Utah Department of Health
Data Suppression Decision Rules Work Group
Report of Guidelines for Data Result Suppression
October 5, 2009

Introduction

Government agencies that collect, store and report data have a responsibility to protect the privacy of individuals and to present reliable results. At the same time, these agencies have a desire to make a broad range of data publicly available in order to inform policy and guide interventions in as timely a manner as possible.

Staff in the Utah Department of Health (UDOH) decided to reassess current procedures and provide guidance to address these issues for its published reports, online static data tables and particularly for its online custom query system output. In fact, this initiative was at least partly motivated by the fact that the UDOH, like many other public health agencies, has increasingly moved to the use of the Web, including a Web-based data query system (WDQS), to disseminate public health data as quickly and widely as possible. There is currently no standardization in how public health agencies, and even different programs in the same agency, address reliability and privacy issues for their WDQSs. Complicating this task is the fact that unlike published reports and static Web information, WDQSs provide less documentation on how to interpret the results of their custom query output.

Background

Several UDOH staff members met in May 2007 to discuss suppression rules for the Behavioral Risk Factor Surveillance System (BRFSS) data on Utah's Indicator-Based Information System for Public Health (IBIS-PH) WDQS known as IBIS-Q. At that time we were planning enhancements to the BRFSS query capability to include small areas, additional cross-tabulations between BRFSS measures and the creation of custom age groups. Staff knew these enhancements would increase the likelihood of small cell sizes, so we decided it was necessary to re-evaluate suppression rules for the BRFSS data and make them consistent for all queries. This grew into a department-wide effort with the involvement of many UDOH staff in order to establish guidelines across programs. See Appendix A, *Utah Department of Health (UDOH) Data Suppression Decision Rules Work Group, Group Charter* for further information about this initiative. There was general support to develop a UDOH policy so that it could be used to explain and justify our methods.

Other developments that prompted this initiative included the Health Insurance Portability and Accountability Act (HIPAA) confidentiality issues around "Protected Health Information" (PHI), new developments in data linkage ability, the need to serve differing audiences with the data, and the sensitivity of the information for certain populations based on size and vulnerability. The updated report of the Federal Committee on Statistics Methodology, *Statistical Policy Working Paper 22 (Second version, 2005)*[i], recommended by CDC's Guidelines Working Group and discussed in the publication *Updated Guidelines for Evaluating Public Health Surveillance Systems*[ii] was reviewed and taken into account in developing the UDOH guidelines.

There exist a variety of data suppression methods including those based on cell size, relative standard error (RSE) and width of the confidence interval (CI) among others. Since being included in IBIS-Q, BRFSS query data have been suppressed if the numerator was less than 5 or the denominator was less than 30. According to the Centers for Disease Control and Prevention's (CDC) BRFSS web site, data in static prevalence tables included on the web site

were not reported if the un-weighted sample size for the denominator was less than 50 or the confidence interval half width was greater than 10 for any cell[iii]. In 2002, the National Center for Health Statistics discussed criteria for suppression employed by major national data systems that are used to track Healthy People 2010 objectives[iv]. Most of the data systems are based on sample surveys and the criteria varied across the data sets. According to the report, many of the systems suppressed cells with a RSE > 30%.

There are alternatives to data suppression including data smoothing, aggregation and artificiality. These are techniques that can be applied after the decision rule has indicated the need for suppression.

Some Important Concepts

Confidence Interval

The margin of error, or confidence interval, describes the range within which one is most likely to find the true value of the statistic (e.g., the true value of a percentage or a mean in a population). In sample surveys, the confidence interval depends on two parameters: the amount of variation in the data and the size of the sample from which the data are generated. In cases where the sample size is close to the population size, the proportion of the population sampled is also a consideration. The 95% confidence interval (for a normal distribution) is 1.96 times the standard error of the statistic (e.g., the percentage or mean).

When examining health event data, or count data, such as the number of cases of disease, death, or hospitalization, where the number of events is small (<20), the Poisson distribution is used to calculate the confidence interval. For data with more than 20 events, the Poisson distribution may be approximated by the normal distribution.

Relative Standard Error

The relative standard error (RSE) also known as the coefficient of variation (CV) is a measure of the variability of the estimate compared with the magnitude of the estimate. It may be thought of as the percentage of the magnitude of the estimate that is subject to random error. The RSE describes the variability in a set of measurements by expressing the standard deviation or standard error as a percentage of the estimate.

When using survey data where measures are typically expressed as percentages (i.e. BRFSS, Youth Risk Behavior Survey (YRBS), Utah Healthcare Access Survey (UHAS), Pregnancy Risk Assessment Monitoring System (PRAMS)), if the estimated percentage is less than or equal to 50%, the RSE is calculated by dividing the standard error of the estimate (SE(R)) by the estimate itself (R). (RSE =100 x (SE(R)/R)). However, if the estimated percentage is greater than 50%, the RSE is calculated by dividing the standard error of the estimate (SE(R)) by one minus the estimate (1-R). (RSE =100 x (SE(R)/(1-R))). Estimates with large RSEs are considered unreliable. See Appendix B for examples.

When using count data (i.e. communicable diseases, births, deaths, hospital discharges, and emergency department data) the relative standard error is calculated based on the rate (R) and the number in the population (P). Rates are typically reported per 100,000 but may use another number in the denominator so that the numerator value is an easily understandable number. To express this number as a percent of the rate use the following calculation: RSE =SQRT(100,000/PR). See Appendix C for examples.

Power to Detect a Difference

In some cases, the data user will want to know whether he/she will be able to detect a difference. For instance, has health insurance coverage in Utah improved since 1996? Or, do parents of children with special health care needs have a more difficult time finding health care

for their child than do parents of other children? The ability to detect a statistically significant difference depends on 1) the magnitude of statistical significance required (usually $\alpha \leq .05$), 2) the sample size (n), and 3) the actual magnitude of the difference. Small samples can be used to detect large differences, especially if the alpha criterion is lenient. But to detect small differences at the conventional level of statistical significance, the sample size must be relatively large.

Context / Use of the Estimate

The context of how the estimate is being used should be taken into account. Some estimates will be used to make decisions that impact a great number of people, whereas other estimates will be used to provide a rough idea of the magnitude of a problem or condition. Perhaps the former situation (high impact) should cause us to use a stricter criterion than the latter.

Methodology

UDOH staff reviewed and discussed some of the recent literature about this topic[v],[vi],[vii]. Using 2004-2006 BRFSS data, staff ran cross tabulations in SAS-Callable SUDAAN in order to compute the standard error, weighted rate and confidence intervals for several indicators, some with low prevalence or limited number of respondents by small area. This information was exported to an Excel spreadsheet and used to calculate the relative standard error and confidence interval length. We were then able to apply and compare various suppression rules including: 1) less than 5 observations in the numerator or 30 in the denominator, 2) the RSE greater than 30%, 3) the RSE greater than 50%, 4) the confidence interval length larger than the estimated percent, and 5) 1/2 the confidence interval length greater than 10% and denominator less than 50.

In addition, a statistician in the Bureau of Health Promotion had an idea to create a RSE-like estimate for Poisson-distributed count data using the following formula: 100 x [(UCL-LCL)/(2 x 1.96 x rate)] where UCL is the upper confidence limit and LCL is the lower confidence limit. The RSEs created using this method were compared to those calculated using the standard error.

The work group also determined that it would be advisable to calculate the RSE for percentages using the actual percentage if it was less than or equal to 50%, and the complement of the percentage for those greater than 50% (i.e. RSE =100 x (SE(R)/R) for $R \leq 50\%$ and RSE =100 x (SE/(1-R)) for $R > 50\%$). This was deemed necessary because measures can be reported as either a percentage of respondents with a specific attribute such as health insurance coverage or the percentage without the attribute such as those without health insurance coverage.

Staff concluded that using rules based on the relative standard error, or RSE (also known as the coefficient of variation) was a statistically sound method that could be implemented within the WDQS.

Suggested Guidelines for Reporting UDOH Data Results

Following are the final recommendations of this work group. We decided to describe the criteria as guides for reporting data. There are two sets of criteria, strict and minimum. Each is suggested for a specific type of use or context. When the criteria are not met, the estimate should be suppressed and not reported. If the estimate is used, it should be used with caution, and there should be full disclosure of the limitations of the estimate. A summary of this information can be found in APPENDIX D as a flow chart. These are considered a minimum standard, and programs that have a higher federal standard or other requirements should use those.

Minimum Criteria

For Reporting Survey Data and Population Event Data:

- ✓ RSE $\leq$ 50%
- ✓ If 30% < RSE $\leq$ 50% an asterisk should be included with a footnote that says: *Use caution in interpreting, the estimate has a relative standard error greater than 30% and does not meet UDOH standards for reliability.

Minimum criteria are to be applied in the following circumstances:

- To inform program decisions that involve small numbers of people,
- To be able to measure gross changes in a measure over time or across groups, or
- To inform the allocation of a small amount of money or other resources.

Strict Criteria

For Reporting Survey Data:

- ✓ $\geq$ 10 cases in the numerator
- ✓ AND a RSE $\leq$ 30%

For Reporting Population Event Data:

- ✓ $\geq$ 20 cases in the numerator and $\geq$100 persons in the population
- ✓ AND a RSE $\leq$ 30%

Strict criteria are to be applied in the following circumstances:

- To inform a policy decision that will impact a large number of people,
- To be able to measure small changes in a measure over time or across groups
- To inform the allocation of a large amount of money or other resources.
- When there is a legislative or agency rule, policy or standard that mandates suppression
- When there is a public expectation that this data will be suppressed

Strict criteria may also be applied if misuse of the data could:

- cause undo public alarm and unwarranted response (e.g., past experience has shown that insufficient data that weakly suggests an increased risk in an area can result in media driven public health policy that consumes public health resources where stronger data would have demonstrated no increased risk).
- result in actual public harm (e.g., could the data be used to miss-guide the public or contradict an established intervention resulting in the public ignoring the safeguards).
- impede agency intervention activities (e.g., similar to above but instead of the intervention just being ignored, the intervention is brought into question or blocked).

Confidentiality Criteria

For purposes of reporting data, certain criteria may be applied that guard against identification of an individual in a public data set.

- ✓ $\geq$ 100 persons in the population (e.g., if you are looking at the causes of deaths among 10-19 year olds, there must be at least 100 persons age 10-19 in the population of interest, if you are looking at early prenatal care among teens giving birth, there must be at least 100 teens who gave birth)
- ✓ $\geq$ 20 cases in the numerator (e.g., if you are looking at the causes of deaths among 10-19 year olds, there must be at least 20 persons age 10-19 in the population of interest who died in the time period, if you are looking at early prenatal care among teens giving birth, there must be at least 20 teens who did not have early prenatal care).

Confidentiality criteria are to be applied when the data publisher wants to protect the confidentiality of members of the population from which the data derive. One must assume that all information is sensitive information. An individual's address, whether a child is immunized, whether a woman received early prenatal care, the cause of death for an individual, or whether a person was hospitalized for a certain condition: these are all pieces of information that can potentially harm an individual or his or her family if they are made public. Certain provisions may be made for the release of this information - and those provisions are beyond the scope of this document.

Issues to take into consideration when evaluating the need to protect confidentiality:

- Vulnerability:
    - Could data subjects (even if they are not identified) be exploited (i.e., criminal predation, commercial contact, scams, etc.)? Remember, that data subjects may be exploited via targeted exploitation if only the areas they live in are identified.
    - Who are the potential exploiters? Does one exist? What kinds of exploitation could occur? How damaging is exploitation (i.e., commercial advertising to targeted neighborhood identified as having problems may be an irritation but may not warrant withholding data).
- Identifiable data:
    - Could data subjects be identified if this data were linked to other publicly available data?
    - Could data subjects be identified if multiple queries of the data at different scales and subsets were conducted to reveal masked information?

Additional things to consider when reporting:

Age-adjusted percentages and rates:

The group decided that when reporting age-adjusted rates, the relative standard error will be calculated based on the crude rate and suppression rules will be applied accordingly. Once the relative standard error has been determined, the age-adjusted rate may be calculated and should apply the same footnote when applicable, based on the relative standard error of the crude rate.

Validity / Defensibility:

- Could the data pass scientific or public scrutiny?
- Does the sample design produce representative data?
- Is there a lot of missing elements in the data?

Context:

- Is the data being delivered in a context-free environment? Data presented in a report where ample opportunity to provide context, discuss the weaknesses and concerns of the data, and provide interpretation of the data should not be subject to as stringent criteria as data provided in a context-free environment (i.e., a Web-based data query system).

---

[i] Statistical Policy Working Paper 22 (Second Version, 2005), retrieved from http://www.fcsm.gov/working-papers/spwp22.html, July 19, 2008

[ii] Updated Guidelines for Evaluating Public Health Surveillance Systems, MMWR R&R July 27, 2001:50(RR13):1-35. Retrieve from http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm, July 19, 2008

[iii] Retrieved from http://apps.nccd.cdc.gov/brfss/index.asp, July 19, 2008.

[iv] Klein RJ, Proctor SE, Boudreault MA, Turczyn KM. Health People 2010 criteria for data suppression. Statistical Notes, no 24. Hyattsville, Maryland: National Center for Health Statistics, June 2002.Access online 7/19/08 at: http://www.cdc.gov/nchs/data/statnt/statnt24.pdf

[v] Rudolph BA, Shah GH, Love D. Small Numbers, Disclosure Risk, Security and Reliability Issues in Web=based Data Query Systems. Journal of Public Health Management and Practice, 2006, 12(2), 176-183.

[vi] Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk, NAHDO-CDC Cooperative Agreement project, CDC Assessment Initiative, December 2004. Accessed online 7/19/08 at: http://nahdo.org/CS/files/folders/technical_publications/entry163.aspx.

[vii] Stoto MA. Statistical Issues in Interactive Web-based Public Health Data Dissemination Systems. RAND Health Working Paper Series: WR-106, October 2003. Prepared for the National Association of Public Health Statistics and Information systems. Accessed online 7/19/08 at: http://www.rand.org/pubs/working_papers/2005/RAND_WR106.pdf

APPENDIX A

Utah Department of Health (UDOH), Data Suppression Decision Rules Work Group

Group Charter

November 6, 2007

Group Goal

The UDOH Data Suppression Decision Rules Workgroup (Data Suppression Group) desires to provide guidance to UDOH programs on decision rules for suppressing UDOH data results to 1) protect privacy of Utahns, and 2) ensure reliability of published data results.

Background

There is an inherent tension between, on the one hand, making data results publicly available and, on the other hand, suppressing data results that are unreliable or might identify individuals who are represented in those data results.

The Case for Releasing Data

Public health data are extremely valuable to inform public policy decision-making and provide for the public good. Accurate and timely data enable necessary resource allocation to public health activities that prevent disease, disability and death. Prevention opportunities at the community level are often the most effective. UDOH programs are regularly asked to release data for small populations, including health districts, other geographic communities, and communities based on other characteristics.

The Risks of Inappropriate Release

Privacy. Public health datasets generally consist of data records that refer to a specific individual (e.g., survey responses, birth certificates, hospital visits, and cases of disease), and contain information that is sensitive in nature. We have an obligation to protect the confidentiality of the data in our custody so that individual Utahns may be protected from harm and embarrassment.

Reliability. We might also do harm by releasing a count or rate that may mislead users because it is not reliable. The most common threat to data reliability across all our health event datasets is the size of the sample or population. Data based on small samples or populations is subject to greater variability (e.g., across time periods). We must reflect our knowledge of the reliability of a data result when it is reported (such as by reporting confidence intervals). Are there also instances in which a data result is so unreliable that it should be suppressed entirely?

Data Suppression Rules

There exist a variety of rules for when to suppress data. For instance, the national PRAMS program uses “30 sample cases in the denominator.” Utah’s IBIS-PH Query system suppresses a data result when there are either fewer than 5 cases in the numerator or 30 in the denominator. Others have used relative standard error (RSE), and a ratio of the numerator to the denominator.

Data suppression rules apply to the question, “When should data be suppressed?” They do not address a related question, “How should the data be suppressed?” The UDOH Data Suppression Decision Rules Work Group will address the first question, but not the second.

Work Group Members

Shelly Wagstaff, DCFHS, Bureau of Health Promotion
Michael Friedrichs, DCFHS, Bureau of Health Promotion
Laurie Baksh, DCFHS, Maternal and Child Health Bureau

Shaheen Hossain, DCFHS, Maternal and Child Health Bureau
Bruce Wood, DHCF
David Jackson, DELS, Communicable Disease Epidemiology Program
Greg Williams, DELS, Communicable Disease Epidemiology Program
Michael Lowe, DELS, HIV/AIDS
Sam LeFevre, DELS, Environmental Epidemiology Program
Diane Hartford, DHSI, Emergency Medical Services
Iona Thraen, DHSI
Barry Nangle, CHD
Brian Paoli, CHD, Office of Public Health Assessment
Lois Haggard, CHD, Office of Public Health Assessment (facilitator)
Kathryn Marti, CHD, Office of Public Health Assessment
Jennifer Wrathall, CHD, Office of Public Health Assessment
Tong Zheng, CHD, Office of Public Health Assessment
Kimberly Partain McNamara, CHD, Office of Public Health Assessment
Carol Masheter, CHD, Office of Health Care Statistics
John Morgan, CHD, Office of Health Care Statistics
Mylitta Barrett, CHD, Office of Vital Records and Statistics
Jeffrey Duncan, CHD, Office of Vital Records and Statistics

Group Work Products

The UDOH Data Suppression Decision Rules Work Group will produce a report that documents our conclusions. The document will contain:

- Decision rules for data result suppression that will take into account a variety of considerations (e.g., privacy, reliability, audience, sensitivity of the information, vulnerability of the population, data linkage, etc.), and can be applied to a variety of data types (e.g., sample surveys, vital records, eligibility and claims databases, etc.) reported in various ways (e.g., online query system, static reports, etc.).
- The document will present a parsimonious solution that will be acceptable department-wide. Failing that goal, a small set of rules will be articulated, along with a decision tree that provides guidance on which rules should be applied under specific circumstances.
- The document will include the justification for the decision rules, and the issues and considerations that led to them.

Group Activities

The group will meet monthly, for no longer than 1 ½ hours to review information and draft the data suppression guidance. It was tentatively decided (9/19) to start by designing decision rules for data release on the IBIS-PH query system.

Group Charter Revision History:

Created	Lois M. Haggard	9/21/07
Revised:	Lois M. Haggard	11/6/07

APPENDIX B

Utah Department of Health, Data Suppression Decision Rules Work Group

October 6, 2009

Calculation of Relative Standard Error (RSE) for Percentages

RSE = Relative Standard Error
R = rate (percentage in this case)
SE (R) = Standard Error of estimated percentage

If the rate is less than or equal to 50% then use this formula:
RSE =100 x (SE(R))/R)

If the rate is greater than 50% then use the following formula:
RSE =100 x (SE(R))/(1-R))

Example #1

The crude estimated percentage of Utah adults aged 85 and older who were current cigarette smokers in 2007 was 7.65% with a standard error of 3.81%.

RSE = 100 x (0.0381/0.0765) = 49.8%

Example #2

The crude estimated percentage of Utah adults aged 85 and older who were non-smokers in 2007 was 92.35% with a standard error of 3.81%.

RSE = 100 x (0.0381/(1-0.9235)) = 100 x (0.0381/0.0765) = 49.8%

Note: You should get the same RSE here since example #2 is reporting the complement of the information in example #1. If we were using the general minimum criteria, both percentages would be reported but would include a footnote stating to use caution in interpreting the result because the estimate has a relative standard error greater than 30% and does not meet UDOH standards for reliability.

APPENDIX C

Utah Department of Health, Data Suppression Decision Rules Work Group

October 6, 2009

Calculation of the Relative Standard Error (RSE) for Count Data

D = # of deaths ~ Poisson
P = # in population where deaths occurred
R = rate per 100,000 = (D/P) *100,000

Var(R) = Var [(D/P) x 100,000]
= $(100,000/P)^2$ Var(D)
= (100,000/P) x ((100,000*D)/P)
= (100,000/P) x R

RSE = 100 x [SQRT(Var(R))]/R
= 100 x [SQRT((100,000/P)*R)]/R
= 100 x SQRT[100,000/(P*R)]

Example #1

D = 388 Alzheimer's disease deaths in Utah in 2006
P = UT Population in 2006 was 2,615,129
R = (388/2,615,129) x 100,000 = 14.84 per 100,000

Var(R) = (100,000/2,615,129) x 14.84 = 0.5675

RSE = 100 x SQRT[100,000/(2,615,129 x 14.84)] = 5.08%

Example #2

D = 5 Alzheimer's disease deaths in Utah adults between the ages 55-64 in 2006
P = 55-64 UT Population in 2006 was 201,340
R = (5/201,340) x 100,000 = 2.48 per 100,000

Var(R) = (100,000/201,340) x 2.48 = 1.23

RSE = 100 x SQRT[100,000/(201,340 x 2.48)] = 44.75%

These two examples illustrate the how the RSE increases in a small population that only includes Utah adults aged 55-64 in Example #2, versus the total Utah population used in Example #1.

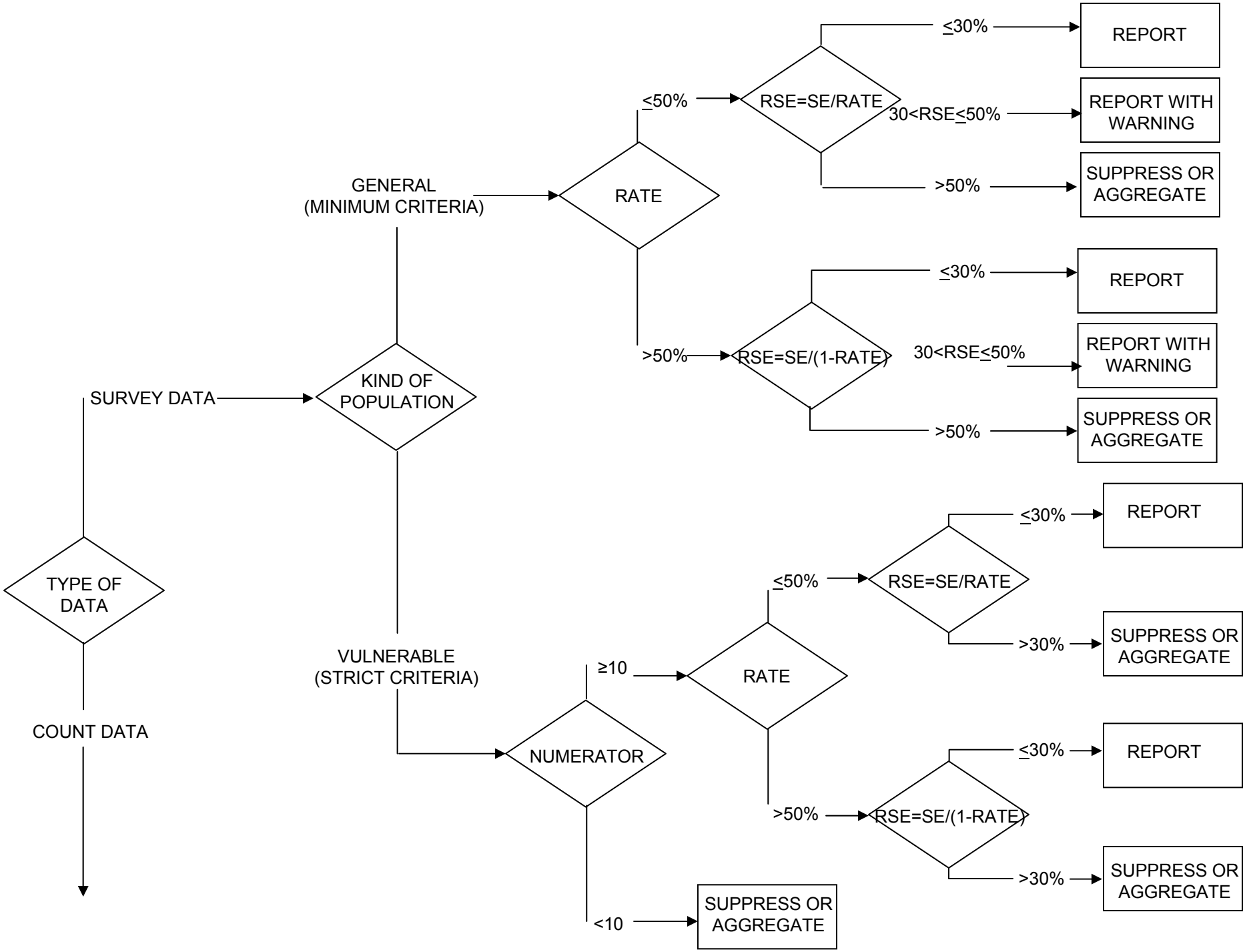
APPENDIX D

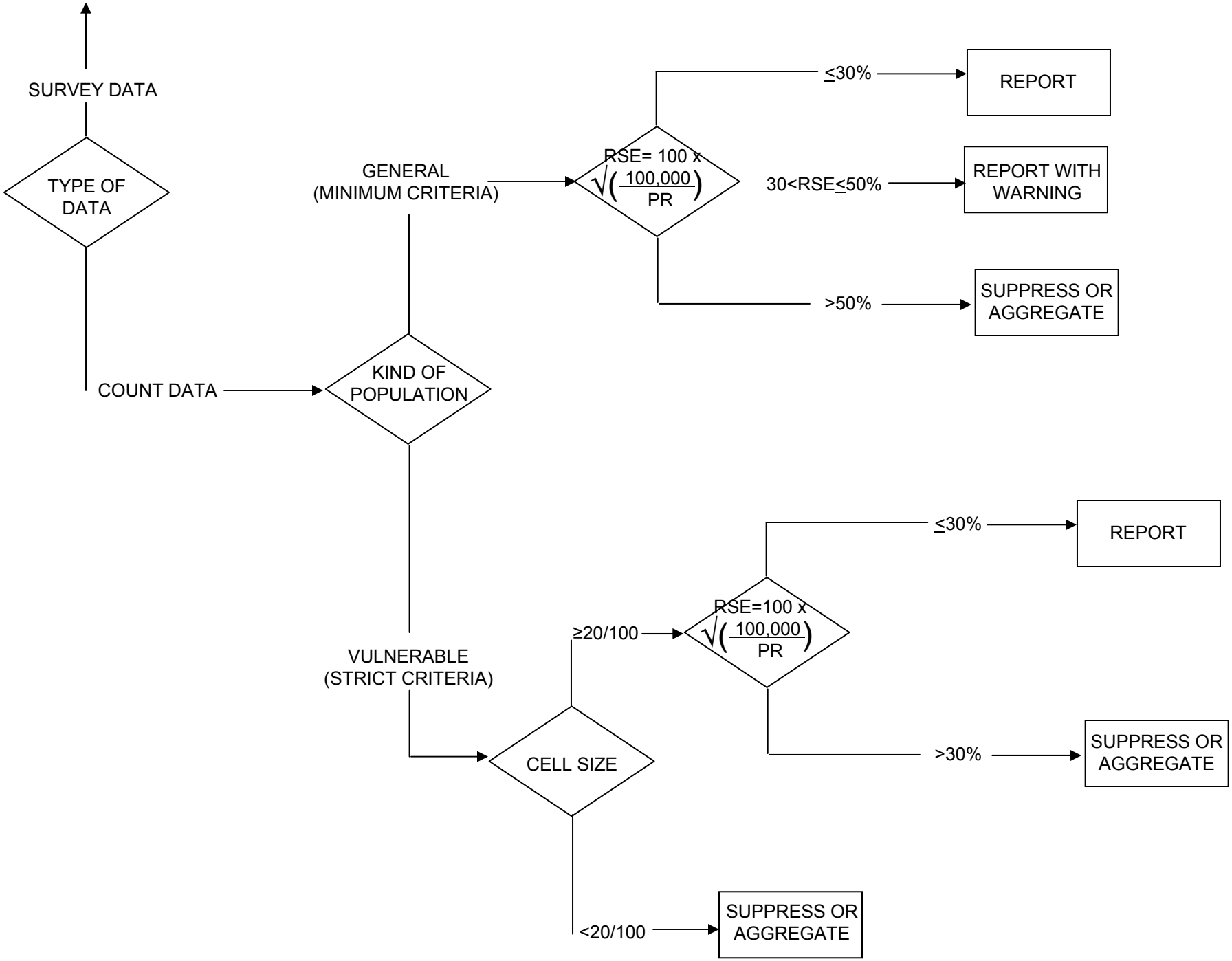
Utah Department of Health

Data Suppression Decision Rules Work Group:

Decision Tree Flow Chart

October 6, 2009


SURVEY DATA
TYPE OF DATA
COUNT DATA
KIND OF POPULATION
GENERAL (MINIMUM CRITERIA)
RATE
≤50%
RSE=SE/RATE
≤30%
REPORT
30<RSE≤50%
REPORT WITH WARNING
>50%
SUPPRESS OR AGGREGATE
>50%
RSE=SE/(1-RATE)
≤30%
REPORT
30<RSE≤50%
REPORT WITH WARNING
>50%
SUPPRESS OR AGGREGATE
VULNERABLE (STRICT CRITERIA)
NUMERATOR
≥10
RATE
≤50%
RSE=SE/RATE
≤30%
REPORT
>30%
SUPPRESS OR AGGREGATE
>50%
RSE=SE/(1-RATE)
≤30%
REPORT
>30%
SUPPRESS OR AGGREGATE
<10
SUPPRESS OR AGGREGATE


SURVEY DATA
TYPE OF DATA
COUNT DATA
KIND OF POPULATION
GENERAL (MINIMUM CRITERIA)
RSE= 100 x √(100,000/PR)
≤30%
REPORT
30<RSE≤50%
REPORT WITH WARNING
>50%
SUPPRESS OR AGGREGATE
VULNERABLE (STRICT CRITERIA)
CELL SIZE
≥20/100
RSE=100 x √(100,000/PR)
≤30%
REPORT
>30%
SUPPRESS OR AGGREGATE
<20/100
SUPPRESS OR AGGREGATE