that data. Although the methods for addressing reliability and confidentiality concerns are related – and may overlap – the rationale for doing so is different. This report will not address the reliability issue.

What Constitutes a Breach of Confidentiality?

Even hypothetical scenarios in which aggregated data reported by OPHD could lead to an incontrovertible breach of confidentiality are extremely rare. But, there is a gradient of risk which depends on the degree to which it could be perceived that a statistical figure on a table can be linked to a specific individual in the population. For that linkage to occur, some external source of information about the underlying population is required. Data linkage cannot be completely prevented but the possibility of it happening can be minimized.

Data linkage

When releasing aggregated data, inadvertent disclosure of private information about an individual may result from combining the released data with publicly available information or detailed external data sources that may or may not be available to the general public. The identity of or private information about an individual may be inferred by someone with access to these sources of information.

The linkage may be deliberate. A resourceful investigator might attempt to match statistical data to Voter Registration records or other public data sets that include names and addresses for a large portion of the population. Or, the linkage may simply be the consequence of general knowledge in a close community. For example, you may know that your neighbor is being treated for cancer, or your child's classmate is frequently hospitalized, or your friend gave birth last year. When combined with certain statistical data about your community this could lead you to deduce other information about these individuals.

Disclosure risk

Because of the many unknowns with potential data linkage, the risk associated with it occuring can't be exactly quantified. But, data linkage is most likely to occur when the combination of attributes that describe confidential data subjects in aggregated data (such as age, sex, race or county of residence) is so specific that there are only a few people in the underlying population who match that description.

Incontrovertible identification of a particular data subject can, in general, only occur when data linkage reveals that there is only one possible person in the population who shares their attributes. Beyond this unlikely occurance there is only a gradient of disclosure risk that is based, at least in part, on the *perception* that an indivdual can be uniquely identified. So, the critical factor in determining the risk of disclosure from release of any particular data point is the size of the underlying population from which that data point was derived. From a data analysis perspective, this is generally thought of as the "denominator" and can be used as a proxy measure of disclosure risk. The smaller the denominator, the greater the risk.

Hypothetical example #1:

Diabetes-related Hospitalizations by Age Group, Sex and County				
County	***Sex***	***Age Group***	***# People Hospitalized***	***# of People in Population***
Fairfax	*Female*	*55 to 64*	*10*	*51,000*
Fairfax	*Male*	*55 to 64*	*30*	*49,000*
Fairfax	*Female*	*65 to 74*	*10*	*30,000*
Fairfax	*Male*	*65 to 74*	*20*	*28,000*

Every combination of the attributes county, sex and age group describes thousands of people who may or may not be the few who were hospitalized for diabetes. Contrast with the same data aggregated differently, below.

Diabetes-related Hospitalizations by Age, Sex and ZIP code				
ZIP code	***Sex***	***Age in Years***	***# People Hospitalized***	***# of People in Population***
54321	*Female*	*60*	*1*	*12*
54321	*Male*	*60*	*3*	*9*
54321	*Female*	*70*	*1*	*1*
54321	*Male*	*70*	*2*	*3*

At least one data point in the above table uniquely describes an individual in the population. Other data points come close to doing so. There are a number of ways in which the one 70 year old female in ZIP code 54321 can be identified. Department of Motor Vehicle records could be used to determine her name – assuming she has a driver's license. Or, her neighbor can probably identify her based on this information. She can be positively identified and confidential information about her – that she was hospitalized for diabetes – has been revealed.

Small underlying population

So, the primary risk to confidentiality comes from reporting data on events that occur within small, well-defined populations.

Hypothetical example #2:

Chlamydia Rates Among Women Age 20-24 by Community			
Place	***Rate per 100,000***	***# of Women Age 20-24 with Chlamydia***	***# of Women Age 20-24***
Community A	*30,000 per 100,000*	*3*	*10*
Community B	*689 per 100,000*	*2*	*290*
Community C	*3,328 per 100,000*	*25*	*751*
Community D	*2,633 per 100,000*	*275*	*10,443*

In Community A it would be easy to identify all 10 women between the ages of 20 and 24. Someone who knows everyone in this small community well might be able to use this knowledge to infer who the 3 with Chlamydia might be. Whether the inference is accurate or not may not be relevent to the perception that these 3 people can be identified.

Caution should be used when determining what constitutes the underlying population. It may depend on how the data are presented.

Hypothetical example #3:

Leading Causes of Death among Native American Men Age 18-24 in Community X			
Cause of Death	***Death Rate per 100,000 among Native American Men Age 18-24***	***Number of Deaths among Native American Men Age 18-24***	***Number of Native American Men Age 18-24***
Total	*140 per 100,000*	*7*	*5,000*
Cancer	*20 per 100,000*	*1*	*5,000*
Suicide	*100 per 100,000*	*5*	*5,000*
Heart Disease	*20 per 100,000*	*1*	*5,000*

At first glance, the underlying population of Native American men age 18 to 24 in Community X is not small (5,000). But, because of the way the table is constructed there is another – much smaller – identifiable underlying population. That is, all persons in this demographic group who died (7).

Information about the deaths of these individuals may have been reported in the newspaper – as a matter of public record - even though their cause of death was not. There is a risk that public information and community knowledge about these 7 individuals could be used to determine their cause of death.

100% of underlying population

A special case in which the underlying population does not have to be small in order for definitive disclosure of confidential information to occur is when 100% of that population are also data subjects. If it were reported that 100% of a population had disease x or engaged in behavior y it would constitute an obvious breach of confidentiality. Fortunately, these circumstances are rare. But, they are more likely to occur in small populations.

Hypothetical example #4:

1st Trimester Prenatal Care by Community			
Place	***% of Women Who Received 1st Trimester Prenatal Care***	***# of Women Who Received 1st Trimester Prenatal Care***	***# of Women Who Gave Birth***
Community A	*75%*	*150*	*200*
Community B	*80%*	*40*	*50*
Community C	*0%*	*0*	*51*
Community D	*50%*	*40*	*80*

Admittedly, the hypothetical figures for Community C (above) are unlikely. But, this illustrates that 0% represents the same disclosure risk as 100% in circumstances where the inverse characteristic represents sensitive information (i.e., not receiving early prenatal care). Reporting that 0% of the population in Community C contracted Chlamydia would not be a problem, though.

A gray area exists with reporting high percentages that are still less than 100%. It doesn't constitute definitive disclosure but it can stigmatize an entire population. In this case, it may be a better assessment of risk to look at the difference between the denominator and the numerator.

Hypothetical example #5:

Smoking During Pregnancy			
Place	***% of Women Who Smoked During Pregnancy***	***# of Women Who Smoked During Pregnancy***	***# of Women Who Gave Birth***
Community C	*98%*	*50*	*51*

While there's always room for doubt whether an individual is part of the 98% in Community C that smoked during pregnancy or the 2% that didn't, everyone in the population is stigmatized by the perception that nearly all smoked. On the other hand, it would be important to know that there's such a high smoking rate in this community in order to develop interventions.

While it may be desirable to avoid stigmatizing an entire population, it should be kept in mind that a very high rate may be a significant finding for surveillance and for identifying communities at risk.

Small number of events

Reporting data for a small number of events occuring within a large population does not represent any risk to confidentiality.

Hypothetical example #6:

Tuberculosis Rates by County			
County	**Rate per 100,000**	**# of Tuberculosis cases**	**# of People in County**
County A	*1.0*	*1*	*100,255*
County B	*1.7*	*2*	*118,360*
County C	*0.3*	*1*	*354,542*
County D	*0.8*	*3*	*383,857*

There's no way to identify the few individuals in each of the counties who had Tuberculosis because they could be any one of thousands of people.

High risk attributes for data linkage

Some attributes – such as demographic characteristics - are easier to link with external information because they are observable, self-evident, constant or more likely to be found in sources of public information. For example, gender is observable and is found in most public data. Birthdate is commonly recorded in public datasets and is a feature of every person that will never change. While ZIP code of residence is readily available information, it isn't a stable feature of an individual as people move. In contrast, clinical features like blood type or a blood glucose level from a specific clinical encounter are not observable traits and are not likely to be found in public datasets.

Hypothetical example #7:

Smoking During Pregnancy			
Place	**% of Women Who Smoked During Pregnancy**	**# of Women Who Smoked During Pregnancy**	**# of Women Who Gave Birth**
Community X	*60%*	*3*	*5*

In this case, it's easy to identify all the women in a small community who give birth. The underlying population can be identified by observable traits. If your friend is one of the very few women in the population who gave birth, you may be able to deduce whether she smoked during pregnancy based on knowledge you have of this entire population in your community – such as their smoking habits prior to pregnancy.

The potential breach of confidentiality is a result of the very small number of people in the underlying population (5) and the ease with which they can all be identified.

Hypothetical example #8:

- *4 out of 10 colorectal cancer diagnoses are late stage cancer*
- *4 out of 10 persons with HIV were exposed by intravenous drug use*
- *4 out of 10 people with type A blood are HIV positive*

In these statements, the underlying population is defined by shared clinical attributes – having colorectal cancer, HIV or type A blood. Because none of these are publicly available information the 10 people in the underlying population are not easily identifiable.

It isn't always clear what constitutes an easily identifiable attribute. But, it's prudent to assume that all demographic characteristics (e.g., age, sex, race) are identifiable traits whereas characteristics that are themselves confidential (e.g., HIV status, cholesterol level) are not.

Threshold of acceptable risk

In all of the above examples, it is assumed that there is some non-quantifiable but non-zero disclosure risk represented by the size of the underlying population as a proxy measure. But, at what population threshold is it no longer possible to know – or find out – enough information about all members of that population to make reasonable deductions about which individuals have experienced a particular health event? Some threshold of acceptable risk must be established. In the absence of a scientfic consensus about such a threshold, these Guidelines recommend continuing to use a denominator of 50. This threshold has been used in OPHD for many years.

Summary of Methods for Protecting Confidentiality

There are a number of methods for controlling the disclosure of private information in aggregated data. Generally, they require specific criteria for determining what constitutes a "sensitive cell" because of it's disclosure risk. Sensitive cells are identified by a threshold which may be the value of the count, population denominator, event denominator, or some combination. A review of the literature relating to confidentiality protection of statistical data reveals little scientific basis to support any particular value for a threshold of acceptable risk.

For clarity, the following terms and definitions will be used throughout the report:

sensitive cell – The value(s) in the data table, chart or map that – if reported – may pose a significant risk of inadvertent disclosure of confidential information.

threshold cell – The cell that contains the threshold value that is used to identify sensitive cells. It may or may not be the same as the sensitive cell. It may or may not be on the same published table, chart or map as the sensitive cell.

Cell Suppression

Cell suppression is a common method in which the values of table cells are not released if they are deemed to be sensitive. If a cell count or denominator is below some threshold, then the sensitive cell value is withheld from published tables. However, there is no scientific guidance or consensus on how to choose an appropriate threshold value. This is generally determined at the discretion of the individual agency or program. If the table being released has two or more dimensions and includes marginal totals, additional cells must also be suppressed to prevent the sensitive cell value from being calculated from the table margins. This is known as complementary suppression. The choice of complementary cells can be somewhat complex, especially in large

tables where there are multiple primary suppressions. The cell suppression method can result in substantial information loss and reduce the usefulness of released statistical data. In addition, it is sometimes possible to combine multiple tables to back-calculate the values of suppressed cells, in which case confidential information may be disclosed.

Data Aggregation

There are several methods to eliminate small numerators and denominators by aggregating groups. Geographic areas, race and age groups or other stratified groups may be rolled up into more general categories to create larger subgroups that increase the size of the cell count and denominator population. For example, instead of reporting pregnancy rates for 10-14 year olds a rate could be calculated for 10-19 year olds. Data may also be aggregated over time, usually by combining years to increase the person-time of the denominator populations. This method allows more data to be released without risk of disclosure, but also limits the detail and utility of the information.

Rounding and Data Perturbation

Other methods described in the literature – such as rounding and data perturbation – have the disadvantage of introducing some level of uncertainty to the data. They are less commonly used than aggregation or cell suppression and will not be discussed further in this report. See references #3 ("Management and Institutional Controls…") and #5 ("Small Numbers, Disclosure Risk,…") in Appendix B: Bibliography for more information on these methods.

Recommended Guidelines for OPHD

General Recommendations

1: Use these guidelines as a basic minimum standard

Programs that collect or manage data have the option of imposing a more stringent standard on data release especially if they want to protect a vulnerable group or particularly sensitive information. Likewise, exceptions to the rule may also be made if there is a compelling public health interest (e.g., sentinel case reporting).

2: Have a written policy

All programs that collect or manage data should have a written policy that defines their confidentiality rule and explains the rationale, especially if it differs from the recommendations of this report.

3: Use denominators as the primary threshold values to identify sensitive cells

The size of the underlying population (or, denominator) is the major factor determining the risk of disclosure of information about an individual in a statistical table, map or chart. The smaller the

subgroup, the more easily inferences can be made about a given individual in that subgroup. Be careful about determining what the true underlying population is.

Because there is no known accepted standard for determining the threshold of cell sensitivity, the "rule of 50" which has been in use by OPHD's Center for Health Statistics for years was chosen.

Note that although the denominator is used as the threshold value, it is actually the numerator itself which is the sensitive cell. When rates are not being calculated the *de facto* denominator is still the threshold value.

4: Events with 100% rates should not be reported

Although rates of 100% are rare, they represent a high risk of breaching confidentiality. Keep in mind that 0% may or may not represent the same risk depending on whether the inverse is a confidential characteristic or not.

Programs may also choose to suppress data with very high rates that are less than 100%. Either select a rate threshold (e.g., 95% or 98%) or use the absolute difference between the denominator and the numerator as a threshold.

5: Aggregate first if possible then suppress sensitive cells

All of the programs in the Public Health Division either use data aggregation, cell suppression, or some combination to protect confidentiality in publicly released data. Other methods such as data perturbation and controlled rounding are generally not used to mask data because of the added complexity and the perception that they may violate the integrity of the data.

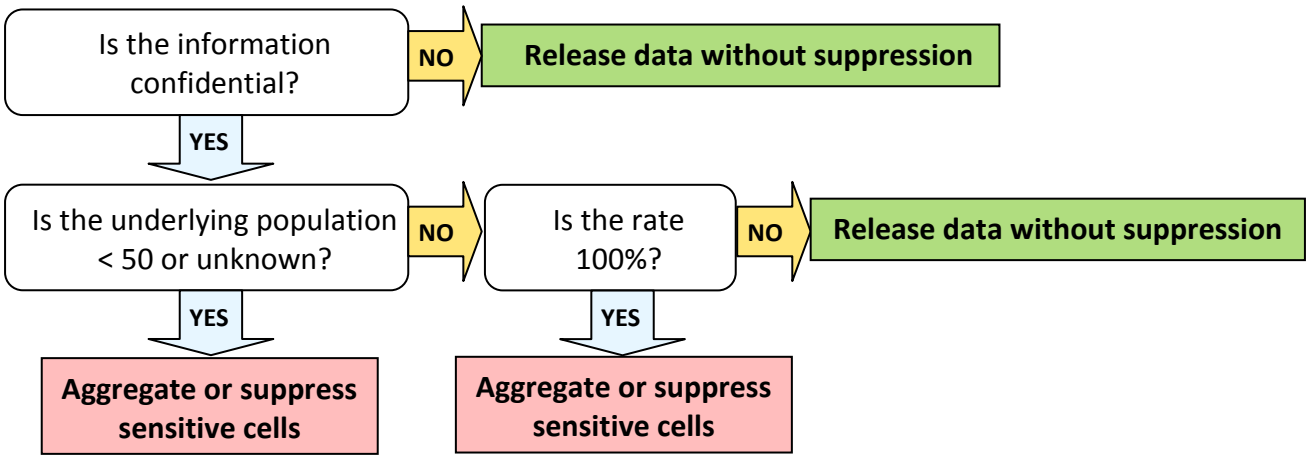
6: Apply complementary suppression to tables with marginal totals

This may not be feasible when a large number of tables are produced for a report or when using a web-based data query system. And, it may not be necessary for all audiences. Complementary suppression prevents data users from being able to back-calculate suppressed values within a table. But, it may not prevent them from being able to perform those calculations across different tables.

7: Apply rules for reporting data to protect reliability after confidentiality concerns are addressed

Concerns about the reliability of the data (not covered in this paper) should also be addressed. But, confidentiality trumps reliability.

Data Release Decision Tree



How to Apply the Guidelines

How to Determine if the Information is Confidential

Although this will vary by dataset, in general almost all of the data that is collected and maintained by OPHD is confidential. A notable exception is simple counts of vital events (number of births and number of deaths) with no other information. All of the other information associated with births and deaths is confidential. Basic demographic information (age, race, sex) is not confidential. See Appendix A for a partial list of confidential and non-confidential data elements.

Ex. Death – NOT confidential
Ex. Death from cancer – confidential
Ex. Cancer diagnosis - confidential

How to Determine the Underlying Population

The underlying population is the group with shared characteristics from which the data subjects are drawn. For convenience, we use the rate denominator or the de facto denominator (i.e., the number we would use if we were calculating a rate). This is not always the same thing as the underlying population.

1. The denominator is frequently an estimate - not an actual count - of persons in that population.
2. The denominator may exclude people who are publicly indistinguishable from people who ARE in the denominator. For example, when calculating "% Low Birth weight" the denominator excludes births for which the birth weight is unknown.

For client data, the underlying population is just the client population – not the total population.

When there is no way to know if the underlying population is less than 50, it should be treated as if it were. For example, if the population from which data subjects are drawn is everyone living in a large unincorporated

area for which there is no current data you may make the assumption based on the size of the area that the de facto denominator is ≥ 50. However, if – for example - the underlying population is all transgendered persons living in Hillsboro it should be treated as unknown.

How to Suppress Sensitive Cells

There is a distinction between the threshold criteria used to identify sensitive cells and what should actually be suppressed. Although the size of the denominator is the primary determinant of sensitivity, it is in itself not sensitive in most cases. Population estimates, total enrolled clients, and birth and death totals do not need to be suppressed.

Tables

Both the count of cases and the crude rate can be sensitive cells when they appear in tables together. Age-adjusted rates or other rates that can't easily be used to back-calculate the count don't need to be suppressed. The table should be footnoted to explain why the data was suppressed. In the example below, the **threshold cell** is shown in a **blue diamond** and **sensitive cells** are **circled in red**.

Prior to Suppression

Subset	Rate per 1,000	95% CI	Count	Pop
Age 0-17	65	(40 to 100)	20	306
Age 18-34	128	(13 to 240)	6	47
Age 35-64	58	(21 to 126)	6	103
Age 65+	19	(15 to 24)	77	3,992

Partial Suppression

Subset	Rate per 1,000	95% CI	Count	Pop
Age 0-17	65	(40 to 100)	20	306
Age 18-34	*	(13 to 240)	*	47
Age 35-64	58	(21 to 126)	6	103
Age 65+	19	(15 to 24)	77	3,992

* Data not reported to protect confidentiality

Full Suppression

Subset	Rate per 1,000	95% CI	Count	Pop
Age 0-17	65	(40 to 100)	20	306
Age 18-34	*	*	*	*
Age 35-64	58	(21 to 126)	6	103
Age 65+	19	(15 to 24)	77	3,992

* Data not reported to protect confidentiality

Rates do not need to be suppressed if the count is not on the same table. However, sensitive counts should still be suppressed even when the threshold cell is not on the same table.

Rate Not Suppressed		
Subset	Rate per 1,000	95% CI
Age 0-17	65	(40 to 100)
Age 18-34	128	(13 to 240)
Age 35-64	58	(21 to 126)
Age 65+	19	(15 to 24)

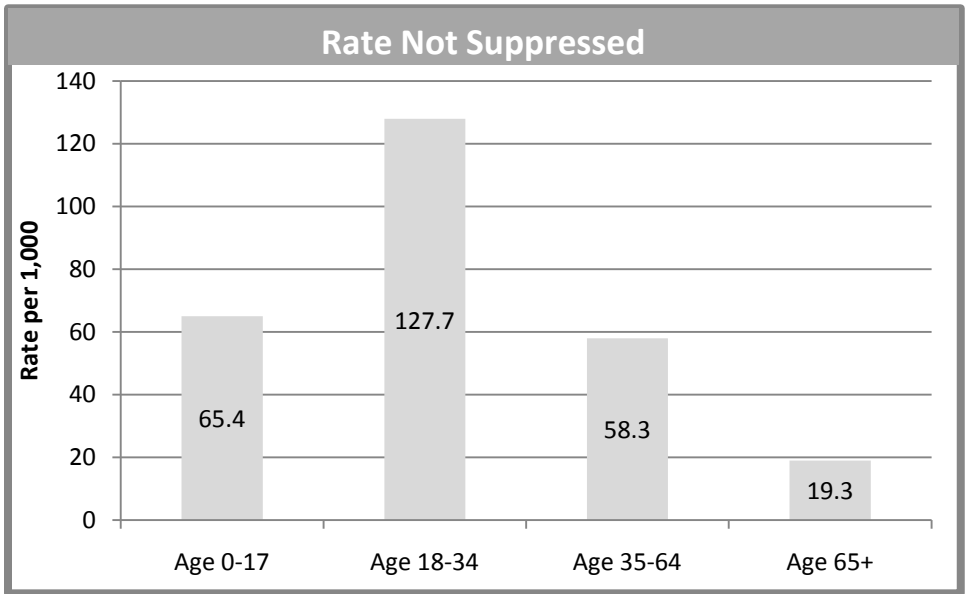
Suppression of Count	
Subset	Count
Age 0-17	20
Age 18-34	*
Age 35-64	6
Age 65+	77

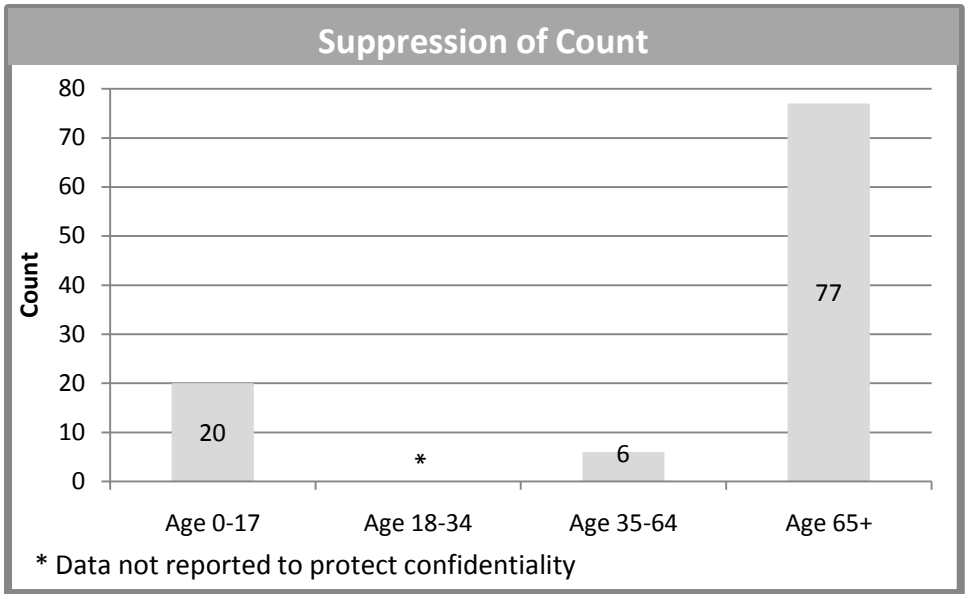
* Data not reported to protect confidentiality

Authors of reports containing multiple data tables may want to consider whether a reader would be able to back-calculate suppressed data by cross-referencing multiple tables.

Charts

When the data are charted instead of being presented in tabular form, suppression only needs to occur if counts are shown.





Maps

When data are displayed on maps, sensitive cells only need to be suppressed for confidentiality reasons when counts are mapped, the legend ranges are narrow enough to identify the exact value of the count in any geographic area, and the size of the underlying population is known.

But, use caution when presenting data in interactive GIS software. The user should not be able to zoom in on address level data.

Complementary Suppression to Prevent Back-Calculating Suppressed Values

It is quite easy for data users to back-calculate the value of a single suppressed cell on a table if the total across sub-categories is also given.

Primary Suppression only				
Subset	**Rate per 1,000**	**95% CI**	**Count**	**Pop**
Age 0-17	65	(40 to 100)	20	306
Age 18-34	*	*	*	47
Age 35-64	58	(21 to 126)	6	103
Age 65+	19	(15 to 24)	77	3,992
Unknown			4	
All Ages	**25**	**(21 to 27)**	**113**	**4,448**

* Data not reported to protect confidentiality

This back-calculation can be prevented by also suppressing the value of at least one other cell even though it is not in itself sensitive. Because this results in the loss of otherwise releasable information, care should be taken to select cells for complementary suppression that provide the least useful information. For example, the table may contain a category for “Other” or “Unknown” which is not particularly meaningful.

Complementary Suppression				
Subset	**Rate**	**CI**	**Count**	**Pop**
Age 0-17	65	(40 to 100)	20	306
Age 18-34	*	*	*	47
Age 35-64	58	(21 to 126)	6	103
Age 65+	19	(15 to 24)	77	3,992
Unknown			*	
All Ages	**25**	**(21 to 27)**	**113**	**4,448**

* Data not reported to protect confidentiality

Analysts should be aware that tables which contain data for rolling average years are also vulnerable to back-calculation.

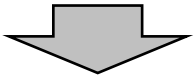
Example Scenarios

In the example scenarios below, the **threshold cell** is shown in a **blue diamond** and **sensitive cells** are **circled in red**.

Scenario 1

Prior to Suppression			
Percent of Births to Women Age 25-29 in which Mother Smoked During Pregnancy, by Race, Oregon County			
Race	**% of Women Age 25-29 Who Smoked During Pregnancy**	**# of Women Age 25-29 Who Smoked During Pregnancy**	**# of Women Age 25-29 Who Gave Birth**
African American	5%	3	62
American Indian	33%	3	9
Asian/Pacific Islander	2%	1	53
White	8%	90	1068
Unknown	16%	8	50
Total	**8%**	**105**	**1242**

- Which traits are confidential information? *Smoked during pregnancy*
- Which traits could be linked with other data sources? *Residence in Oregon County; female; age 25-29; race/ethnicity; had a baby recently*
- Risk of disclosure of confidential information? *All women who had a baby in this place/age/race demographic group are identifiable. Community knowledge about the underlying population may be used to infer which ones did or did not smoke during pregnancy*.
- Decision: *Suppress sensitive cell based on small denominator and complementary cell to prevent back-calculation. Suppress rates in order to prevent back-calculation of counts.*

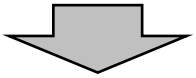


Complementary Suppression			
Percent of Births to Women Age 25-29 in which Mother Smoked During Pregnancy, by Race, Oregon County			
Race	**% of Women Age 25-29 Who Smoked During Pregnancy**	**# of Women Age 25-29 Who Smoked During Pregnancy**	**# of Women Age 25-29 Who Gave Birth**
African American	5%	3	62
American Indian	*	*	9
Asian/Pacific Islander	2%	1	53
White	8%	90	1068
Unknown	*	*	50
Total	**8%**	**105**	**1242**
* Data not reported to protect confidentiality			

Scenario 2

Prior to Suppression		
Percent of New HIV/AIDS Diagnoses, by Transmission Category, Oregon County		
Transmission Category	**Percent of New HIV/AIDS Diagnoses**	**# of New HIV/AIDS Diagnoses**
MSM	60%	12
Intravenous Drug Use	5%	1
Transfusion	10%	2
Perinatal	5%	1
Other	20%	4
Total	**100%**	**20**

- Which traits are confidential information? *HIV/AIDS diagnosis; Transmission category*
- Which traits could be linked with other data sources? *Residence in Oregon County; gender (relevant to MSM transmission only); age (relevant to Perinatal transmission only)*
- Risk of disclosure of confidential information? *Although it isn't explicitly stated on the table, the entire population from which these cases are drawn is about 1,000,000. The only possible risk of disclosure comes from the unlikely possibility that someone knows who all of the 20 people (out of 1,000,000) who were diagnosed with HIV are and can deduce information about how they were exposed based on this community knowledge.*
- Decision: *There is no need to suppress data because the risk is exceedingly small.*

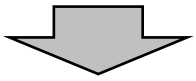


No Suppression		
Percent of New HIV/AIDS Diagnoses, by Transmission Category, Oregon County		
Transmission Category	**Percent of New HIV/AIDS Diagnoses**	**# of New HIV/AIDS Diagnoses**
MSM	60%	12
Intravenous Drug Use	5%	1
Transfusion	10%	2
Perinatal	5%	1
Other	20%	4
Total	**100%**	**20**

Scenario 3

Prior to Suppression			
Syphilis Cases by Type, Oregon			
Type	**Crude Rate per 100,000**	**# of Syphilis Cases**	**Total Population**
Early	1.6	60	3,844,195
Late Latent and Other/Unknown	0	0	3,844,195
Congenital	0	0	3,844,195
Total	**1.6**	**60**	**3,844,195**

- Which traits are confidential information? *Syphilis; type of syphilis*
- Which traits could be linked with other data sources? *Residence in Oregon*
- Risk of disclosure of confidential information? *The total underlying population is very large – almost 4 million people. But, there is another underlying population: those 60 with syphilis. If it were possible to identify all of them then the type of syphilis they had would also be disclosed because 100% of the cases were early syphilis. But, the chances of that happening are remote. In order to uncover one piece of confidential information someone would need to already know another piece of equaly confidential information.*
- Decision: *There is no need to suppress data because the risk is exceedingly small.*

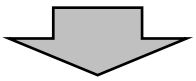


No Suppression			
Syphilis Cases by Type, Oregon			
Type	**Crude Rate per 100,000**	**# of Syphilis Cases**	**Total Population**
Early	1.6	60	3,844,195
Late Latent and Other/Unknown	0	0	3,844,195
Congenital	0	0	3,844,195
Total	**1.6**	**60**	**3,844,195**

Scenario 4

Prior to Suppression		
Deaths among Native American Men Age 18-24, by Cause, Oregon County		
	Percent of Deaths	**# of Deaths**
Cancer	0%	0
Heart Disease	0%	0
Suicide	100%	5
Accidents	0%	0
Other	0%	0
Total	**100%**	**5**

- Which traits are confidential information? *Cause of death*
- Which traits could be linked with other data sources? *Residence in Oregon County; male; age 18-24; Native American race/ethnicity*
- Risk of disclosure of confidential information? *Deaths are public record. All 5 individuals can easily be identified. So, their cause of death has definitively been disclosed.*
- Decision: *Supress the 100% cell and use complementary suppression to prevent back-calculation.*

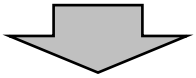


Complementary Suppression		
Deaths among Native American Men Age 18-24, by Cause, Oregon County		
	Percent of Deaths	**# of Deaths**
Cancer	0%	0
Heart Disease	0%	0
Suicide	*	*
Accidents	0%	0
Other	*	*
Total	**100%**	**5**
* Data not reported to protect confidentiality		

Scenario 5

Prior to Suppression					
Percent of 11th Graders Reporting Binge Drinking, By Race/Ethnicity, Oregon School District (source: OHT)					
Gender	**% Reporting Binge Drinking (weighted)**	**# Reporting Binge Drinking (weighted)**	**# Reporting Binge Drinking (unweighted)**	**# Participating Students**	**# Participating Schools**
African American	15%	6	6	40	
American Indian	13%	3	3	24	
Asian/Pacific Islander	29%	15	15	52	
Hispanic	10%	6	6	60	
White	23%	45	45	200	
Total	**20%**	**75**	**75**	**376**	**1 out of 1**

- Which traits are confidential information? *Binge drinking*
- Which traits could be linked with other data sources? *Race/ethnicity; grade; school and school district*
- Risk of disclosure of confidential information? *In most types of surveys, it isn't possible to identify who did and didn't participate in the survey out of the entire population. But, because of the sampling methodology for the OHT survey, participants can be identified if it is known which school(s) in the school district participated because, theoretically, all students in that school are surveyed. If there is only 1 school in the school district, all participants are personally known to everyone associated with that school. Because the number of participants in certain demographic groups is small, personal knowledge of these groups may lead to disclosure of their alcohol use behaviors.*
- Decision: *Treat this data as if it were full count data rather than survey data and suppress sensitive cells based on small denominators. Because more than one cell is being suppressed, further complementary suppression isn't necessary. Suppress percents in order to prevent back-calculation of counts.*



Suppressed					
Percent of 11th Graders Reporting Binge Drinking, By Race/Ethnicity, Oregon School District (source: OHT)					
Gender	**% Reporting Binge Drinking (weighted)**	**# Reporting Binge Drinking (weighted)**	**# Reporting Binge Drinking (unweighted)**	**# Participating Students**	**# Participating Schools**
African American	*	*	*	40	
American Indian	*	*	*	24	
Asian/Pacific Islander	29%	15	15	52	
Hispanic	10%	6	6	60	
White	23%	45	45	200	
Total	**20%**	**75**	**75**	**376**	**1 out of 1**
* Data not reported to protect confidentiality					

If there are other non-participating schools in the school district and the estimates for this school are weighted to the demographic profile of all schools, it might be possible to report the data without suppression as long as there's no way to back-calculate the unweighted values.

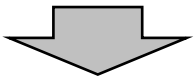
Prior to Suppression					
Percent of 11th Graders Reporting Binge Drinking, By Race/Ethnicity, Oregon School District (source: OHT)					
Gender	**% Reporting Binge Drinking (weighted)**	**# Reporting Binge Drinking (weighted)**	**# Reporting Binge Drinking (unweighted)**	**# Participating Students**	**# Participating Schools**
African American	20%	8	6	40	
American Indian	8%	2	3	24	
Asian/Pacific Islander	27%	14	15	52	
Hispanic	12%	7	6	60	
White	22%	44	45	200	
Total	**20%**	**75**	**75**	**376**	**1 out of 5**

No Suppression (but unweighted values not reported)			
Percent of 11th Graders Reporting Binge Drinking, By Race/Ethnicity, Oregon School District (source: OHT)			
Gender	**% Reporting Binge Drinking (weighted)**	**# Participating Students**	**# Participating Schools**
African American	20%	40	
American Indian	8%	24	
Asian/Pacific Islander	27%	52	
Hispanic	12%	60	
White	22%	200	
Total	**20%**	**376**	**1 out of 5**

Scenario 6

Prior to Suppression			
Age-Adjusted Lung Cancer Death Rate Among Men, By Race/Ethnicity, Oregon County			
Race/Ethnicity	**Age-Adjusted Rate per 100,000**	**# of Deaths Among Men**	**Total Male Population**
American Indian/Alaska Native NH	236	1	780
Asian/Pacific Islander NH	1,720	1	30
Black NH	747	1	248
Hispanic	41	2	12,292
White NH	80	15	9,846

- Which traits are confidential information? *Lung cancer*
- Which traits could be linked with other data sources? *Sex; race/ethnicity; residence in Oregon County*
- Risk of disclosure of confidential information? *Deaths are public record. It might be possible for someone in a close-knit community to know everyone in this demographic group who died and by process of elimination determine who died of lung cancer.*
- Decision: *Suppress sensitive cell based on small denominator. Complementary suppression isn't necessary because neither a total count nor unknown race count are given. No need to suppress age-adjusted rate because count can't be back-calculated from it.*



Suppressed			
Age-Adjusted Lung Cancer Death Rate Among Men, By Race/Ethnicity, Oregon County			
Race/Ethnicity	**Age-Adjusted Rate per 100,000**	**# of Deaths Among Men**	**Total Male Population**
American Indian/Alaska Native NH	236	1	780
Asian/Pacific Islander NH	1,720	*	30
Black NH	747	1	248
Hispanic	41	2	12,292
White NH	80	15	9,846
* Data not reported to protect confidentiality			

Scenario 7

Prior to Suppression			
Prevalence Rate of HIV/AIDS Among Male Children Age 1-14, By Community			
Community	**Rate per 100,000**	**# of Cases Among Males Age 1-14**	**# of Males Age 1-14**
Community A	26.1	6	22,948
Community B	20.0	1	5,000

- Which traits are confidential information? *HIV/AIDS*
- Which traits could be linked with other data sources? *Sex; age; residence in Community A or B*
- Risk of disclosure of confidential information? *It may be commonly known that there is one child in a particular school in Community B who is frequently hospitalized. Because the underlying population is large there's no way that anyone could reasonable infer that the child they know who is frequently ill is the only one in the population who could be the pediatric HIV/AIDS patient.*
- Decision: *No need to suppress.*

No Suppression			
Prevalence Rate of HIV/AIDS Among Male Children Age 1-14, By Community			
Community	**Rate per 100,000**	**# of Cases Among Males Age 1-14**	**# of Males Age 1-14**
Community A	26.1	6	22,948
Community B	20.0	1	5,000

Scenario 8

Prior to Suppression			
HIV/AIDS Among Male Veterans, By Community			
Community	**# of Cases Among Male Veterans**	**# of Male Veterans**	**# of Males**
Community A	2	100	22,948
Community B	4	5	5,000

- Which traits are confidential information? *HIV/AIDS*
- Which traits could be linked with other data sources? *Sex; residence in Community A or B; veteran status*
- Risk of disclosure of confidential information? *Veterans in this small community may know each other well. The true underlying population is the number of male veterans – not total males. This is a small population that is easily identifiable using public data sources.*
- Decision: *Suppress sensitive cell based on small denominator.*

Suppressed			
HIV/AIDS Among Male Veterans, By Community			
Community	**# of Cases Among Male Veterans**	**# of Male Veterans**	**# of Males**
Community A	2	100	22,948
Community B	*	5	5,000
* Data not reported to protect confidentiality			

Appendix A: Examples of confidential and non-confidential data elements

The following lists are intended to be representative of types of confidential and non-confidential data elements. The lists are not intended to be exhaustive. Specific items of importance will vary by program.

Confidential Information:

- Cause of death
- Disease status
- Medical test results
- Medical diagnoses
- Drug use
- Pregnancy
- Prenatal care
- Birth weight
- Sexual preference

Non-Confidential Information:

- Vital Events:
  - Births
  - Deaths
- Demographic Information:
  - Age
  - Education
  - Ethnicity
  - Gender/sex
  - Insurance source
  - Marital status
  - Month or year of birth or death
  - Occupation
  - Race
  - Residence in census tract, city, county or ZIP code
  - Veteran status
  - Population of census tract, city, county, ZIP code, state

Appendix B: Bibliography

1. **Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.** http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

2. **Healthy People 2010 Criteria for Data Suppression.** Richard J. Klein, Suzanne E. Proctor, Manon A. Boudreault, Kathleen M. Turczyn, CDC Statistical Notes, no 24. Hyattsville, Maryland: National Center for Health Statistics, July 2002. http://www.cdc.gov/nchs/data/statnt/statnt24.pdf

3. **Information-Theoretic Disclosure Risk Measures in Statistical Disclosure Control of Tabular Data.** Josep Domingo-Ferrer, Anna Oganian, Vicenc Torra, Proceedings from 14th International Conference on Scientific and Statistical Database Management, 2002 (pp.227-231).

4. **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.

5. **NCHS Staff Manual on Confidentiality.** National Center for Health Statistics, 2004. http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

6. **Small Numbers, Disclosure Risk, Security, and Reliability Issues in Web-based Data Query Systems**, Barbara A. Rudolph, Gulzar H. Shah, and Denise Love, Journal of Public Health Management Practice, 2006, 12(2), 176-183.

7. **Statistical Issues in Interactive Web-based Public Health Data Dissemination Systems.** Michael A. Stoto, RAND Health Working Paper Series: WR-106, October 2003. Prepared for the National Association of Public Health Statistics and Information systems. http://www.rand.org/pubs/working_papers/WR106.html

8. **STATISTICAL POLICY WORKING PAPER 22, Report on Statistical Disclosure Limitation Methodology**. Federal Committee on Statistical Methodology, Statistical and Science Policy Office of Information and Regulatory Affairs Office of Management and Budget, December 2005. http://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf

Appendix C: Guidelines used in other states

A search of the literature cited in Appendix B has found no nationally accepted standards for disclosure control protocols to prevent breaches of confidentiality. The following are examples of small number reporting guidelines developed in other state health departments.

1. **Department of Public Health Confidentiality Procedures.** Massachusetts Department of Public Health, Revised October 1, 2012 (p.40-42). http://www.mass.gov/eohhs/gov/departments/dph/programs/admin/privacy/privacy-confidentiality/confidentiality-policy-and-procedures.html

2. **Guidelines for the Public Release of Public Health Data**. Division of Public Health Services, Health Statistics and Data Management Section, State of New Hampshire Department of Health and Human Services, September 30, 2008. http://www.dhhs.nh.gov/dphs/hsdm/documents/publichealthdata.pdf

3. **Guidelines for the Release of Public Health Data Derived from Personal Health Information**. Office of Epidemiology and Scientific Support, Montana Department of Public Health and Human Services, July 2011. http://www.dphhs.mt.gov/publichealth/epidemiology/documents/Guidelines_Reporting_PHI.pdf

4. **Guidelines for Working with Small Numbers.** Washington State Department of Health, Revised October 2012. http://www.doh.wa.gov/DataandStatisticalReports/DataGuidelines.aspx

5. **Guidelines for Working with Small Numbers.** Health Statistics Section, Colorado Department of Public Health and Environment. http://www.cohid.dphe.state.co.us/smnumguidelines.html (adapted from Washington State guidelines)

6. **Report of Guidelines for Data Result Suppression**. Utah Department of Health, Data Suppression Decision Rules Work Group, October 5, 2009. http://health.utah.gov/opha/IBIShelp/DataSuppression.pdf

Appendix D: Glossary

Aggregated data – Data released by OPHD that are presented as a statistical summary – such as a frequency or rate – rather than as a list of individual persons or events. Sometimes also referred to in this report as "statistical data".

Aggregation – Creation of a statistical summary of data by grouping individual persons or events with common attributes such as age group, geographic area or time period and calculating a summary measure for each group such as frequency or rate.

Cell suppression – A disclosure control method that involves withholding the release of sensitive data points – or, cells – on a table, chart or map of aggregated data.

Client data – A dataset that contains health or encounter records of all the individuals who participate in or receive services from a particular health or social service agency (e.g., WIC).

Complementary suppression – Suppression of non-sensitive cells on an aggregated data table in order to prevent the back-calculation of suppressed sensitive cells from marginal totals.

Confidentiality – The concept that private information obtained from an individual will be kept secure and not disclosed without the authorization of that individual.

Data linkage – The use of publicly available information, detailed external data sources, or general community knowledge about a population to infer the identity of or confidential information about individuals represented in statistical data released by OPHD.

Disclosure control methods – A variety of methods that data analysts can apply to aggregated data prior to release in order to minimize the risk of disclosing confidential information about identifiable individuals.

Disclosure risk - A measure of the risk of disclosing confidential information about or the identity of any individual represented in data released by OPHD. In this report, the term specifically applies to the risk associated with any particular data point in aggregated data.

Full count data – A dataset that theoretically contains every instance of a qualified event (e.g., death, birth, cancer or communicable disease diagnosis, immunization record, health care encounter).

Marginal totals – Totals across sub-categories on an aggregated data table.

Perturbation – A collection of disclosure control methods that alter the cell values of statistical tables to introduce uncertainty, thereby protecting sensitive cells. Values in the underlying dataset are either swapped, shifted randomly, or synthesized, using the same underlying distribution. Some of these methods preserve the marginal totals while others cause shifts in the entire dataset. The values are slightly modified but the distribution of the data is preserved.

Reliability – The statistical stability of a rate or estimate.

Rounding – A disclosure control method in which all cells in the table are rounded to the nearest value of multiples of a selected integer. This introduces uncertainty into the reported counts to protect the sensitive cells. Controlled rounding preserves the rounded marginal totals.

Safe Harbor method of de-identification –One of the two methods described by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule to protect the confidentiality of Protected Health Information prior to release. This method involves removing 18 types of data fields considered to be personal identifiers. While aggregated data produced by OPHD rarely includes the most explicit of these identifiers – such as name and Social Security number – it frequently does include identifiers such as county or ZIP code of residence. The other method described by HIPAA – Expert Determination – allows for a person with appropriate knowledge and experience to assess the level of risk that an individual who is a subject of the data can be identified and apply methods for rendering the data not individually identifiable.

Sensitive cell – The value(s) in the data table, chart or map that – if reported – may pose a significant risk of inadvertent disclosure of confidential information.

Statistical data – Data released by OPHD that are presented as a statistical summary – such as a frequency or rate – rather than as a list of individual persons or events. Sometimes also referred to in this report as "aggregated data".

Survey data – A dataset that contains information gathered from asking questions of randomly selected respondents. For complex surveys such as the Behavioral Risk Factors Surveillance System (BRFSS), these respondents are weighted so that estimates can be made to infer the behavior of an entire population (such as all adult Oregonians).

Threshold cell – In aggregated data, the cell that contains the threshold value that is used to identify sensitive cells. It may or may not be the same as the sensitive cell. It may or may not be on the same published table, chart or map as the sensitive cell.

Threshold of acceptable risk – The cut-off point beyond which the disclosure risk for a particular data point is considered to be too high to release the data.

Underlying population – Everyone in a population defined by specific characteristics (such as age, sex, race, county of residence) who is at risk of experiencing a certain health event and, therefore, might be a data subject in aggregated data released by OPHD. For example, the underlying population for a teen pregnancy rate is the total population of teenage females in the geographic area of interest. From a data analyst perspective, this population is generally thought of as the denominator in a summary measure such as a rate.