STANDARDS FOR WORKING WITH SMALL NUMBERS

Version 2c: March 2025

ARKANSAS DEPARTMENT OF HEALTH

Primary Contact

Namvar Zohoori, MD, MPH, PhD
Chief Science Officer
Namvar.Zohoori@Arkansas.gov

Secondary Contact

Austin Porter, DrPH, MPH
Deputy Chief Science Officer
Austin.Porter@Arkansas.gov

This document is based on information from several CDC, NCHS and other publications (please see Reference list), and a similar document from the Washington State Department of Health, to whom we are grateful.

Table of Contents

Introduction

Purpose

These standards and recommendations for the reporting of data with small numbers have been developed to promote good professional practice among staff involved in data analysis and reporting activities within the Arkansas Department of Health (ADH), and to protect confidentiality when protected health information is reported outside the agency. This document describes standards for presentation of static and interactive query-based tabular data. ***The standards represent minimum requirements that ADH staff must implement***. This document also discusses statistical accuracy and makes recommendations for addressing statistical reliability. Unlike the standards, the recommendations are not mandatory. While these standards and guidelines are intended for audiences of differing levels of training, they assume a basic knowledge of epidemiology and biostatistics.

Scope of the "Standards for Working with Small Numbers"

The ADH routinely makes 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 ADH staff who release Agency 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 Arkansas State laws and regulations and department policies. **Federal and state laws and regulations and ADH policies supersede standards provided in this document.** As specified in data sharing agreements, these standards also apply to non-ADH data analysts, including researchers, who receive record-level ADH data for rerelease in aggregated form to the public. In rare circumstances, 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 ADH also releases 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 and ADH Science Advisory Committee policies if the data are used for research.

Summary (At the very least, read this page!)

Small Numbers Standards (Figure 1)

Population Data: ADH 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" or one of the exceptions below are met.
2. Suppress rates or proportions derived from those suppressed counts.
3. Use secondary suppression as needed to ensure 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 indices) might be more identifiable than those in the middle. If needed, analysts need to top- or bottom-code (see Glossary) the highest and lowest categories within a distribution to protect confidentiality.

Survey Data: ADH 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:

1. Annual (12 months) statewide, county or combined multiple-county counts, or rates or proportions based on **1–9** events *with no further stratification* (for example, age, gender, race, month, etc.).
2. Statewide counts, or rates or proportions based on **1–9** events derived from at least three full years (36 months) of combined data, *with no temporal breakdown*.
3. Statewide counts, or rates or proportions based on **1–9** average monthly events, derived from at least **five** full years (60 months) of combined data, with *no other breakdowns.*
4. County-level counts, or rates or proportions based on **5–9** events derived from at least three full years (36 months) of combined data, *with no temporal breakdown*.
5. Facility- or provider-specific data to those facility personnel or providers for the purpose of quality improvement (see page 14 for details).

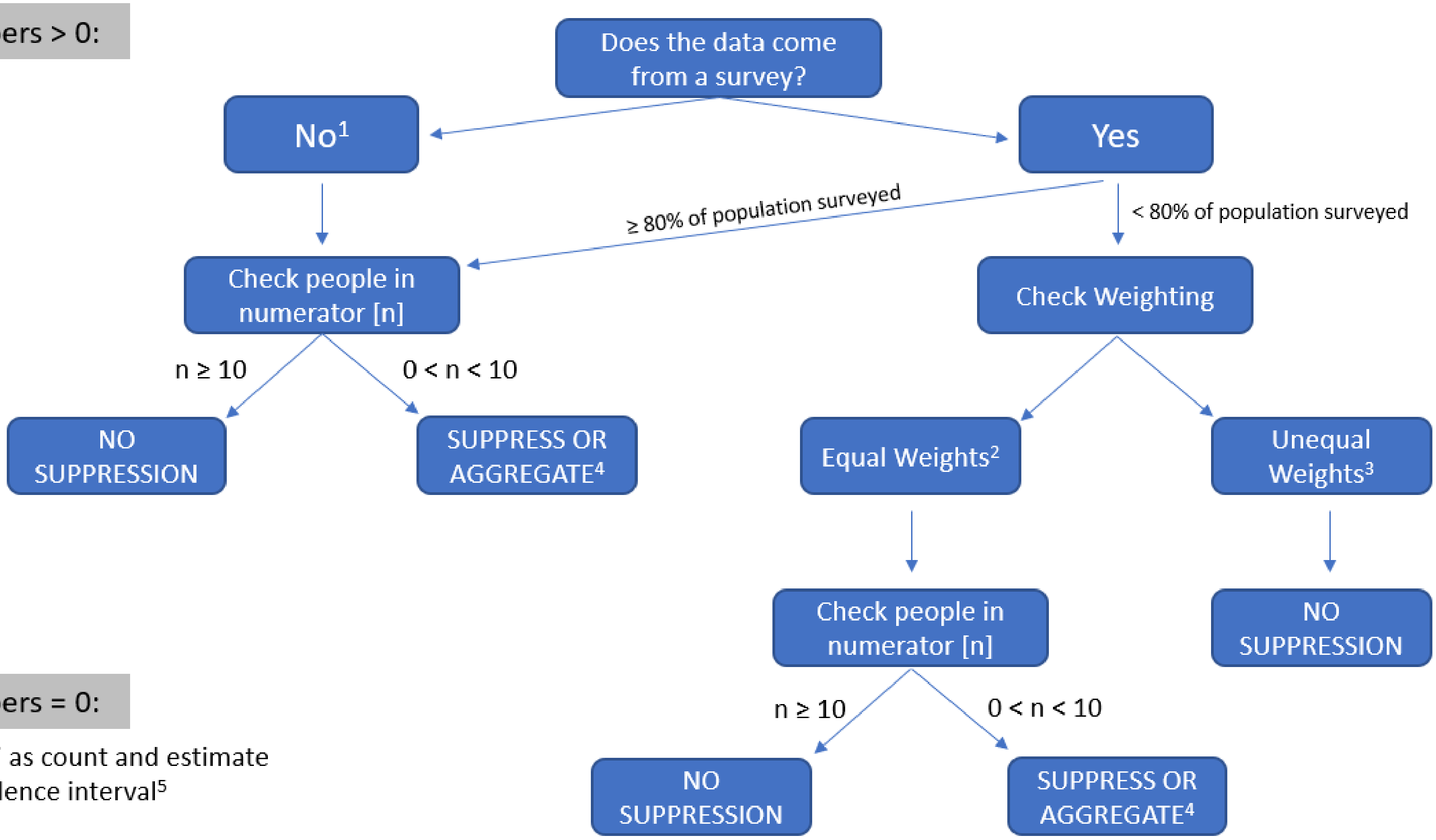
With approval from the Office of the Chief Science Officer (see page 14), additional case-by-case exceptions to the suppression rule can be made, so that the public may receive information when public concern is elevated, protective actions are warranted, or both.

Reliability Recommendations (Figure 2)

- Include notation indicating rate instability when the relative standard error (RSE) of the rate or proportion is 25% or higher. 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.

Figure 1: ADH Data Presentation for the Public – Small Numbers Standard

For numbers > 0:



For numbers = 0:

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

[1] Examples include birth and death data, hospital discharge data, notifiable conditions reports, etc.

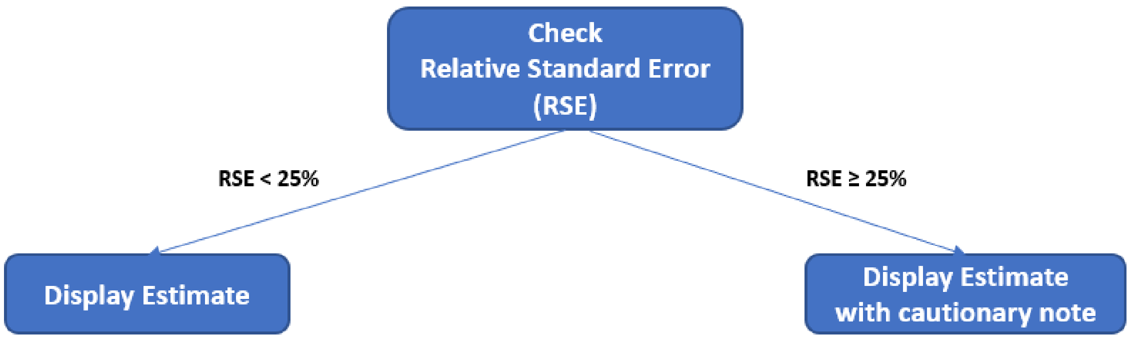
[2] Examples include 3rd Grade Basic Screening Survey

[3] Examples include Behavioral Risk Factor Surveillance System, Pregnancy Risk Assessment Monitoring 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 2.996

Figure 2: ADH Data Presentation for the Public – Reliability Recommendation



Check Relative Standard Error (RSE)

RSE < 25%

Display Estimate

RSE ≥ 25%

Display Estimate with cautionary note

Calculation of RSE?

For data with a binomial distribution:

- Numerator = A
- Denominator = B
- Proportion, r = A/B
- Standard Error = SE = $\surd(r(1-r)/B)$
- Relative SE, RSE = 100(SE/r)

For data with a Poisson distribution:

- A = count of events
- B = population
- Rate = A/B
- SE of the rate = $\surd(rate(1 - rate))/population = \surd A/B$
- RSE = 100(SE/rate) which simplifies to $100(\surd A/A)$.

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?

Generally, a confidentiality breach is defined as a loss or unauthorized access, use or disclosure of confidential information. As a condition of employment, all ADH staff are bound by agency policies and rules, as applicable to various programs, to protect data confidentiality. In addition, staff are required to maintain and disclose any Protected Health Information (PHI), as defined in the federal regulations, in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Standards (U.S.C. 1320d et seq.) and its implementing regulations including, but not limited to, 45 Code of Federal Regulations (CFR) parts 142, 160, 162 and 164. Staff must also comply with any other applicable federal law and regulation. It is the responsibility to all staff, particularly those handling Protected Health Information (PHI), to be familiar with these laws, policies, and rules.

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 within on one margin (i.e. row or column) of the table allows a table reader to ascertain which category the person falls within 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 question 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. 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 these standards?

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). As of this writing, NCHS requires that all data originating from NCHS and released by CDC (such as in tables produced by online query systems WONDER[1] and WISQARS[2] suppress counts that are less than 10, as well as rates and proportions based on counts less than 10. NCHS adopted this standard 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 are even more stringent than this—they specify that no cell (and no statistic based on a cell) of 10 or less may be displayed.[3]

In contrast to these standards, the ADH 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 (see Page 15) when zero counts for one category result in identifying all of the members of the group with a sensitive characteristic. For example, in a school 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 the data tables that will be released.

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

> *ZIP Code 47863 overlaps counties A and B. In 2020, 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 2020, 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. The*

> ***Appendix*** *provides a detailed explanation of this example and the effects of suppressing counts of 1-4 compared to 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 (see Page 14). 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, the entire database is considered restricted confidential information. In other systems, many but not all data items are confidential. In yet other systems, none of the items are confidential. Survey data often contain confidential information and may also contain information that could be used to identify an individual (such as when there is a small number 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.[4] 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, was 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 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 Privacy Rule, located at 45 CFR Part 160 and Subparts A and E of Part 164 in the federal Health Insurance Portability and Accountability Act (HIPAA)[5], defines direct identifiers as:

- Names;
- Postal address information, other than town or city, State, and zip code;
- Telephone numbers;
- Fax numbers;
- Electronic mail addresses;
- Social security numbers;

- Medical record numbers;
- Health plan beneficiary numbers;
- Account numbers;
- Certificate/license numbers;
- Vehicle identifiers and serial numbers, including license plate numbers;
- Device identifiers and serial numbers;
- Web Universal Resource Locators (URLs);
- Internet Protocol (IP) address numbers;
- Biometric identifiers, including finger and voice prints; and
- Full face photographic images and any comparable images.

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)

It is possible to link data using only indirect identifiers.[6-8] 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. See Section “Why do we have these standards” on page 8 for examples of Federal standards.

ADH standards require suppression when the number of cases or events in a cell is 1-9 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 1-9 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. There are a number of publicly available variables that are visible and, therefore, pose increased risk of disclosure when combined with other data. 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.[9] Whenever possible, data aggregation is the preferred method in order to minimize suppression.

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	3-digit ICD code	“Selected cause” tabulation
Geography	Ordinal (spatial)	ZIP code, census tract	County	State

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.

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 be 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 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 may not be meaningful for addressing certain public health problems in Arkansas. 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. 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, if for a department release of asthma data, it is not possible to achieve adequately large numbers in annual county-level data showing both age-specific and gender-specific counts and rates, the analyst may opt to omit the gender-specific data, and display only tables of age-specific data, on the grounds that intervention programs may not target 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:

1. Statewide, county or multiple county counts and rates or proportions based on counts 1-9 for an entire year **without any additional stratification**.
2. Statewide counts, or rates or proportions based on 1–9 events derived from at least three full years (36 months) of combined data, **with no temporal breakdown**.
3. Statewide counts, or rates or proportions based on **1–9** average monthly events, derived from at least **five** full years (60 months) of combined data, with *no other breakdowns*
4. County-level counts, or rates or proportions based on 5–9 events derived from at least **three** full years (36 months) of combined data, **with no temporal breakdown**.
5. 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 authorized personnel at the facility). These data should not be posted on the ADH website. However, programs should consider that once produced, these data may be subject to public disclosure requests.

In addition, agency standards allow for case-by-case exceptions, with advance approval from the Office of the Chief Science Officer. To request an exception, send an email to the Chief Science Officer (Namvar.Zohoori@Arkansas.gov) and the Deputy Chief Science Officer (Austin.Porter@Arkansas.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 ADH employees the response time will be one to three days commensurate with the urgency.

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 a school survey, a zero count in a cell such as "did not use alcohol in the past 30 days" provides meaningful information about all 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:

- Small denominators. Although there is no strict rule about the size of the denominator (population from which the event counts are derived), and the risk of disclosure depends primarily on the size of the numerator, data analysts must consider this to make sure that there is no inadvertent potential for disclosure.

  NOTE: The department routinely publishes data by county. Based on data from the 2021 American Community Survey, 2022 population estimates, 41 Arkansas counties had populations less than 20,000; about 12 of those had populations less than 20,000 person-years when combining three years of data (i.e., 2020–2022). Even though some counties have small populations, most 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 the above 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 NCHS Staff Manual on Confidentiality[10] 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."*

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

Assessing Statistical Issues

Rates, proportions, and percentages need to be checked for their stability so that trends over time and between geographic areas or persons can be evaluated with reasonable confidence. Instability can arise from small numerators (number of cases or events) or small denominators (populations or subpopulations). Note that statistical stability and confidentiality go hand-in-hand. The numerator required to generate an RSE less than a required cutoff may be well above the cutoff level for suppressing cell counts. Suppression of any rate or measure due to instability also protects confidentiality.[11]

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 ADH Office of the Chief Science Officer considers these 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 any of these recommendations, unless required to do so by their respective funding agencies.

Two approaches are outlined below: The Relative Standard Error (RSE), and a newer multistep approach now employed for select NCHS data sources and years. The RSE approach is a simpler and quicker method and is recommended by ADH as the default approach, but where possible or appropriate, or if required by their funding sources, data analysts and programs may choose to use the multistep approach.

Relative Standard Error Approach

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-know 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 may use different approaches. Unless otherwise required by funding agencies, the Department recommends marking data with RSEs greater than 25% as unreliable.

How do I calculate the RSE? At the most basic level, for data with a binomial distribution, the RSE of an estimate is obtained by dividing the standard error of the estimate, SE(r), by the estimate itself, r. This quantity is expressed as a percentage of the estimate and is calculated as:

- Percent RSE = 100 x [SE(r) / (r)]

When, however, data follow a Poisson distribution, the percent RSE is calculated as follows:

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

Estimates with large RSEs (greater than 25%) are considered unreliable, and should be flagged as such.

A simplified method for statistical reliability 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%. For binomial distributions, however, this simplified method is accurate only 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, for binomial distributions, the simplified method is conservative: it over-annotates some results as not reliable, when the numerator and denominator numbers are small.

Multistep Approach

This approach is explained in more details in the references, and those wishing or needing to use this approach are directed to those sources for more information. According to NCHS (see Appendix II of Health, United States, 2019)[12], starting with Health, United States, 2017, data presentation standards for proportions are used for selected National Center for Health Statistics (NCHS) data sources and years. The multistep standards are described in the report, National Center for Health Statistics Data Presentation Standards for Proportions (2017).[13] This multistep approach is based on three factors: minimum denominator sample sizes; the absolute and relative widths of a 95% confidence interval (calculated using the Clopper–Pearson method and adapted for complex surveys by Korn and Graubard[14]); and degrees of freedom. Using these standards, estimates identified as statistically unreliable (or whose complementary proportions are unreliable) are suppressed or flagged. This approach performs well for proportions near 0 or 1, incorporates information from the complex survey design including effective sample sizes, and is generally conservative (i.e., a 95% Clopper–Pearson confidence interval includes the true proportion more than 95% of the time).

> This multistep approach was identified after a review of current standards, the purpose and scope of data collection, and advances in statistical methodology. The use of the Korn–Graubard modification of the Clopper–Pearson confidence interval for proportions is considered an improvement over the commonly used Wald confidence interval, which is known for its under-coverage (i.e., a 95% Wald confidence interval includes the true proportion less than 95% of the time). The multistep approach has been applied to estimates starting with the 2015–2016 and 2013–2016 National Health and Nutrition Examination Surveys and the 2016 National Health Interview Survey. The reliability of estimates for prior data years was evaluated based on relative standard errors. For more information on the multistep approach, see Parker et al. 2017.[13]

Recommendations to address statistical issues

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%. 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 above). 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. 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 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 ten or fewer observations.

A 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, hospitalization rates, registry data, and vital records, for some border counties may be subject to an undercount if the dataset is used without inclusion of Arkansas residents hospitalized out-of-state.

Glossary/Definitions

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.

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” (see below).

Number of events: The number of persons or events represented in any given cell of tabulated data (e.g., numerator).

Population data: Data collected from an entire population with a specific characteristic. Examples include hospital discharge data, vital statistics, and most disease registries.

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 includes, 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.

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.

Survey Data: Data collected from a sample of a population, usually intended to be representative of a specific population. Examples include Behavioral Risk Factor Surveillance System (BRFSS), Youth Risk Behavior Surveillance System (YRBSS), Pregnancy Risk Assessment Monitoring System (PRAMS).

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

1. WONDER Multiple Cause of Death 1999-2009. Atlanta, GA: Centers for Disease Control and Prevention; https://wonder.cdc.gov/wonder/help/mcd.html#Assurance%20of%20Confidentiality. Accessed January, 23, 2023.

2. WISQARS <http://www.cdc.gov/injury/wisqars/fatal_injury_reports.html>

3. Centers for Medicare and Medicaid Services, AGREEMENT FOR USE OF CENTERS FOR MEDICARE & MEDICAID SERVICES (CMS). (https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/Data-Use-Agreement.pdf)

4. 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.

5. U.S. Department of Health & Human Services. The HIPPA Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. Accessed Jan. 15, 2023.

6. 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.

7. 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

8. 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. (https://ephtracking.cdc.gov/technicalNotes)

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

10. 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 17, 2023.

11. National Environmental Public Health Tracking. Centers for Disease Prevention and Control. Technical Notes and Instructions for the Environmental Public Health Tracking Network. https://ephtracking.cdc.gov/technicalNotes. Accessed May 17, 2023.

12. National Center for Health Statistics. Health, United States, 2019. Hyattsville, MD. 2021. DOI: https://dx.doi.org/10.15620/cdc:100685. Accessed January 23, 2023.

13. Parker JD, Talih M, Malec DJ, Beresovsky V, Carroll M, Gonzalez JF Jr, et al. National Center for Health Statistics data presentation standards for proportions. National Center for Health Statistics. Vital Health Stat 2(175). 2017. Available from: https://www.cdc.gov/ nchs/data/series/sr_02/sr02_175.pdf.

14. Korn, E. L., & Graubard, B. I. (1998). Confidence Intervals for Proportions with Small Expected Number of Positive Counts Estimated from Survey Data. Survey Methodology, 24, 193-201. https://www150.statcan.gc.ca/n1/pub/12-001-x/1998002/article/4356-eng.pdf.

Appendix: Detailed Example of Disclosure Risk

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

> ZIP Code 47863 overlaps counties A and B. In 2020, 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 2020, 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). 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.

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.

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

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 2020, 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.

Table 2: Number of 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: Number of Births by County of Residence	
County	Births
County A	450
County B	1421