

Guidelines for Working with Small Numbers

Revision Date: October 15, 2012

Primary Contact: Juliet VanEenwyk, Ph.D., State Epidemiologist for Non-Infectious Conditions

Secondary Contact: Steven C. Macdonald, Ph.D., Surveillance Epidemiologist, Environmental Public Health Division

Purpose

What is new, and how does this affect public health assessment?

Scope of the "Guidelines for Working with Small Numbers"

Why are small numbers a concern in public health assessment?

What constitutes a breach of confidentiality?

Why do we question the reliability of statistics based on small numbers?

Why do we have guidelines rather than standards?

Guidelines for Working with Small Numbers

General Considerations

Assessing Confidentiality Issues

Examine denominator size for each cell

Examine numerator size for each cell

Consider the proportion of the population sampled

Consider the nature of the information

How to Reduce Risk of Confidentiality Breach

General approach

Aggregation

Cell suppression

Other methods

Group identification

Recommendations to Protect Confidentiality

Assessing and Addressing Statistical Issues

Relative standard error

Increase numerator size for rare events or denominator size for samples

Include confidence intervals

Recommendations to Address Statistical Issues

Glossary

References

Resources

Relevant Policies, Laws and Regulations

Appendix 1: Rule-based use of suppression and aggregation

Purpose

The Assessment Operations Group in the Washington State Department of Health (department) works with local health jurisdiction to develop guidelines related to data collection, analysis and use in order to promote good professional practice among staff involved in assessment activities within the department and in local health jurisdictions in Washington. While the guidelines are intended for audiences of differing levels of training, they assume a basic knowledge of epidemiology and biostatistics. They are not intended to recreate basic texts and other sources of information; rather, they focus on issues commonly encountered in public health practice and, where applicable, refer to issues unique to Washington State.

What is new and how does this affect public health assessment?

These guidelines have been expanded to include survey as well as population-based data. They also incorporate the use of relative standard error when assessing statistical stability.

Scope of the “Guidelines for Working with Small Numbers”

The department and local health jurisdictions routinely make aggregated health and related data available to the public. Historically, these data were presented as static tables. Over the past decade, however, interactive Web-based data query systems allowing users to build their own tables have become more common. The department and local health jurisdictions also release files containing record-level data. The following guidelines apply to releases of aggregated population-based and survey data available to the public other than those mandated by law. Releases include both static data tables and graphics, such as charts and maps, as well as tables and graphics produced through interactive query systems. The guidelines do not apply to release of record-level data. Release of record-level data is governed by federal and state disclosure laws, which can be specific to a dataset, and by Institutional Review Boards if the data are used for research.

Why are small numbers a concern in public health assessment?

Public health policy decisions are fueled by information, which is often in the form of statistical data. Questions concerning health outcomes and related health behaviors and environmental factors often are studied within small subgroups of a population, because many activities to improve health affect relatively small populations. Additionally, continuing improvements in the performance and availability of computing resources, including geographic information systems, and the need to better understand the relationships among environment, behavior and health have led to increased demand for information about small populations. These demands are often at odds with the need to preserve privacy and data confidentiality. Small numbers also raise statistical issues concerning the accuracy, and thus usefulness, of the data.

What constitutes a breach of confidentiality?

A breach of confidentiality occurs when analysts release information in a way that allows an individual to be identified and reveals confidential information about that person (that is, information which the person has provided in a relationship of trust, with the expectation that it will not be divulged in an identifiable form). The following guidelines provide cues to situations that present high risk for a breach of confidentiality and suggestions on how to reduce this risk. In addition to these guidelines, analysts should be familiar with relevant federal and Washington State laws and regulations and department policies. (See [Relevant Policies, Laws and Regulations](#).) **Federal and state laws and regulations and department policies supersede guidance provided in this document.**

Why do we question the reliability of statistics based on small numbers?

Estimates based on a random sample of a population are subject to sampling variability. Rates and percentages based on full population counts are also subject to random variation. (See [Guidelines for Using Confidence Intervals for Public Health Assessment](#) for a short discussion of variability in population-based data.) The random variation may be substantial when the measure, such a rate or percentage, has a small number of events in the numerator or a small denominator. Typically, rates based on large numbers provide stable estimates of the true, underlying rate. Conversely, rates based on small numbers may fluctuate dramatically from year to year, or differ considerably from one small place to another, even when differences are not meaningful. Meaningful analysis of differences in rates between geographic areas or over time requires that the random variation in rates be quantified; this is especially important when rates or percentages are based on small numerators or denominators.

Why do we have guidelines rather than standards?

It is generally easier for data analysts to conform to standards than to apply guidelines in deciding whether to publish information based on small numbers. Several factors, however, make it difficult to establish standards that protect confidentiality and provide reliable estimates while also maximizing the availability of health and related data. For example:

- Different public health datasets have different laws and rules governing confidentiality.
- The feasibility of scrutinizing tables to assure protection of confidentiality differs depending on the number of tables produced. For example, the data analyst might be able to maintain confidentiality with smaller minimum numbers when publishing a handful of static data tables that can be inspected individually and in combination than when developing an interactive query system capable of producing hundreds of tables.

We have not found nationally accepted standards for suppression of data due to potential breaches of confidentiality or statistical stability. Different units, for example, use different methods even within a single federal agency. The lack of a single national standard, perhaps, speaks to the problems inherent in such an approach. Thus, data analysts need to use judgment in determining whether aggregated data available to the public protect confidentiality and are precise and stable enough to allow users to draw reasonable conclusions.

Guidelines for Working with Small Numbers

General Considerations

These guidelines address both confidentiality and statistical issues in working with small numbers. In some department data systems, such as the AIDS registry, the entire database is considered confidential. In other systems, such as the birth certificate system, many but not all data items are confidential. In yet other systems, none of the items are confidential, such as most records in the death certificate system. Survey data often contain confidential information and may also contain information that could be used to identify an individual (e.g., there might be small numbers of individuals with a particular visible characteristic in a small geographical area). A first step in using these guidelines is to determine if the datasets you are working with contain confidential or potentially identifiable information. If so, the following section on protecting confidentiality is relevant. Otherwise, you need only concern yourself with the statistical issues section.

Assessing Confidentiality Issues

With population-based data, most problems with confidentiality occur when the population from which the events arise (i.e., denominator) is small, but the number of events (i.e., numerator)

might also be important. For example, if there are 5,000 individuals in a specific age-race-sex group in a single county, the likelihood of identifying a single individual from data in a published table is quite small. In smaller populations, it is more likely that an individual might be identifiable. However, even in larger populations, it is conceivable that a single individual might be identifiable, if there are only one or two individuals with some special characteristic. For example, in a modest sized community, it may be commonly known that there is only one child who is frequently hospitalized, and a table showing that this community has one case of pediatric HIV-AIDS could unintentionally allow knowledgeable residents to infer the child's illness. Similarly, if a unique individual, such as one of the parents of the frequently hospitalized child described above, were drawn into a survey, knowledgeable residents might infer the illness of the child from survey data indicating one child with HIV-AIDS in that community. Thus, the same cautions for population data generally apply to survey data as well.

Examine denominator size for each cell. Prior to disseminating tables derived from datasets that contain confidential information, analysts should consider the size of the denominators, i.e., the population size represented in each cell, row or column in the table. Caution should increase as the population sizes shrink, because the risk of violating confidentiality increases when data are tabulated for small groups, as might occur for example, when analyzing data by racial categories in small geographic areas.

Among the several national standards for minimum denominator size we identified, none were relevant to all data dissemination situations faced by state and local health agencies in Washington State.

- For population-based tabular data, the federal [Office of Management and Budget's \(OMB\) 1999 Checklist on Disclosure Potential of Proposed Data](#) recommends assessing the risk of a confidentiality breach for populations under 100,000. (OMB 1999) The National Center for Health Statistics (NCHS) references this stipulation in its [2004 Staff Manual on Confidentiality](#). (NCHS 2004) Based on this recommendation, the vast majority of tables using Washington State population-based data would need to be evaluated for their potential to breach confidentiality.
- The Health Insurance Portability and Accountability Act (HIPAA) provides a guide to sharing record-level data that is also relevant to data tables. HIPAA allows sharing of records for geographic areas (3-digit ZIP codes) containing more than 20,000 people in combination with omitting identifiers that are unique to individuals (e.g., name, Social Security number), all elements of dates except for year, and single years of age for those over 89 years. (NIH 2004) Although data-sharing agreements further protect confidentiality by limiting how record-level HIPAA data can be used and assign responsibility of protecting confidentiality to the data analyst who publishes the data, allowing record-level data to be shared for geographic areas with more than 20,000 people suggests that release of aggregate data in which the base population is 20,000 may be adequate to protect confidentiality.
- We have found one example of a national practice where suppression is based only on the numerator and thus, no minimum denominator size is required. The National Program of Cancer Registries United States Cancer Statistics website provides annual state-level cancer incidence and death data by race, ethnicity, and sex. [Technical notes](#) for this site state, "The cell suppression threshold value of 16, which was selected to reduce misuse and misinterpretation of unstable rates and counts in this report, is more than sufficient to protect patient confidentiality." (NPCR 2008) Website maps indicate that states can also request data suppression, but neither the technical notes nor the website explains reasons states choose to suppress data.

Examine numerator size for each cell. Data analysts should consider the number of events in each cell of a table to be released (i.e., the numerator for a rate calculation). As with denominator assessment, there is no single national standard for determining when small numerators might lead to breaches of confidentiality. In fact, disclosing that there has been one case of a disease in a state or county might not breach confidentiality if no other detail is given. Small numerators are

of increasing concern for confidentiality if there are also small numbers of individuals with the reported characteristic(s) in the population. If the characteristic is observable (e.g., distinctive physical characteristics) or the participants in the survey are known, risk for identification may be further increased. For data tables, the [2004 NCHS Staff Manual on Confidentiality](#) requires:

- No single cells containing all observations of a row or column.
- At least five observations for a row or column total in a cross-tabulation.
- At least five observations total.

NCHS might change this guidance when the 2004 manual is updated. Since May 2011, the CDC interactive query system, WONDER, has suppressed birth and death data if there are not at least 10 observations. (WONDER 2012) Other groups at CDC use different criteria. For example, the Environmental Public Health Tracking Network currently suppresses rates based on non-zero counts less than six.

We do not recommend automatic suppression of tables or cells within tables due to small numbers. Rather, we recommend that if the number of cases or events in a cell is less than 10, the data analyst consider the likelihood of a breach of confidentiality. A count of no events in the cell is unlikely to be a threat to confidentiality unless it provides meaningful information about the remaining 100% of participants, but a count of one to nine events may be a threat to confidentiality. We selected this cut point to be consistent with what will likely be the new standard for NCHS when it releases a new staff manual.

Consider the proportion of the population sampled. For survey data, the potential for breaches of confidentiality decreases as the proportion of the population in the sample decreases. The [NCHS Staff Manual on Confidentiality](#) states that less than 10% might generally be safe, but cautions that there could be exceptions, as for example, when so much detail is presented that an individual with unusual characteristics could be identified. If the sampling probabilities are large and the pool of potential survey participants in the population are known (e.g., students in a school), this may increase the risk of identification, especially when the numerator or denominator is small. The U.S. Office of Management and Budget's [Checklist on Disclosure Potential of Proposed Data Releases](#) classifies the former decennial census long form sample of about 17% of households as comprising a "large portion of the population," thus requiring disclosure risk assessment before releasing data tables.

Consider the nature of the information. The [Checklist on Disclosure Potential of Proposed Data Releases](#) identifies examples of variables that are visible and, therefore, pose increased risk of disclosure. Examples include income and related variables such as property value and rent or mortgage payments; unusual occupation; unusual health condition; very old age; and race or ethnicity. Physical characteristics such as obesity are also visible and might increase risk of individual identification.

How to Reduce the Risk of a Confidentiality Breach

General Approach. The general approach to privacy protection involves what has been termed "computational disclosure control," which includes both aggregation of data values in the dataset before analysis, and cell suppression in a table after analysis (Sweeney 1997). Web-based query systems, such as that developed by the Washington Tracking Network (WTN), aggregate data using rule-based static and dynamic parameter control in order to minimize suppression. Appendix 1 outlines the aggregation and suppression rules used by the WTN to protect confidentiality.

Aggregation. Aggregation of data values is appropriate for fields with large numbers of values, such as dates, diagnoses and geographic areas; it is the primary method used to create tables with no small numbers as denominators or numerators. Granularity refers to the degree of detail or precision in data, or the fineness with which data fields are subdivided. The following table shows examples.

		Granularity: Aggregation		
Field	Type	<i>Fine</i>	<i>Medium</i>	<i>Coarse</i>
Age	Continuous	Year of birth	5-year age group	10-year age group
Date of occurrence	Continuous	Month	Year	Multiple years combined
Diagnosis	Nominal	Complete ICD code	Three-digit ICD	"Selected cause" Tabulation
Geography	Ordinal (spatial)	Zip code, census tract	County	State

In addition to considering each field on its own, aggregation should consider each field in combination with others. When numbers are large, data are commonly disaggregated across multiple fields, resulting in release of multiple data tables. However, when numbers are small, protecting confidentiality often requires limiting the number of fields which are disaggregated simultaneously, resulting in release of fewer data tables. When numbers are tiny, tables may be limited to those where only one field is disaggregated at a time.

Cell suppression. When it is not possible, or desirable, to create a table with no small numbers as denominators or numerators, then cell suppression is used. "Primary" cell suppression is used to withhold data in the cell that fails to meet the threshold, followed by secondary or complementary suppression of three other cells in order to avoid inadvertent disclosure through subtraction. Note that cell suppression is a method of last resort, due to the often unavoidable side-effect of suppressing releasable data values as a consequence of complementary suppression, and due to the amount of labor necessary to implement the method. The following table shows an example of complementary suppression. In this example, even if all the cells except for the cell in the upper left (0–34 Black) meet the threshold for release, data in three additional cells need to be suppressed.

Age	Black	White	Other	Total
0–34	Suppress	30	Suppress	60
35–64	Suppress	60	Suppress	150
65+	70	90	80	240
Total	120	180	150	450

If the value of the information in all cells is not the same, data analysts should suppress cells that provide less useful information. In the previous table, "other" includes a diversity of racial groups and such aggregation is usually not meaningful for addressing public health problems in Washington State. In the same table, suppressing information for the two youngest age groups might be best, if the condition is one that primarily affects older individuals. Alternatively, if the goal of the table is to provide data for targeting prevention to middle-aged people, complementary suppression of data for the youngest and oldest age groups might be preferable.

Suppression algorithms for protecting confidentiality are best based on a combination of denominator and numerator values. For example, the CDC Environmental Public Health Tracking Network currently starts with a denominator rule threshold of 100,000 for displaying counts. At 100,000 or more, all counts can be displayed. When the denominator is less than 100,000, the rule specifies that counts are displayed only if there are no events or six or more events. (See [Appendix 1.](#))

Other methods. When neither of these methods (aggregation of data values to create coarser granularity or cell suppression) is satisfactory, the data analyst might want to omit certain fields from analysis entirely. For example, for a department release of asthma data, it was not possible to achieve adequately large cell denominators in annual county-level data showing both age-specific and gender-specific counts and rates. Those publishing the data opted to omit the gender-specific data, and display only tables of age-specific data, on the grounds that no

intervention programs targeted groups differently on the basis of gender, but most intervention programs target age groups differently.

Group identification. Data in a table provides information on the probability that someone in a defined group has a given characteristic. The [2004 NCHS Staff Manual on Confidentiality](#) describes this as “probability-based” disclosure. The manual recommends suppression if a table reveals “that a highly specific group had an extremely high probability of having a given sensitive characteristic...” (p.15) The manual also notes that “only in unusual circumstances could any such disclosure be considered unacceptable.” (p.15) While suppression due to probability-based disclosure would be rare in public health data tables, data analysts should consider this issue when publishing confidential information, especially when the prevalence of a sensitive characteristic in specific group is high.

Recommendations to Protect Confidentiality

The following guidelines can be used to alert data analysts to situations that require particular attention to avoid breaches of confidentiality. **They are not requirements for suppressing data.** For example, the department routinely publishes data by county. In 2010, nine counties had populations less than 20,000 and three had populations less than 20,000 person-years when combining three years of data (i.e., 2009–2011). Even though some counties do not meet the 20,000 threshold, most department programs are comfortable publishing numbers or rates by county when the population denominator is the entire county population. However, programs carefully evaluate the potential for breaches of confidentiality when considering publishing the same data by demographic characteristics, because denominators shrink when considering subpopulations within counties. Depending on the type of data and the types of demographic characteristics, programs might conclude that there is not a risk for a breach of confidentiality and they can safely publish the data. Alternatively, they might conclude there is a risk of inadvertent disclosure and decide not to publish such tables at all or not publish for selected counties.

- Evaluate the risk for a breach of confidentiality for denominators less than 100,000. Be especially cautious with denominators less than 20,000.
- Be cautious when reporting counts less than 10.
- Be cautious when reporting a specific confidential characteristic of a population if a very high proportion of the population has this characteristic.
- When producing multiple tables from the same dataset, be careful that users cannot derive confidential information through a process of subtraction.
- If data are suppressed, provide an indicator (e.g., asterisk) in the suppressed cell and a legend under the table explaining the reason for suppression.

Assessing and Addressing Statistical Issues

Relative standard error. The relative standard error (RSE) provides a measure of reliability for statistical estimates. The RSE is computed by dividing the standard error of the estimate by the estimate and multiplying by 100 to convert it to a percentage. When the RSE is large, the estimate is imprecise. In these instances, the data analyst needs to balance issues of the “right to know” with presenting data that might be misleading.

There is no single national standard for deciding when the RSE is so large that one should not present the data. Federal agencies and even units within a single federal agency might use different approaches. For example, within the Centers for Disease Control and Prevention:

- A 2009 NCHS report suppressed data with RSEs greater than 40% and noted that data with RSEs of 30–40% were unreliable. (Fryar 2009)

- A 2010 NCHS publication suppressed data when RSEs were greater than 50% and noted that estimates with RSEs of 30–50% were unreliable. (NCHS 2010)
- A 2011 NCHS publication suppressed data based on sample size, but not RSEs. Estimates with RSEs of 30–59% were marked as unreliable. (Bercovitz 2011)
- CDC Environmental Public Health Tracking Network displays all rates that are not suppressed for confidentiality protection. Rates with RSEs of 30% or greater are annotated as unreliable. (NEPHTN 2008) (See [Appendix 1.](#))
- The National Program of Cancer Registries suppresses data due to concerns about the statistical stability when the number of events is less than 16, stating that a count of fewer than about 16 results in an RSE of about 25%. (NPCR 2008)

Different programs at the department use different practices based on RSE. Currently, some programs do not publish data when RSEs are greater than 30%. In contrast, the Washington Tracking Network follows standards for the CDC Environmental Public Health Tracking Program and marks data with RSEs greater than 30% as unreliable, but does not suppress data for statistical reasons. (See [Appendix 1.](#)) A middle ground is to suppress data with RSEs above a given cut point, such as RSEs of 40, 50 or 60% as in the NCHS examples given above, and mark as unreliable data with RSEs between 30% and the cut point. The approach taken by different data analysts might vary depending on the primary audience and purpose of the publication.

Increase numerator size for rare events and sample size for samples. As the proportion of data suppressed or annotated as unreliable increases, the value of the data table decreases. Increasing the numerator for population data based on a Poisson distribution and the sample size for surveys will improve the stability of the estimate and reduce the RSE. Techniques to improve stability within a fixed sample size or population include the following aggregation methods:

- Combining multiple years of data
- Collapsing data categories
- Expanding the geographic area under consideration

Include confidence intervals. We recommend including confidence intervals when presenting rates, especially when the RSE is large. (See [Guidelines for Using Confidence Intervals for Public Health Assessment.](#)) Generally, based on a Poisson distribution for rare events, rates based on fewer than 12 events have an RSE over 30% and wide confidence intervals. For example, an infant death rate of nine per 1,000, based on nine deaths in a population of 1,000 live births, has an RSE of 33% and a Poisson-based 95% confidence interval between four and 17. This is not very precise information and if the data are presented, users need to know this.

In instances where it is not feasible to incorporate confidence intervals into a data table, we recommend that data analysts:

- Report the numerator and denominator on which the rate is based, for example in a legend or table subtitle.
- Flag rates with RSEs greater than 30% (if these data are not suppressed) and include a footnote to indicate that data are unreliable and imprecise.

Recommendations to Address Statistical Issues

- Include confidence intervals to show the extent of variation that might occur by chance.
- When RSEs are greater than 30% mark these data as unreliable.
- Consider suppression as a method of last resort when data are so unreliable and imprecise that they cannot be used effectively for planning programs or informing policy

decisions. If data are suppressed, provide an indicator (e.g., asterisk) in the suppressed cells and a legend under the table explaining the reason for suppression.

- Consider using geographic modeling, including using Bayesian smoothing, as an alternative to suppression. A discussion of this method, however, is beyond the scope of these guidelines.

Glossary

Confidential data/information: Information that an individual or establishment has provided in a relationship of trust, with the expectation that it will not be divulged in an identifiable form. The confidentiality of specific data elements or information in individual databases or record systems may be defined by federal or state laws or regulations, or policies or procedures developed for those systems.

Confidentiality breach: An unauthorized release of identifiable or confidential data/information, which may result from a security failure, intentional inappropriate behavior, human error or natural disaster. A breach of confidentiality may or may not result in harm to one or more individuals.

Individually identifiable data/information: Data/information that identifies, or is reasonably likely to be used to identify, an individual or an establishment protected under confidentiality laws. Identifiable data/information may include, but is not limited to, name, address, telephone number, Social Security number and medical record number. Data elements used to identify an individual or protected establishment can vary depending on the geographic location and other variables (e.g., rarity of person's health condition or patient demographics). For purposes of this guideline, "identifiable information" includes [potentially identifiable information](#).

Number of events: The number of persons or events represented in any given cell of tabulated data (e.g., numerator). (See [Guidelines for Using and Developing Rates for Public Health Assessment](#).)

Population or sample size: The total number of persons or events included in the calculation of an event rate (e.g., denominator). (See [Guidelines for Selection of Population Denominators](#).)

Potentially identifiable information: Information that does not contain direct identifiers, such as name, address or specific dates, but provides information that could be used in combination with other data to identify individuals.

Rate: A measure of the frequency of an event per population unit. (See [Guidelines for Using and Developing Rates for Public Health Assessment](#).) In these guidelines the terms rate, proportion and percent are interchangeable.

Sensitive personal information: Whereas confidential personal information means information collected about a person that is readily identifiable to that specific individual, sensitive personal information extends beyond that to information which may be inferred about individuals, where that information is associated with some stigma. Examples are certain diseases, health conditions or health practices. The sensitivity of certain personal information may vary between communities.

References

Bercovitz A, Moss A, Sengupta M, et al. An overview of home health aides: United States, 2007. National health statistics reports; no 34. Hyattsville, MD: National Center for Health Statistics; 2011. <http://www.cdc.gov/nchs/data/nhsr/nhsr034.pdf> Accessed October 12, 2012.

Fryar CD, Merino MC, Hirsch R, Porter KS. Smoking, alcohol use, and illicit drug use reported by adolescents aged 12-17 years: United States, 1999-2004. National health statistics reports; no 15. Hyattsville, MD: National Center for Health Statistics; 2009. <http://www.cdc.gov/nchs/data/nhsr/nhsr015.pdf>. Accessed October 12, 2012.

National Center for Health Statistics. *Table 1. Health insurance coverage status, coverage type, and selected characteristics, for persons of all ages, January-June 2010*. Hyattsville, MD: National Center for Health Statistics; 2010. http://www.cdc.gov/nchs/data/health_policy/Health_Insurance_Selected_Characteristics_Jan_June_2010.pdf. Accessed June 21, 2012.

National Environmental Public Health Tracking Network (NEPHTN). *Data Re-release Plan Version 2.5*. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Environmental Health, Division of Environmental Hazards and Health Effects, Environmental Health Tracking Branch; 2008. http://ephtracking.cdc.gov/docs/Tracking_Re-Release_Plan_v2.5.pdf. Accessed September 20, 2012.

National Institutes of Health. *Research Repositories, Databases, and the HIPAA Privacy Rule*. Washington, DC: U.S. Department of Health and Human Services, National Institutes of Health; Posted January 12, 2004 (revised: 7/02/04). http://privacyruleandresearch.nih.gov/research_repositories.asp. Accessed October 15, 2012.

NCHS Staff Manual on Confidentiality. Hyattsville, MD: Department of Health and Human Services, Public Health Service, National Center for Health Statistics; 2004. <http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf>. Accessed January 2, 2011.

NPCR Technical Notes: Statistical Methods: Suppression of Rates and Counts. Atlanta, GA: Centers for Disease Control and Prevention, National Program of Cancer Registries, Division of Cancer Prevention and Control, National Center for Chronic Disease Prevention and Health Promotion; 2008. http://www.cdc.gov/cancer/npcr/uscs/technical_notes/stat_methods/suppression.htm. Accessed September 20, 2012.

OMB Checklist on Disclosure Potential of Proposed Data Releases. Washington, DC: Office of Management and Budget; 1999. <http://www.fcs.m.gov/committees/cdac/index.html>. Accessed September 20, 2012.

WONDER Multiple Cause of Death 1999-2009. Atlanta, GA: Centers for Disease Control and Prevention; 2012. http://wonder.cdc.gov/wonder/help/mcd.html#Assurance_of_Confidentiality. Accessed July 18, 2012.

Resources

Klein RJ, Proctor SE, BVoudreault MA, Turczyn KM. Healthy People 2010 criteria for data suppression. Statistical Notes, no 24. Hyattsville, MD: National Center for Health Statistics; June 2002.

NCHS Staff Manual on Confidentiality. Hyattsville, MD: Department of Health and Human Services, Public Health Service, National Center for Health Statistics; 2004. <http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf>. Accessed January 2, 2011.

OMB Checklist on Disclosure Potential of Proposed Data Releases. Washington, DC: Office of Management and Budget; 1999. <http://www.fcs.m.gov/committees/cdac/index.html>. Accessed September 20, 2012.

Sweeney L. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine & Ethics*. 1997;25:98-110.

Relevant Policies, Laws and Regulations

[Release of Confidential Information: Department Policy 17.006](#) (link accessible to department employees only)

[Medical records—health care information access and disclosure: Chapter 70.02 RCW](#)

[Public records act. Chapter 42.56 RCW](#)

[Executive Order on Public Records Privacy Protections: EO 00-03.](#)

Vital records

- Requesting a listing or file of vital records with personal identifiers: [WAC 246-490-030](#)
Requesting vital records information without personal identifiers: [WAC 246-490-020](#)

The following examples, provided by the department data custodians, include the major datasets used for assessment in Washington.

Birth records: [RCW 70.58.055](#) and [WAC 246-491-039](#)

Death records: [RCW 9.02.100](#) and [WAC 246-490-110](#) (deaths related to abortion), [WAC 246-491-039](#) (fetal death records), [RCW 70.24.105](#) (deaths related to HIV-AIDS).

HIV/AIDS and other communicable disease data: [RCW 70.24.105](#) and [WAC 246-101](#).

Hospital discharge data: [RCW 43.70.052](#) and [WAC 246-455-080](#)

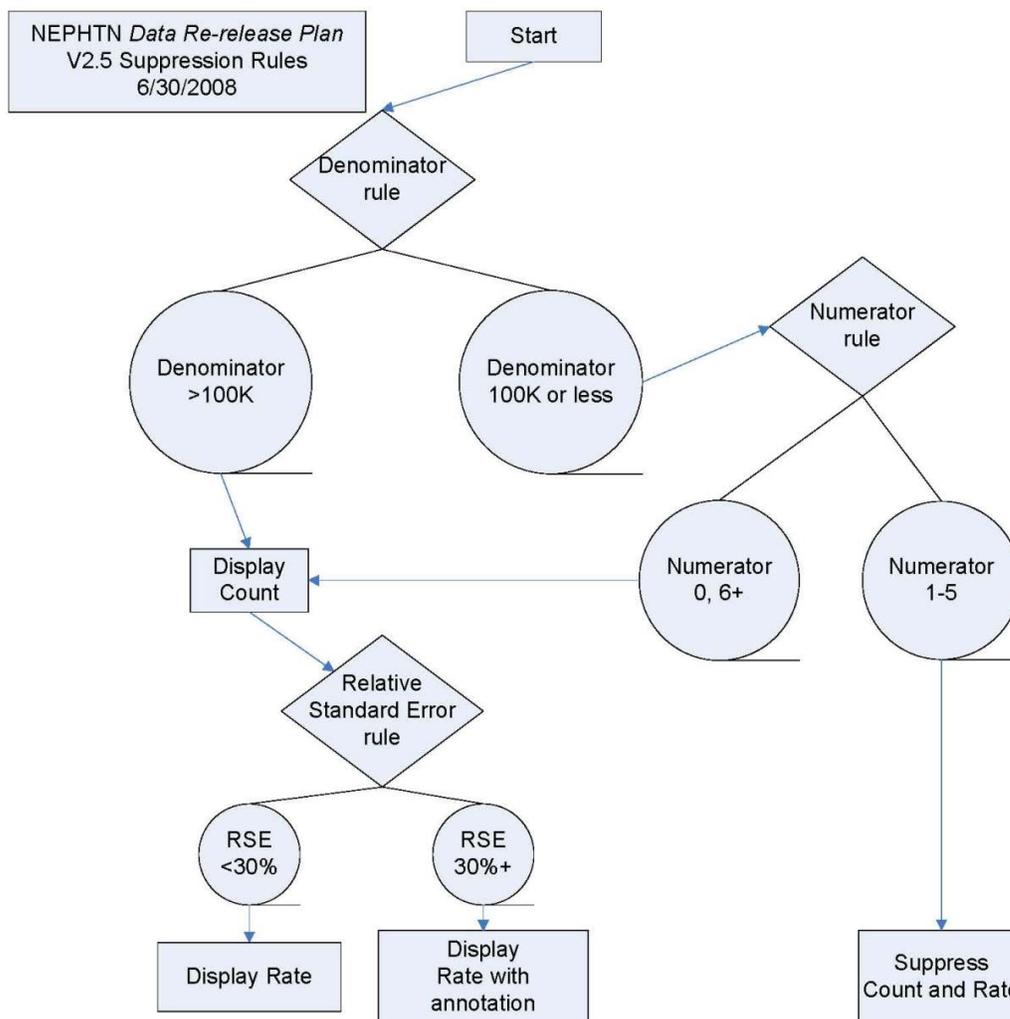
Cancer registry data: [RCW 70.54.250](#) and [WAC 246-102-070](#)

Appendix 1

Rule-based use of suppression and aggregation

The Washington Tracking Network (WTN) has an online data query system which displays data in tables, charts and maps, accessible by the public. In order to avoid automated production of tables where most rows are suppressed due to small numbers, WTN supplements its suppression rules with aggregation rules. The purpose is to aggregate data using static and dynamic parameter control in order to minimize suppression.

WTN follows the suppression rules developed by the CDC National Environmental Public Health Tracking Network (NEPHTN 2008). This flowchart depicts how these rules operate.



These rules apply to each row in every table, suppressing non-zero cells where the count is less than six, unless the denominator is greater than 100,000. The rules also result in annotating data where the RSE is 30% or higher.

When a health event is relatively rare, application of these rules can result in tables with many rows of suppressed data. Users find these tables to be extremely frustrating. Small subpopulations invariably lead to small numbers. Aggregation yields larger numbers, although stratification is needed to focus analysis, so a balance is desirable.

Fields in a dataset are commonly termed “parameters” in the context of data query systems. Parameter control can be achieved through use of static methods (within a parameter) or dynamic methods (between parameters). Dynamic parameter control is also termed “adaptive stratification.” Optimal parameter control includes protocol-driven use of both static and dynamic methods.

With static parameter control, some strata can be blocked by design, limiting tables to those based on greater aggregation. Examples are: displaying only multi-year data, not annual data; or, displaying only multi-county data, not county-level data. Parameters can also be excluded entirely, as when a dataset field is not relevant to program planning or evaluation. The static parameter control design rules should be reviewed with data stewards and program partners, who may want to make refinements. The key basis for the application of static parameter control design rules is program/planning utility.

The story of the asthma data online query system developed jointly by American Lung Association of Washington (ALAW) and the Washington State Department of Health (department) in the early 2000s is illustrative. The data shared by the department with ALAW for the query system potentially could have contained very tiny numbers, if stratified by age and gender simultaneously. The department proposed to share only one of these fields, but not both. ALAW members and department asthma program staff decided that, because intervention and prevention programs differ by age (there are programs for children and programs for adults), but not by sex, they wanted to see age strata in the data tables. The department excluded the gender parameter.

WTN rules for static parameter control start with count-based thresholds for Stratum Exclusion:

Spatial

- if <200 cases/year, then only multi-county regions available (no single county display)

Temporal

- if <400 cases/year, then only 5-year rollup available (no single year or 3-year rollup)
- if <800 cases/year but 400+ cases/year, then only 3-year rollup available (no single year)

Consultation with data stewards and program partners has often modified these rules. For example, in order to display annual data, greater spatial aggregation can be used. Once these rules are decided upon, they become static.

With dynamic parameter control, disaggregation is dependent on interactive query choices; in other words, adaptive stratification is interdependent, conditional on whether other parameters are aggregated. With small numbers, we want more aggregation; with larger numbers, we want less aggregation. WTN separates various topic areas in differing levels for adaptive stratification, termed AS Levels.

- With an AS1 (very small numbers), only one stratification parameter is available at a time; for example, if user selects disaggregation by geography, then the remainder of parameters are fully aggregated.

- With an AS3 (mid-range numbers), three stratification parameters are available at a time; for example, if user selects disaggregation by geography, time and gender (e.g., annual county-level by gender), then the remainder of parameters are fully aggregated.
- With an AS5 (large numbers), five stratification parameters are available at a time; for example, if user selects disaggregation by geography, time, age group, gender and race (e.g., annual county-level by age, race and gender), then the remainder of parameters are fully aggregated.

The WTN thresholds for Adaptive Stratification are:

- AS1 = < 100 cases per year statewide
- AS2 = 100-499 cases per year statewide
- AS3 = 500-999 cases per year statewide
- AS4 = 1000-4999 cases per year statewide
- AS5 = 5000-99,999 cases per year statewide
- AS6 = 100,000+ cases per year statewide

This WTN practice is a rule-based protocol. Thresholds between adjacent levels of Adaptive Stratification are independent of topic area (i.e., standardized across all topic areas).