Department of Health Agency Standards for Reporting Data with Small Numbers

Revision Date: May 2018

Primary Contact: Cathy Wasserman, PhD, MPH, State Epidemiologist for Non-Infectious Conditions
Secondary Contact: Eric Ossiander, PhD

Purpose

The Assessment Operations Group in the Washington State Department of Health (department) develops standards and 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 standards and 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?

This document describes recently adopted department standards for presentation of static and interactive query-based tabular data. The standards differ from the previous guidelines in that they represent minimum requirements that department staff must implement. This document also discusses statistical accuracy and makes recommendations for addressing statistical reliability. Unlike the standards, the recommendations are not mandatory. The department has a policy governing the sharing of confidential information both within and external to the department, Policy 17.006. (Link accessible to department employees only). This policy was revised in 2017 and now incorporates these standards for data reporting.

Scope of the “Standards 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 public users to build their own tables have become more common. These standards must be used by department staff who release department population-based or survey data in aggregated form available to the public. These releases include both static data tables and graphics, such as charts and maps, as well as tables and graphics produced through interactive query systems. In addition to these standards, analysts need to 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 standards provided in this document.** As specified in data sharing agreements, these standards also apply to non-departmental data analysts who receive record-level department data for rerelease in aggregated form to the public. In rare circumstances, such as with the Healthy Youth Survey, the department shares record-level data collected in partnership with other entities for rerelease in aggregated form. In these instances, other standards might apply.

The department and local health jurisdictions also release files containing record-level data. These standards do not apply to release of record-level data to the public. Release of record-level data is governed by federal and state disclosure laws, which can be specific to a dataset, as well as by Institutional Review Boards if the data are used for research.

Summary

Small Numbers Standards

Population Data: Department staff who are preparing confidential data for public presentation must:

1. Suppress all non-zero counts which are less than ten, unless they are in a category labeled "unknown."
2. Suppress rates or proportions derived from those suppressed counts.
3. Use secondary suppression as needed to assure that suppressed cells cannot be recalculated through subtraction.
4. When possible, aggregate data to minimize the need for suppression.
5. Individuals at the high or low end of a distribution (e.g., people with extremely high incomes, very old individuals, or people with extremely high body mass indexes) might be more identifiable than those in the middle. If needed, analysts need to top- or bottom-code the highest and lowest categories within a distribution to protect confidentiality. (See Glossary.)

Survey Data: Department staff preparing data for public presentation must:

1. Treat surveys in which 80% or more of the eligible population is surveyed as population data, as described above.
2. Treat surveys in which less than 80% of the eligible population is surveyed as follows:
    a. If the respondents are equally weighted, then cells with 1–9 respondents must be suppressed and top- and bottom-coding need to be considered.
    b. If the respondents are unequally weighted, so that cell sample sizes cannot be directly calculated from the weighted survey estimates, then there is no suppression requirement for the weighted survey estimates, but top- and bottom-coding might still be needed to protect confidentiality.

Exceptions to these standards include release of:

- Annual statewide, county or multiple county counts, or rates or proportions based on 1–9 events with no further stratification.
- Facility- or provider-specific data to facility personnel or providers for the purpose of quality improvement.

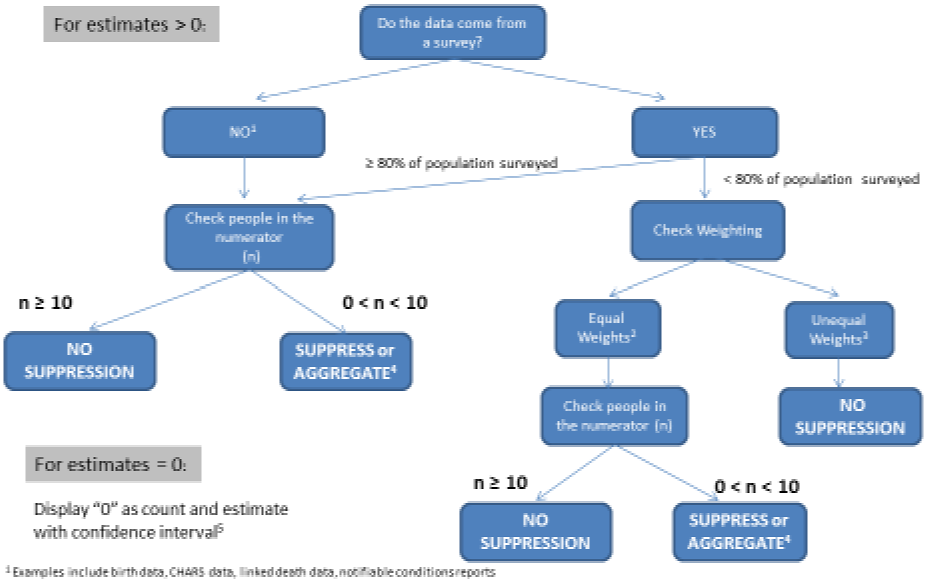
With approval from the Office of the State Health Officer, additional case-by-case exceptions to the suppression rule can be made, so that the public may receive information when public concern is elevated, protective actions are warranted or both.

Reliability Recommendations

- Include notation indicating rate instability when the relative standard error (RSE) of the rate or proportion is 25% or higher, but less than an upper limit established by the program. Suppress rates and proportions with RSEs greater than the upper limit; include notation to indicate suppression due to rate instability.
- Minimize the amount of unstable and suppressed data by further aggregating data, such as by combining multiple years or collapsing across categories.
- Include confidence intervals to indicate the stability of the estimate.

The standards and reliability recommendations are concisely represented in the following diagram which is downloadable as a separate pdf.

DOH Data Presentation for the Public – Small Numbers Standard



For estimates > 0:

Do the data come from a survey?

NO[1]

YES

≥ 80% of population surveyed

< 80% of population surveyed

Check people in the numerator (n)

Check Weighting

n ≥ 10

0 < n < 10

NO SUPPRESSION

SUPPRESS or AGGREGATE[4]

Equal Weights[2]

Unequal Weights[3]

Check people in the numerator (n)

NO SUPPRESSION

n ≥ 10

0 < n < 10

NO SUPPRESSION

SUPPRESS or AGGREGATE[4]

For estimates = 0:

Display "0" as count and estimate with confidence interval[5]

[1] Examples include birth data, CHARS data, linked death data, notifiable conditions reports

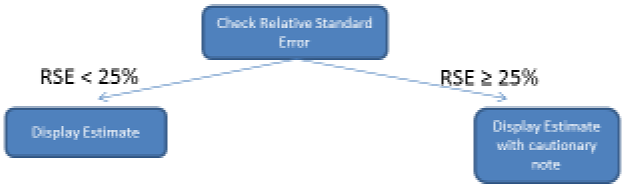
[2] Examples include Healthy Youth Survey

[3] Examples include Behavioral Risk Factor Surveillance System, Pregnancy Risk Assessment System

[4] Exceptions include annual state- or county-specific counts or rates with no stratification.

[5] 95% Poisson confidence interval for 0 is 0 to 3/n.

DOH Data Presentation for the Public – Reliability Recommendation



Check Relative Standard Error

RSE < 25%

RSE ≥ 25%

Display Estimate

Display Estimate with cautionary note

Calculation of the RSE

Depending on whether the data follow a Poisson or Binomial statistical distribution, methods for calculation of the RSE differ.

When data follow a Poisson distribution, the RSE is calculated as follows:
- A = count of events
- B = population
- Rate = A/B
- SE = Standard Error = SE of the rate = $\sqrt{(rate(1-rate))/population}$ = $\sqrt{A/B}$
- Percent RSE = 100(SE/rate) which simplifies to 100($\sqrt{A}$/A.
- Note that counts of 16 or less will have RSE > 25%

When data follow a Binomial distribution, the RSE is calculated as follows:
- A = numerator
- B = denominator
- Proportion = A/B
- SE = Standard Error = $\sqrt{(proportion(1-proportion))/B}$
- Percent RSE = (SE/Proportion)100

Background

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 which are at the highest risk of developing adverse health outcomes. 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 protect privacy and confidentiality. Small numbers also raise statistical issues concerning accuracy, and thus usefulness, of the data.

What constitutes a breach of confidentiality?

Department policy 17.005 defines a confidentiality breach as a loss or unauthorized access, use or disclosure of confidential information. (Link accessible to department staff only.) In the context of this document, 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). In data tables, a breach of confidentiality can occur if knowing which category a person falls in on one margin (i.e. row or column) of the table allows a table reader to ascertain which category the person falls in on the other margin. The section "Working with Small Numbers" below describes situations that present high risk for a breach of confidentiality and how to reduce this risk.

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

Estimates based on a 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 as 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, subpopulations 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 a new standard?

Our adoption of a standard requiring the suppression of cells reporting between 1 and 9 events is primarily based on the practice of the federal Centers for Disease Control and Prevention (CDC) National Center for Health Statistics (NCHS). NCHS requires that all data originating from NCHS and released by CDC (such as in tables produced by online query systems WONDER <http://wonder.cdc.gov/> and WISQARS <http://www.cdc.gov/injury/wisqars/fatal_injury_reports.html>) suppress counts that are less than 10, as well as rates and proportions based on counts less than 10. NCHS adopted this standard in 2011 after finding that a previous rule of suppressing cell counts between 1 and 4 failed to prevent disclosure of an individual's information. Instructions in Section 9 of the Centers for Medicare and Medicaid Services' (CMS) data use agreement specify the same suppression rule: no cell (and no statistic based on a cell of) 10 or less may be displayed (https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/Data-Use-Agreement.pdf). In contrast to these standards, the department standard allows release of tabular data where the count is zero, on the basis that a count of no events is, in many circumstances, unlikely to be a threat to confidentiality. However, data analysts need to be aware of the potential for group identification when zero counts for one category result in identifying all of the members of the group with a sensitive characteristic. For example, with Healthy Youth Survey data for a specific school, a count of 0 for no drug use would indicate that all students used drugs, breaching their trust that their responses would be kept confidential.

It is impossible to absolutely guarantee against disclosure risk when releasing data, because it is impossible to know how much outside information is available to the data user. Data users may have information from personal knowledge of people in the population from which the data were drawn, from searching for information on the Internet, or from other tables of similar data released by different agencies, or by the same agency at different times. Additionally, we cannot always anticipate or analyze all of the data tables that will be released.

Here we illustrate disclosure risk with an example from birth data. These are real Washington State data, but to prevent disclosure of sensitive data we have changed the county names and ZIP Codes.

> ZIP Code 47863 overlaps counties A and B. In 2005, there were 82 births to mothers whose resident ZIP Code was 47863; 81 of those mothers lived in County B, and 1 lived in County A. For the sake of this example, we pretend that no other ZIP Codes overlap the two counties. Let's say that one agency has provided, or posted on the Internet, a table that shows the number of prior pregnancies for birth mothers by resident ZIP Code, and another agency has provided or posted the same data by county of residence. By adding up the births for all ZIP Codes in County B, including 47863, a data user could ascertain that there was only 1 birth to a mother from County A who lived in ZIP Code 47863. If the data user happened to know this woman (say, as a neighbor), then the data user would know the number of her prior pregnancies. We can guard against this type of disclosure by suppressing some cells. In 2005, some of the ZIP Codes in County B had fewer than 10 births, and a rule requiring suppression of those numbers would make it harder for the data user to figure out how many births were in the overlap area. Appendix 1 provides a detailed explanation of this example and the effects of suppressing counts of 1-4 and 1-9.

Although we cannot guarantee that a rule requiring the suppression of counts between 1 and 4 will lead to disclosure of sensitive data, or that a rule requiring suppression of counts between 1 and 9 will prevent it, it is clear that the 1-9 rule will make disclosure substantially less likely. Additionally, data analysts should be aware of the considerations and approaches described below so they can minimize the risk of a breach of confidentiality despite adhering to the minimum standards. Some programs may need to adopt more stringent rules as program-specific standard practice. If the program needs to request an exception to the agency standard, the issues described below should be considered and addressed in the exception request. Protecting confidentiality starts with understanding the considerations that have gone into developing the standards, which are discussed below.

Working with Small Numbers

General Considerations

These standards and recommendations address both confidentiality and statistical issues in working with small numbers. In some data systems, such as the HIV/AIDS data system, the entire database is considered restricted confidential information (Category 4 data - link accessible to department employees only). 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 (such as when there are a small numbers of individuals with a visible characteristic in a small geographical area). If the datasets you are working with contain confidential or potentially identifiable information, the following sections on protecting confidentiality are relevant. Otherwise, only the sections on statistical issues are relevant.

Assessing Confidentiality Issues

Risk of disclosure depends almost entirely on the size of the numerator, as inferred from papers in the conference proceedings of a UNESCO-sponsored conference in 2014 (Domingo-Ferrer, Ed. 2014). Even in large populations it is conceivable that a single individual might be identifiable if there are few individuals with some special characteristic. For example, independent of the size of the community, if some residents of a community know of a child who is frequently hospitalized and an agency publishes a table showing that the community has one pediatric hospitalization and it is for pediatric HIV-AIDS, this table 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.

Know the identifiers. Data analysts should assess each field in the dataset to determine whether it is a "direct identifier" or an "indirect identifier". These terms are admittedly somewhat imprecise and can vary by dataset. Direct identifiers uniquely identify a person. Thus, direct identifiers are never publicly released and except in rare circumstances (for example, when license numbers are assigned sequentially such that a number can be used to estimate the length of time a provider has practiced) are not applicable to aggregated data. Indirect identifiers refer to group identity and are commonly presented when reporting aggregated public health data. Several examples of direct and indirect identifiers follow.

The federal Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (section 164.514(e)) (National Institutes of Health 2004) defines direct identifiers as:

- Name
- Street name or street address or post office box
- Telephone and fax numbers
- Email address
- Social security number
- Certificate/license numbers
- Vehicle identifiers and serial numbers
- URLs and IP addresses
- Full-face photos and other comparable images
- Medical record numbers, health plan beneficiary numbers, and other account numbers
- Device identifiers and serial numbers
- Biometric identifiers, including finger and voice prints

Indirect identifiers are fields which, when combined with other information, can be used to uniquely identify a person. Examples include:

- Detailed demographic information (e.g., age, gender, race, ethnicity)
- Detailed geographic information (e.g., census tract of residence, 5-digit ZIP code)

- Hospital name or location
- Detailed employment information (e.g., occupational title)
- Exact date of event (e.g., birth, death, hospital discharge)

WAC 246-455 defines direct and indirect identifiers for Comprehensive Hospital Abstract Reporting System (CHARS) data. In this case, direct identifiers include:

- Patient first name
- Patient middle name(s)
- Patient last name
- Social security number
- Patient control number or medical record number
- Patient zip code + 4 digits
- Dates that include day, month and year
- Admission and discharge dates in combination

The WAC defines indirect identifiers as information that may identify a patient when combined with other information. Indirect identifiers include:

- Hospital or provider identifiers
- 5-digit ZIP code
- County, state and country of residence
- Dates that include month and year
- Admission and discharge hour
- Secondary diagnosis, procedure, present on admission, external cause of injury, and payer codes
- Age in years
- Race and ethnicity

Datasets can be linked using only indirect identifiers (Hammill and colleagues, 2009; Pasquali and colleagues, 2010; Lawson and colleagues, 2013). Although aggregated data presented in tabular format are unlikely to be used in this fashion and the data standards outlined in this document are designed to minimize risk, no standard can absolutely guarantee against disclosure risk. Thus, to avoid presenting data that risk a breach of confidentiality, analysts should examine each field for its potential to allow users to identify a person.

Examine numerator size for each cell. Data analysts should consider the number of events in each cell of a table to be released and numerators when the data released are rates or proportions. 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.

Examples of CDC standards include the:

- 2004 NCHS Staff Manual on Confidentiality (NCHS 2004) that requires:
  - No single cells containing all observations of a row or column.
  - At least five[1] observations for a row or column total in a cross-tabulation.
  - At least five[1] observations total.
- Interactive query system, WONDER. Since May 2011 WONDER has suppressed birth and death data if there are not at least 10 observations (WONDER 2012).
- Environmental Public Health Tracking Network currently suppresses rates based on non-zero counts less than six, (EPHTN 2008) unless the data originate from a program with stricter suppression rules. For example, Environmental Public Health Tracking Network suppresses mortality data with fewer than 10 observations consistent with NCHS standards.

The department standards require suppression when the number of cases or events in a cell is less than 10 to reduce 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. A data analyst may choose a higher threshold, if other information indicates a greater likelihood of a possible breach of confidentiality in a specific situation.

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. Surveys that include 80% or more of the eligible population should be treated in the same way as population data. Surveys of facilities or surveys conducted within facilities, such as schools, sometimes fall into this category. If the survey includes less than 80% of the eligible population, and if the identity of the respondents is kept private, then the risk of disclosing identifying information is far lower than for population data, particularly if weighted survey estimates are presented, instead of respondent cell sizes.

Consider the nature of the information. The U.S. Census bureau uses the *Checklist on Disclosure Potential of Data* that identifies examples of variables that are visible and, therefore, pose increased risk of disclosure (U.S. Census 2013). 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 Meet the Standard 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 those developed by the Washington Tracking Network (WTN) and the Washington State Cancer Registry, aggregate data using rule-based static control, dynamic parameter control or both in order to minimize suppression. Appendix 2 outlines the aggregation 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

[1] In 2011, NCHS changed its standard to at least 10 observations, but has not reissued its Staff Manual on Confidentiality.

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**	*Fine*	*Medium*	*Coarse*
Age	Continuous	Year of birth	5-year age group	10-year age group
Date of occurrence	Continuous	Month and year	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.

Data analysts also need to consider whether individuals in extreme categories, such as extremely high income, high body mass index or very old age, are identifiable. For example, in a table presenting body mass index (BMI) by another health outcome, even 10 people in the group with the highest BMI might identifiable. In these instances, top- and bottom-coding need to be considered as a special case of aggregation. In the example of BMI by a health condition, the analyst might truncate all categories greater than 40 kg/m^2 to a single category of greater than 40 kg/m^2. Similarly, HIPAA specifies that all ages 90 and older must be aggregated into a top-coded category of 90 and older.

Cell suppression. When it is not possible, or desirable, to create a table with no small numbers, 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 (also termed “complementary”) suppression of three other cells in order to avoid inadvertent disclosure through subtraction. Secondary cell suppression is a method of last resort, due to the often unavoidable side-effect of suppressing releasable data values, and due to the amount of labor necessary to implement the method. The following table shows an example of secondary 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 to prevent the ability for back-calculating the suppressed cell.

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. The software program tau-ARGUS uses mathematical algorithms to perform secondary suppression in a way that assures that the suppressed data cannot be uncovered by back calculation. (tau-ARGUS 2014) However, tau-ARGUS may be difficult to use. If data are suppressed, the data analyst should provide an indicator (e.g., asterisk) in the suppressed cell and a legend under the table explaining the reason for suppression.

Omission of stratification variables. 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 numbers 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 many intervention programs target age groups differently.

Exceptions to the Small Numbers Standard

The following small numbers are allowed to be reported on a regular basis:

- Statewide, county or multiple county counts and rates or proportions based on counts 1-9 for an entire year without additional stratification.
- Facility or provider-specific data to be used for quality improvement purposes. Such information may be prepared and shared with data reporters (i.e., providers or other personnel at the facility). These data should not be posted on the department website. Programs should consider that once produced, these data may be subject to public disclosure requests.

In addition, the agency standards allow for case-by-case exceptions, with advance approval from senior management. To request an exception, the data analyst emails the State Epidemiologist for Infectious Disease (Scott.Lindquist@doh.wa.gov) and the State Epidemiologist for Non-Infectious Conditions (Cathy.Wasserman@doh.wa.gov) with the subject line: Small Numbers Exception Request. The email must contain the following information:

- Brief description of the health data that are being released.
- Identifiers by which the data are stratified.
- Rationale for the exception, including why aggregation is not an acceptable approach.
- The value of the numbers (or numerators for rates) that will be released (e.g. counts of 6-9 events).
- Why releasing counts (or rates based on numerators less than 10) will not compromise confidentiality.

The maximum response time for planned periodic reporting, such as annual data reports, will be 10 business days. In a public health emergency, such as described below, for department employees the response time will be one to three days commensurate with the urgency. Local Health Officers will determine release of data in county-specific emergency situations.

Two examples of situations when an exception would likely be approved are:

- In a cluster investigation, intense public interest often combines with very small numbers of cases. In order to be responsive to the community and allay fear, the department may decide it is important to make an exception to the standard while still protecting privacy.
- Similarly, in a public health emergency such as a communicable disease outbreak or other all-hazards incident, case counts may be released when the numbers are very small. This should be done in the context of an imminent public health threat, such as person-to-person spread of disease, where immediate action is indicated to protect public health.

When releasing small numbers to the public in the context of the above exceptions, the department recommends:

- Limiting the amount of information released in order to protect the identity of the person(s) involved.
- Reporting at most the person's gender, decade of age and county of residence. For minors, ages should be reported as <18 unless there is a compelling public health rationale for a different aggregation of ages.

Considerations for Implementing Suppression Rules that Exceed the Standards. There are some situations in which complying with the standard might not sufficiently protect confidentiality. For example, in a small school with high participation in the Health Youth Survey, a zero count in a cell such as "did not use alcohol in the past 30 days" provides meaningful information about the students who took the survey with the understanding that their answers were confidential. **Data analysts and programs are responsible for assessing data for potential breaches of confidentiality even when complying with the standard.**

Situations that require particular attention to avoid breaches of confidentiality even when complying with the standards include:

- Denominators less than 20,000. Although the risk of disclosure depends primarily on the size of the numerator, most governmental groups responsible for maintaining data confidentiality place constraints on the size of the denominator, as well. For example, The National Center for Health Statistics' Staff Manual on Confidentiality prohibits releasing record-level data for geographic areas with fewer than 100,000 people, effectively limiting tabular data to geographies with 100,000 people or more (NCHS 2004). HIPAA allows sharing of record-level data for 3-digit ZIP code areas containing at least 20,000 people if several other conditions are met (e.g., suppression of all elements of dates except for year, and grouping single years of age for people over 89 years into a single category). Thus, HIPAA effectively limits release of aggregated data to areas with more than 20,000 people (National Institutes of Health 2004).

  NOTE: The department routinely publishes data by county. Based on the Washington State Office of Financial Management's April 1, 2017 population estimates, nine counties had populations less than 20,000; three of those had populations less than 20,000 person-years when combining three years of data (i.e., 2015–2017). Even though some counties do not meet a 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 should carefully evaluate the potential for breaches of confidentiality when publishing data with denominators of subpopulations less than 20,000. 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 data that meet standards for counts and numerators. 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 subpopulations.
- Counts less than 20.
- Reporting a specific confidential characteristic of a population if a very high proportion of the population has this characteristic. This is called "group identification." Data in a table provide 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" and describes the problem as follows:

  > Data in a table may indicate that members of a given population segment have an 80-percent chance of having a certain characteristic; this would be a probability-based disclosure as opposed to a certainty disclosure of information on given individuals. In a sense, every published table containing data or estimates of descriptors of a specific population group provides probability-based disclosures on members of that group, and only in unusual circumstances could any such disclosure be considered unacceptable. It is possible that a situation could arise in which data intended for publication would reveal that a highly specific

> group had an extremely high probability of having a given sensitive characteristic; in such a case the probability-based disclosure perhaps should not be published. (NCHS 2004, p. 15)

- Producing multiple tables from the same dataset; in this case, be careful that users cannot derive confidential information through a process of subtraction.

Assessing and Addressing Statistical Issues

The following recommendations offer approaches to decrease the likelihood that some data users might misinterpret data that are unstable due to small numbers. The recommendations are based on practices followed by many units within the Centers for Disease Control and Prevention. The department's Assessment Operations Group considers the general approaches as best practices when releasing aggregated data to the public. However, unlike the mandatory standards outlined above, data analysts and programs can decide when and how to implement these recommendations.

What is the relative standard error? The relative standard error (RSE) provides a measure of reliability (also termed "statistical stability") for statistical estimates. When the RSE is large, the estimate is imprecise and we term such rates or proportions "unstable" or "not reliable." In these instances, the data analyst needs to balance issues of the right-to-k now with presenting data that might be misleading.

There is no single national standard for deciding when the RSE is large enough to need annotation or so large that one should suppress the data. Federal agencies and even units within a single federal agency use different approaches. For example, within the Centers for Disease Control and Prevention:

- The National Center for Health Statistics (NCHS) May 2017 publication "Health, United States, 2016" annotates survey data with RSEs of 20–30% and suppresses data with RSEs greater than 30% (or cells with fewer than 100 survey respondents) (NCHS 2017).
- The National Program of Cancer Registries (NPCR) 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 2014).
- Cells with denominators less than 50 or RSEs greater than 30% are suppressed in CDC's 2015 online publication "2014 BRFSS Asthma Call-Back Survey Prevalence Tables." There is no cut point for marking data as unreliable, but confidence intervals are provided around all estimates so that a reader can note the potential range of estimate variability (CDC 2015).
- The National Health Interview Survey's January 2014 online tables, "Health insurance coverage by coverage status, type, selected characteristics and age, January-June 2013," suppress data when RSEs are greater than 50% and note that estimates with RSEs of 30–50% are unreliable, as in Table 2 with data for persons aged 0–18 (NHIS 2014).
- A 2011 NCHS publication suppressed data with fewer than 30 cases due to lack of statistical reliability, and marked estimates based on 30–59 cases, or more than 59 cases with RSEs greater than 30%, as unreliable (Bercovitz 2011).
- CDC Environmental Public Health Tracking Network national portal continues to use its 2008 standard of displaying all rates that are not suppressed for confidentiality protection (i.e. cells with 6 or more observations or the minimum number of observations required by the data provider, such as 10 for NCHS data). Rates with RSEs of 30% or greater are annotated as unreliable. Although not explicitly stated, rates based on fewer than 6 observations have RSE greater than about 45% (NEPHTN 2008).

As with CDC, different programs at the department use different practices for suppressing or annotating data due to concerns over statistical instability and the concomitant potential for misinterpretation of data. 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 because of statistical instability.

How do I calculate the RSE? Depending on whether the data follow a Poisson or a Binomial statistical distribution, methods for calculating standard errors (SE), and hence RSEs, differ.

When data follow a Poisson distribution, the percent RSE is calculated as follows. Note that the Poisson-based calculation of RSE does not use the population.

- Notation:
  - $\text{A} = \text{count of events}$
  - $\text{B} = \text{population}$
  - $\text{Rate} = \text{A}/\text{B}$
  - $\text{SE of the rate} = \sqrt{(\text{rate}(1 - \text{rate}))/\text{population}} = \sqrt{\text{A}/\text{B}}$
- $\text{Percent RSE} = 100(\text{SE}/\text{rate})$ which simplifies to $100(\sqrt{\text{A}}/\text{A}.$

A simplified method can be used: any result of a rate calculation where the count of events is less than 17 is not reliable, because rate calculations where the count of events is 16 or less result in RSEs higher than 25%.

When data follow the Binomial distribution, the RSE is calculated as follows:

- Notation:
  - $\text{A} = \text{numerator}$
  - $\text{B} = \text{denominator}$
  - $\text{Proportion} = \text{A}/\text{B}$
  - $\text{SE} = \text{Standard Error} = \sqrt{(\text{proportion}(1 - \text{proportion}))/B}$
- $\text{Percent RSE} = (\text{SE}/\text{Proportion})100$

The simplified method for identifying proportions that are not reliable is accurate for the binomial distribution when the denominator is more than 1000. When the denominator is smaller, the simplified method results in more proportions being labeled as not reliable than if the full RSE calculation was used. Thus, the simplified method is conservative: it over-annotates some results as not reliable, when the numerator and denominator numbers are small.

Recommendations to address statistical issues include annotating or suppressing data based on the RSE and including confidence intervals.

- ***Use a notation to annotate estimates when RSEs are greater than 25% and less than an upper limit established by the program***. Rates or proportion with RSEs greater than the upper cut point should be suppressed. For Poisson distributions, this recommendation simplifies to annotation with counts of 16 or less (see section on using the Poisson distribution to calculate RSEs below). Given the requirement to suppress data with 9 observations or less, for Poisson distributions, suppression would occur with RSEs greater than 33%. The flag can be an asterisk or other symbol or the designation “NR” for “not reliable” next to the rate or proportion in the table. Suppression could be indicated by an “NA” for “not available.” A footnote should explain the notation. Maps can use similar annotation if rates are displayed or maps can indicate estimates with wide variability and suppression using coloring or shading, such as diagonal hatched shading for “not reliable” and white for “not available" and a legend explaining the meaning of the color or shading.

- ***Reduce the amount of annotated and suppressed data due to instability of the estimate.*** As the proportion of data suppressed or annotated as unreliable increases, the value of the data table decreases. Increasing the numerator 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 (CIs) when presenting rates and proportions.*** (See Guidelines for Using Confidence Intervals for Public Health Assessment.) CIs give users an understanding of the stability of an estimate independent of annotation. CIs can be displayed on tables in numeric form or visually on charts and line graphs. Online query systems might automatically display CIs or CIs might only be displayed when the user selects a “Display CI” button. A “hover-over pop-up” which uses a small window to separately display the rate or proportion with its CI for each data-point on an online chart is another possible method. CIs might be less important on line graphs, such as a graph that shows rates by year, because the year-to-year variation is visible from the line, itself. If there is no practical method for including CIs on maps, using shading to show estimates that are not reliable may need to suffice. One approach may be to vary the gap between cross hatching on a map, such that it narrows as the reliability decreases. In this instance, highly unreliable data will be essentially blacked out and the underlying color indicating the rate will not be visible. Although including confidence intervals implicitly shows the stability of the estimate, data analysts should consider annotation even when displaying confidence intervals, if some of the intended audience might not understand the meaning of confidence intervals.

The approach to annotating and suppressing data might vary depending on the primary audience and purpose of the publication. For example, when there is increased concern over statistical stability, a public health program may decide that program-specific practice will be to routinely annotate data with RSEs between 22% and 30% and suppress data with RSEs greater than 30%. For Poisson-based rates, this simplifies to annotation of rates based on counts between 11 and 20 and suppression of estimates based on 10 or fewer observations.

Note on bias. The issue of bias differs from the issue of the precision of estimates in that bias is *non-random* error. When the data analyst is aware that the count of persons or events is systematically over- or under-ascertained in the data, we recommend that the data user be informed by annotating the data as “not available” or “not reliable” due to bias. The data analyst must distinguish these annotations from those for statistical instability. For example, agency studies have shown that hospitalization rates for some border counties are subject to an undercount if the CHARS dataset is used without inclusion of Washington residents hospitalized out-of-state. The WTN metadata explains this in its Caveats section:

> Without reciprocal agreements with abutting states, statewide measures and measures for geographic areas (e.g., counties) bordering other states may be underestimated because of health care utilization patterns. The Tracking Network rules currently call for exclusion of hospitalization data obtained from adjacent states, regardless of whether a state has reciprocal agreements with the adjacent states. In eight Washington counties, hospitalization rates are biased due to county residents traveling out-of-state for hospital care. For these counties (Asotin, Clark, Cowlitz, Garfield, Klickitat, Pacific, Skamania, and Wahkiakum), more than 15% of hospitalizations of county residents occur outside of the state. The bias from this undercount is judged to be excessive, and WTN annotates these hospitalization rates with the NR designation.

Glossary

Bottom-coding: Bottom-coding of a variable places a lower limit for aggregation of a variable such that any value less than the lower limit is included in that category. For example, if a data analyst wanted to provide rates of heart attacks by 5-year age groups, the analyst might decide to aggregate all ages 30 and under into a bottom-coded category of 30 or younger to protect the confidentiality of the few relatively young people who have heart attacks.

Confidential data/information: For the purposes of these standards and recommendations confidential information includes all 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. For these standards and recommendations, confidential information includes, but is not limited to, information that is exempt from public disclosure as described in department policies 17.005 and 17.006. (Links accessible to department staff only)

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, phone 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).

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. Potentially identifiable information include, but is not limited to, indirect identifiers as described in statues and administrative codes.

Rate: A measure of the frequency of an event per population unit. Rates include an element of time (average speed while driving is a rate expressed as miles per hour, cancer mortality might be expressed as deaths per person-year at risk, etc.) We also often call indicators rates although we are actually referring to proportions (e.g. an attack rate is the proportion of people who develop disease after exposure to a pathogen; the smoking rate is the proportion of people surveyed who reported smoking.) These guidelines hold for both rates and proportions. (See Guidelines for Using and Developing Rates for Public Health Assessment.).

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. Sensitive personal information includes, but is not limited to, restricted confidential information defined in department policy 17.006. (Link accessible to department staff only.)

Top-Coding: Top-coding of a variable places an upper limit for aggregation of a variable such that any value greater than the upper limit is included in that category. For example, HIPAA specifies that categories of ages greater than 90 years not be published, but rather aggregated and recorded as 90 or older to prevent identification.

References

Bercovitz A, Moss A, Sengupta M, et al. An overview of home health aides: United States, 2007. *National Health Statistics Reports*. 2011 May; 34. http://www.cdc.gov/nchs/data/nhsr/nhsr034.pdf. Accessed May 23, 2017.

Centers for Disease Control and Prevention (CDC). 2014 BRFSS Asthma Call-back Survey Prevalence Tables. 2015. Available at https://www.cdc.gov/brfss/acbs/2014_tables.html. Accessed August 29, 2017. See, for example, Table 8b available at https://www.cdc.gov/brfss/acbs/2014/prevalence_tables/table8b.html , accessed August 29, 2017.

Domingo-Ferrer J, Ed. Privacy in Statistical Databases. UNESCO Chair in Data Privacy, International Conference, PSD 2014, Ibiza, Spain, September 17-19, 2014. Proceedings. Switzerland, Springer International Publishing. 2014; 24- 35, 36-47, 48-61, 62-78, 79-88.

Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. *Am Heart J*. 2009 Jun;157(6):995-1000. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732025/. Accessed May 23, 2017.

Lawson EH, Ko CY, Louie R, Han L, Rapp M, Zingmond DS. Linkage of a clinical surgical registry with Medicare inpatient claims data using indirect identifiers. *Surgery*. 2013 Mar;153(3):423-30.

National Center for Health Statistics. Health, United States, 2016 with Chartbook on Long-term Trends in Health. Hyattsville, MD. 2017. https://www.cdc.gov/nchs/hus/index.htm. Accessed July 19, 2017.

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 May 23, 2017.

National Health Interview Survey. Health insurance coverage status, coverage type, and selected characteristics, for persons of aged 0–18, January-June 2013. Centers for Disease Control and Prevention. (2014) https://www.cdc.gov/nchs/health_policy/health_insurance_selected_characteristics_jan_jun_2013.htm. Accessed August 28, 2017.

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 May 23, 2017.

National Center for Health Statistics. *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 May 23, 2017..

National Program Cancer Registries (NPCR). United States Cancer Statistics;Technical Notes; Suppression of Rates and Counts. Centers for Disease Control and Prevention, Division of Cancer Prevention and Control, 2014. http://www.cdc.gov/cancer/npcr/uscs/technical_notes/stat_methods/suppression.htm. Accessed May 23, 2017.

Pasquali SK, Jacobs JP, Shook GJ, et al. Linking clinical registry data with administrative data using indirect identifiers: implementation and validation in the congenital heart surgery population. *Am Heart J*. 2010 Dec;160(6):1099-104. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3011979/. Accessed May 23, 2017.

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

Tau-ARGUS. 2014. Version 4.1. Argus Open Source Project. Available at http://neon.vb.cbs.nl/casc/tau.htm. Accessed November 1, 2017.

U.S. Census Bureau. Checklist on Disclosure of Potential Data. 2013. Available at https://www.census.gov/srd/sdc/drbchecklist51313.docx. Accessed May 23, 2017.

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 May 23, 2017.

Resources

Klein RJ, Proctor SE, Boudreault 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 May 23, 2017..

Federal Committee on Statistical Methodology*.* Confidentiality and Data Access Committee (CDAC) Resources for Confidentiality and Data Access Information. Including *OMB Checklist on Disclosure Potential of Proposed Data Releases.* Washington, DC: Office of Management and Budget; 1999. https://fcsm.sites.usa.gov/committees/cdac/cdac-resources/ Accessed May 23, 2017.

Statistics Netherlands. Statistical Disclosure Control: τ-ARGUS home page. http://neon.vb.cbs.nl/casc/..%5Ccasc%5Ctau.htm. Accessed May 23, 2017.

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)

Responsibilities for Confidential Information: Department Policy 17.005 (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. (http://www.digitalarchives.wa.gov/GovernorGregoire/execorders/recpriv/recpriv2.htm)

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 several 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.

Cancer registry data: RCW 70.54.250 and WAC 246-102-070

Appendix 1: Detailed Example of Disclosure Risk

Here we illustrate disclosure risk with an example from birth data. These are real Washington State data, but we have changed the county names and ZIP Codes to prevent disclosure of sensitive data.

> ZIP Code 47863 overlaps counties A and B. In 2005, there were 82 births to mothers whose resident ZIP Code was 47863; 81 of those mothers lived in County B, and 1 lived in County A. For the sake of this example, we pretend that no other ZIP Codes overlap the two counties. Let's say that one agency has provided, or posted on the Internet, a table that shows the number of prior pregnancies for birth mothers by resident ZIP Code, and another agency has provided or posted the same data by county of residence. By adding up the figures for all ZIP Codes in County B, including 47863, a data user could ascertain that there was only 1 birth to a mother who lived in ZIP Code 47863 in County A. If the data user happened to know this woman (say, as a neighbor), then the data user would know the number of her prior pregnancies. We can guard against this type of disclosure by suppressing some cells. In 2005, some of the ZIP Codes in County B had fewer than 10 births, and a rule requiring suppression of those numbers would make it harder for the data user to figure out how many births were in the overlap area. A detailed explanation of the effects of suppressing counts of 1-4 or 1-9 is provided below.
>
> In practice, we cannot anticipate or analyze all of the data tables that will be released. We cannot guarantee either that a rule requiring only the suppression of counts between 1 and 4 will lead to disclosure of sensitive data, or that a rule requiring suppression of counts between 1 and 9 will prevent it. However, it is clear that the 1-9 rule will make disclosure substantially less likely.

First, we have a list of ZIP Codes by county, which shows that one ZIP Code (47863) lies in both counties A and B:

Table 1: ZIP Codes by County

County A	47863
County A	47864
County A	47865
County A	47866
County A	47867
County A	47868
County A	47869
County A	47870
County A	47872
County B	47863
County B	47873
County B	47883
County B	47884
County B	47885
County B	47886
County B	47887
County B	47888
County B	47889
County B	47890
County B	47892
County B	47893

County B	47894
County B	47895
County B	47896

Let's say that we have tables showing births by resident ZIP Code, and no data suppression (Table 2 [column 2]) and births by county of residence (Table 3). Since the sum of births in ZIP Codes that fall wholly or at least partially in County B (ZIPs 47873-47896 plus ZIP 47863) is 1,422, we can deduce that there is just one birth in those ZIP Codes that is not in County B (because the total for County B in Table 3 is 1,421), and therefore just one birth to a County A resident living in ZIP Code 47863. In any set of tables lacking any suppression that showed characteristics of births (such as the number of prior pregnancies) by resident ZIP Code and by county of residence, a data user could identify the characteristics of that single birth.

Table 2: Births by ZIP Code

ZIP Code	Births	Births (counts of 1-4 suppressed)	Births (counts of 1-9 suppressed)
47863	82	82	82
47864	1	*	*
47865	3	*	*
47866	34	34	34
47867	1	*	*
47868	2	*	*
47869	7	7	*
47870	398	398	398
47872	3	*	*
47873	148	148	148
47883	14	14	14
47884	596	596	596
47885	150	150	150
47886	43	43	43
47887	1	*	*
47888	3	*	*
47889	8	8	*
47890	9	9	*
47892	11	11	11
47893	2	*	*
47894	25	25	25
47895	229	229	229
47896	101	101	101

Table 3: Births by county of residence

County	Births
County A	450
County B	1421

Now let's say that we have suppressed the data in all cells having a count between 1 and 4 (see column 3 in Table 2). The sum of births in non-suppressed ZIP Codes that fall wholly or at least partially in County B is

1,416. A data user can see that the counts in the three ZIP Codes which are wholly in County B have been suppressed, and, knowing the suppression rule, can deduce that there were between 1,419 (i.e., the sum of 1,416 and 3, assuming 1 birth in each suppressed ZIP code) and 1,428 (1,416 plus 12, which assumes 4 births in each suppressed ZIP Code) births in ZIP Codes that fall wholly or at least partially in County B. Since there were 1,421 births to County B residents, the data user can deduce that there were 0 to 7 births to County A residents living in ZIP Code 47863.

The 3 County B ZIP Codes in which data were suppressed had 1, 3, and 2 births. Note that if, by happenstance, these ZIP Codes had all had 4 births, then the total number of births in County B would have been 1,427, and this total would have been shown in the births by county table. Then the data user, knowing that there were between 1,419 and 1,428 births in County B ZIP Codes, could deduce that there were 0 or 1 births in County A in Zip Code 47863. If the data user knew a County A mother who lived in ZIP 47863 and gave birth in 2005, then the data user would know that was the only such mother. Additionally, this suppression rule does not suppress counts of 0, so any combination of 0 or 4 births among those 3 ZIP Codes would have allowed the data user to reach that same conclusion.

Now let’s say that we have suppressed the data in all cells having a count between 1 and 9 (see fourth column in Table 2). The sum of births in non-suppressed ZIP Codes that fall wholly or at least partially in County B is 1,399. A data user can see that the counts in 5 ZIP Codes in County B have been suppressed, and, knowing the suppression rule, can deduce that there were between 1,404 (i.e., the sum of 1,399 and 5, assuming 1 birth in each suppressed ZIP Code) and 1,444 (1,399 plus 45, which assumes 9 births in each suppressed ZIP Code) births in ZIP Codes that fall wholly or at least partially in County B. Since there were 1,421 births to County B residents, the data user can deduce that there were 0 to 23 births to County A residents living in ZIP Code 47863. An alternative realization of these data that would allow a data user to identify an individual mother as the only mother in County A in ZIP Code 47863 would require each of these 5 ZIP Codes to have either 0 or 9 births. This would be far less likely to happen than the scenario above, which only required 3 ZIP Codes to have 0 or 4 births.

Appendix 2: *Washington Tracking Network Rule-Based Use of 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 goal is to aggregate data using static and dynamic parameter control in order to minimize suppression. As of August 2017, WTN uses static parameter control with dynamic parameter control still in the planning stages. Dynamic parameter control methods have been implemented in the Washington State Cancer Registry website (https://fortress.wa.gov/doh/wscr/WSCR/Query.mvc/Query), and on the national Tracking Network portal (http://ephtracking.cdc.gov/).

When a health event is relatively rare, application of suppression 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 (temporal aggregation); or, displaying only multi-county data, not county-level data (spatial aggregation). 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 separate 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)
- if <100 cases/year, then only state-level available (no multi-county regions or 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)

From	To	Temporal	Spatial
800	above	Single year	County
400	<800	3-year rollup	County
200	<400	5-year rollup	County
100	<200	5-year rollup	MCR
	<100	5-year rollup	statewide

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 into 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 a 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 a 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 a 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).