HPCDP Guidelines for Reporting Reliable Numbers

Version 2
Final: March 7, 2018

Introduction

The Health Promotion and Chronic Disease Prevention (HPCDP) section developed the *Guidelines for Reporting Reliable Numbers* to ensure that only reliable estimates from HPCDP's surveillance system are reported. This document is not a primer for statistics and is focused only on the reliability of data. This means that it only pertains to estimates. It does not pertain to the release of counts of individuals or release of identified or de-identified data, which are issues related to confidentiality. The Oregon Public Health Division's Guidelines for *Reporting Small Numbers to Protect Confidentiality* covers confidentiality.

In Summary

- ✓ **Decision 1:** Use these guidelines to determine reliability and reporting of estimates
- ✓ **Decision 2:** Do not report estimates of zero or 100%.
- ✓ **Decision 3:** Do not use complementary suppression if only reporting estimates.
- ✓ **Decision 4:** Do use complementary suppression if also reporting the number of events. If reporting the number of events use the Oregon Public Health Division's *Guidelines for Reporting Small Numbers to Protect Confidentiality* for methods to protect confidentiality or the Oregon Cancer Registry guidelines for protecting confidentiality for cancer data.
- ✓ **Decision 5:** If the age-adjusted rate or percentage is reported with warning or suppressed, then the crude rate or percentage is also reported with warning or suppressed, and vice versa.

The purpose of these guidelines is to inform HPCDP analysts about reporting estimates with different reliabilities. Specifically, these guidelines define when:

1. Estimates are reported as reliable
2. Estimates need to be reported with a notation that they are potentially unreliable
3. Estimates need to be suppressed because they are statistically unreliable

Failure to follow these guidelines could result in the reporting of estimates that are not representative of the population or that are misleading. Providing reliable estimates is essential as it builds trust in the results provided by HPCDP. Releasing unreliable data could make partner organizations discount provided estimates.

Methods

The two types of data used by HPCDP, survey and full count, use different methods for estimating reliability. The definitions of these types of data are included below.

A survey dataset contains information gathered from asking questions of randomly selected respondents. Complex surveys such as the Behavioral Risk Factors Surveillance System (BRFSS) weight (adjust) data so that estimates are generalizable to the entire population. Estimates calculated from survey data are subject to sampling error.

A full count dataset theoretically contains every instance of a qualified event (for example death, birth, cancer diagnosis, or hospitalization) for that population. Because every event is available, these data are not subject to sampling error. However, these data are subject to random errors (measurement errors, processing errors, missing data, and other forms of errors). Because of random errors, estimates based on a small number of events may be unreliable.

To determine whether estimates calculated from survey or full count data sources should be reported, reported with warning, or not reported (i.e., suppressed), each step in these guidelines needs to be completed for every estimate unless otherwise indicated (see Figure 1 or Figure 2 for flowcharts).

If estimates are determined to be unreliable, or need to be reported with a warning, there are some approaches you may consider to improve/allow reporting. The easiest option is to combine groups to increase sample size. For example, if you were running smoking rates by age groups, and the 18-24 age group would be suppressed, you could instead group to an 18-34 age span. Another option is to combine multiple years of data (for example three years of BRFSS datasets) if the 18-24 age group specifically is needed. Other approaches are possible, so consult your fellow analysts and appropriate statistical reference for additional guidance.

Guidelines for reporting Survey Data

Step 1: Determine if the estimate is calculated on a full population (i.e., the denominator is every survey respondent such as all Oregonians) or a subpopulation (i.e., the denominator is a county, a race group, or a group of people who share a common trait). Proceed to Step 2a or b.

Step 2a: If full population, determine if the denominator is greater than or equal to (GE) 50. If yes, proceed to Step 3. If not, suppress estimate. Indicate suppression with a symbol.

Step 2b: If subpopulation, determine if the denominator is greater than or equal to (GE) 30. If yes, proceed to Step 3. If not, suppress estimate. Indicate suppression with a symbol.

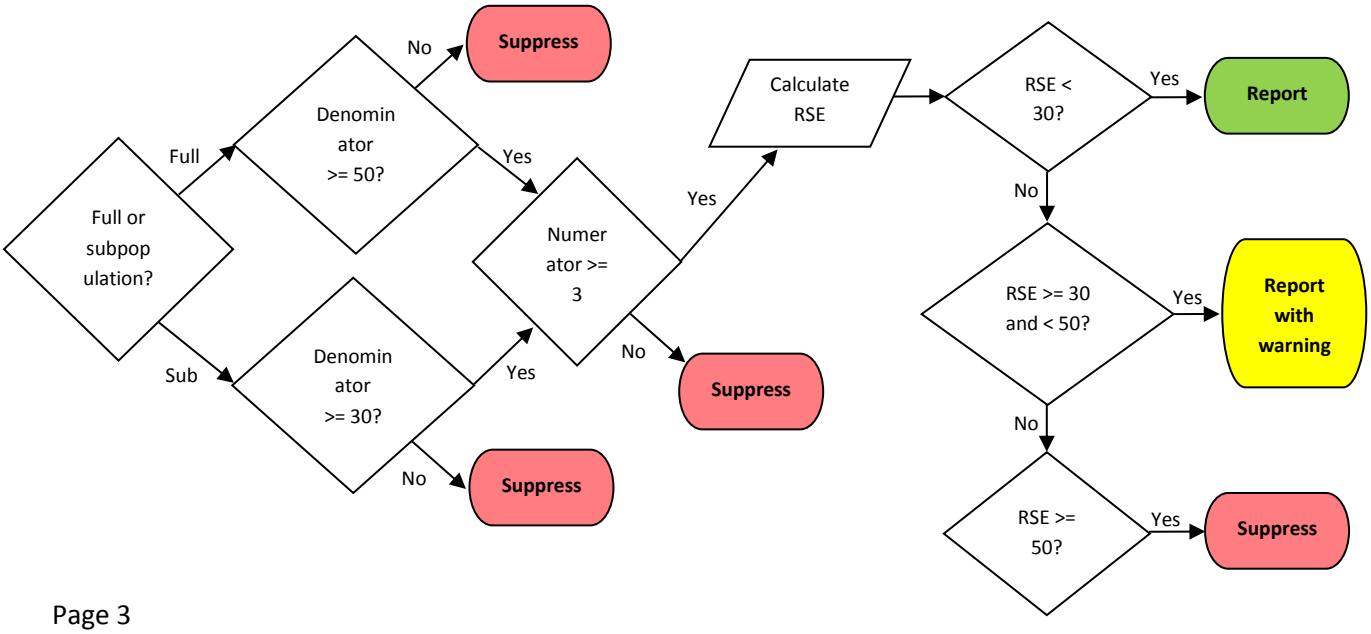
Step 3: For either full or subpopulation, if the numerator is greater than or equal to 3, proceed to step 4. If not, suppress estimate. Indicate suppression with a symbol.

Step 4: Calculate the relative standard error (RSE). Directions are on page 4.

Step 5: Apply the following logic to each RSE.

- RSE < 30%: Report estimate.
- RSE >= 30% and RSE <50%: Report estimate with a warning, and include a notation on the same page as the estimate. Examples of notations include "*This number may be statistically unreliable and should be interpreted with caution."* OR "*May be statistically unreliable; interpret with caution*."
- RSE >= 50%: Do not report estimate, and include a notation on the same page as the estimate. Examples of notations include "*This number is suppressed because it is statistically unreliable.*" OR "*Suppressed; statistically unreliable*."

Figure 1. Flowchart for determining whether and how to report survey data estimates



Guideline for reporting Full Count Data

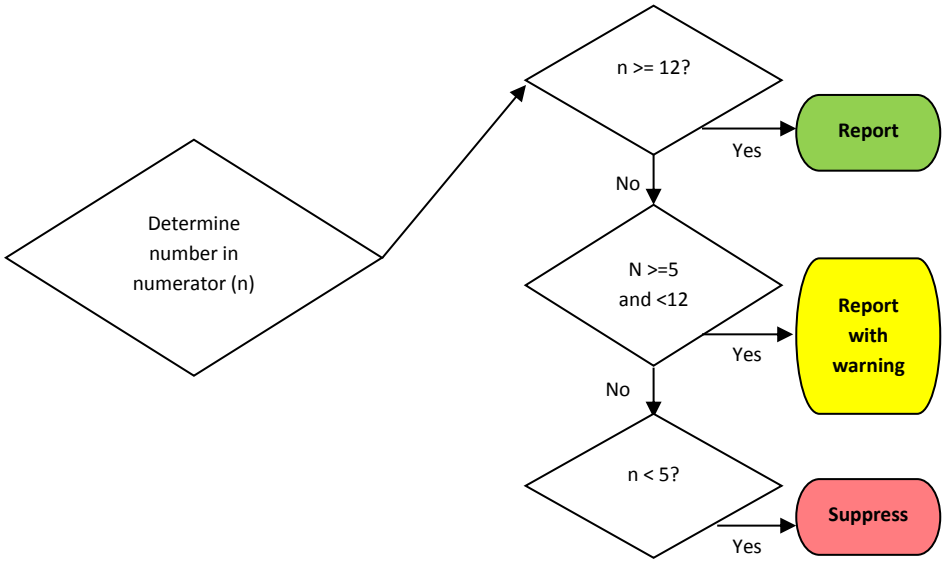
Step 1: Determine the numerator (count of events) for your analysis.

Step 2: Apply the following logic to the numerator (n)

- n >=12: Report estimate.
- n>=5 and n <12: Report estimate with a warning, and include a notation such as "*This number may be statistically unreliable and should be interpreted with* caution." OR "May be statistically unreliable; interpret with caution."
- n <5: Do not report estimate and include a notation such as "*This number is suppressed because it is statistically unreliable*." OR "*Suppressed; statistically unreliable*."

Note: In the guidelines for reporting Survey data (previous page), reliability is based on RSE. For Full Count data however, the reliability is based on the number in the numerator. While this may appear to be a discrepancy in methodology, it is not because the numerator from a full count data set actually provides the RSE. See page 5 under the section called "**How is the RSE calculated for full count data**" for information on how RSE is calculated from full count datasets.

Figure 2. Flowchart for determining whether and how to report full count estimates



Relative Standard Error (RSE) or Coefficient of Variance (CV)

What is the RSE?

RSE is a measure of the variability of an estimate compared with its size. RSE is the percentage of the size of the estimate that is subject to random error. The RSE describes the variability in a set of measurements by expressing the standard error as a percentage of the estimate.

The most basic formula for estimating the standard error is:

Standard error (SE) = standard deviation (x_i) / square root (N)

Where:

x is a set of measurements and N is the sample size

This formula is an example. For **survey datasets** such as the BRFSS, this equation is not sufficient because the design of the survey is a significant factor. For most **full count datasets**, HPCDP assumes that events are from a Poisson probability distribution. For a Poisson probability distribution, standard error is based on the numerator.

How is the RSE calculated for survey data?

RSE = 100 * (SE(X)/X)

Where:

SE(X) is the standard error of your estimate and X is your estimate

How is the RSE calculated for full count data?

RSE = 100 * square root (1/n)

Where:

n = the count of events (numerator)

What statistics software can calculate and report the RSE for survey data?

SAS: Add the "CV" option to your code so that it reports the coefficient of variance.

SPSS: HPCDP does not have the complex samples module so calculation of RSE is not possible.

STATA: Add the "cv" option to your code so that it reports the RSE.

Examples: Survey Data

Example 1

Analytic question: Among people who use electronic cigarettes, what is the distribution of smoking behavior?

Stata syntax

```
** DATA FILE
use "I:\HPCDP_Surveillance-Final\Annual BRFSS\HPCDP_BRFSS2016.dta", clear
svyset [pweight=wtrk_ac_aco], strata(denst)

** USE tbsmokehx3 and Yes on currently chew tobacco
svy: tab tbsmokehx3,subpop (if tbecignow==1) obs ci cv percent format(%9.3fc) stubwidth(28)
```

Stata Output (portions deleted for clarity)

CALC: Hx/Status (Never, Former, Current)	percentages	cv	lb	ub	obs
Never Smoker	12.26	20.84	8.07	18.21	26.00
Former Smoker	23.98	12.53	18.59	30.35	70.00
Current Smoker	63.76	5.55	56.58	70.37	167.00
Total	100.00				263.00

Results

Step 1: This calculation is on a subpopulation (people who currently use electronic cigarettes)

Step 2b: The denominator for this subpopulation is 263 (highlighted in the "Total" row). Since 263 is greater than 30, next look at the RSE.

Step 3: As calculated by the Stata syntax, the relative standard errors (shown in the "cv" column) for the estimates range from 5.55 to 20.84.

Step 4: The RSE for Current Smoker is 5.55 (highlighted in the "cv" column), so this estimate (63.76%) is reported without a warning.

Example 2

Analytic question: Among women, what proportion smoked full-sized cigar in last month?

Stata syntax

```
** DATA FILE
use "I:\HPCDP_Surveillance-Final\Annual BRFSS\HPCDP_BRFSS2016.dta", clear
svyset [pweight=wtrk_c_c], strata(denst)

** USE full-sized cigar and women
svy: tab tbcigarlg, subpop (if sex==2) obs cv format(%9.3fc) stubwidth(28)
```

Stata Output (portions deleted for clarity)

```
-------------------------------------------------------------------------------
Smoked full-sized cigar      |
in last month                |        proportions                cv          obs
-----------------------------+-------------------------------------------------
                      Yes    |             0.003            74.397         2.00
                      No     |             0.997             0.229     1,621.00
                             |
                        Total |             1.00                        1,623.00
-------------------------------------------------------------------------------
  Key:  proportions         =  cell proportions
        cv                  =  coefficients of variation of cell proportions
        obs                 =  number of observations
```

Step 1: This calculation is on a subpopulation (women)
Step 2b: The denominator for this subpopulation is 1,623 (highlighted in the "Total" row). Since 1,623 is greater than 30, next look at the RSE.
Step 3: RSE (CV) is calculated by the Stata syntax.
Step 4: The RSE for currently smoked full-sized cigar is 74.397 (highlighted in the "cv" column), so this estimate (0.3%) is suppressed.
Note that the numerator of 2 (respondent females who smoked full-sized cigars in the last month) would also have required suppression for this particular analysis.

Examples: Full Count Data

Example 1

Analytic question: What was the Oregon rate of diabetes deaths among African-American females in 2006?

Count of events = 10
Rate = 41.64 diabetes deaths out of 100,000 Oregon African-American female residents

Step 1: The numerator for this question is 10.
Step 2: The numerator is greater than five but less than 12; so this estimate (41.64 per 100,000) is reported with a warning.

Example 2

Analytic question: What was the rate of hospitalization with a primary diagnosis of asthma in Lake County in 2007?

Count of events = 1
Rate = 1.2 asthma hospitalizations out of 10,000 Lake county residents

Step 1: The numerator for this question is 1.
Step 2: The numerator is less than 5, so - this estimate (1.2 per 10,000) is suppressed.

Frequently Asked Questions

Question 1: *What does it mean when a number is "unreliable" and why do we care?*

Both survey and full count data are subject to various forms of error. The effect of errors, such as sampling errors in a survey, is increased when an estimate (such as the prevalence of tobacco use) is based on a small number of survey respondents. The consequence is the estimate could fluctuate dramatically from year-to-year and not be representative of the population under study. Thus, the estimate is "unreliable."

It is critical that HPCDP only release reliable estimates. Providing reliable estimates is essential as it builds trust in the results from HPCDP's surveillance system. Releasing unreliable data could make partner organizations discount data released by HPCDP.

Question 2: *Should we be concerned about a breach of confidentiality with survey data?*

The answer is no, with a caveat explained in the next paragraph. Individuals in population-based surveys represent the health status, risk factors, and opinions of a large number of persons with similar demographic characteristics. Therefore, their information is weighted to the number of people they represent. Because of this, releasing information on a small numbers of individuals does not constitute a breach of confidentiality. Even so, if you follow the reliability criteria, estimates based on small numbers will not be released.

The caveat is that any dataset or estimate must be de-identified. This is rarely a problem with survey data. However, for some surveys there may be identifying information within the dataset. Examples of identifying information are names, phone numbers, dates of birth, home addresses, or any information that would make it possible to match a record in a dataset to an individual. Identifying fields should be removed from the dataset used for analysis and never reported.

Question 3: *Should we be concerned about a breach of confidentiality with full count data?*

Yes! However, this document does not address methods to protect against breach of confidentiality. Follow the Oregon Public Health Division's Guidelines for *Reporting Small Numbers to Protect Confidentiality*.

Question 4: *What is the difference between full population and subpopulation in survey data?*

Full population is when the denominator is every respondent in the survey who answered a question. Examples from the BRFSS would include the prevalence of diabetes in Oregon or the percentage of people in Oregon who are obese. In Oregon, the BRFSS has a number of survey versions. If the question of interest is only on some of those versions, and the denominator is all respondents on those versions, this is still considered full population. Another example is a special survey, such as a stand-alone version of the BRFSS that is only conducted in a single county. If the denominator for a question (such as obesity in that county) is for all respondents of the survey, that is also considered full population.

Subpopulation is when only specific segments of the survey respondents make up the denominator. For example, when investigating the percentage of Oregonians who are obese by gender, age groups, race, county of residence, having a particular disease, or any other subcategory, these would all be considered subpopulations. Using the gender example for females, the percentage obese is calculated using the numerator as the number of females who are obese divided by the denominator as the number of females in the survey. The denominator is only females, not all survey respondents, and is therefore a subpopulation.

Question 5: *How should I report numbers of individuals from survey data?*

With survey data, the raw frequencies are usually not informative. Generally, only the number that the individuals represent is useful. For example, 21 individuals surveyed in the BRFSS could represent 2,500 Oregonians. When releasing this information you would provide the number 2,500, not 21. If raw frequencies are released, it should be made clear to the audience what the frequencies represent. This is not an issue of reliability but of helping readers correctly interpret the information.

Question 6: *If one estimate is suppressed, should all estimates in a table or graph be suppressed?*

This question pertains to a situation where one response category receives a warning for reliability or is suppressed, but the other(s) do not. One possible extension of the reliability rules would be to warn or suppress all other levels of the question. This approach is often called complementary suppression.

The second survey data example (see page 7) shows the percentage of Oregon women who currently smoked full-sized cigar. In this example, which is a yes/no question, the "yes" answer is suppressed. If you used complementary suppression the 'no" answer would also be suppressed even though the RSE is acceptable.

HPCDP does NOT use complementary suppression when only the estimate is reported. It is important for HPCDP analysts to provide as much quality information as possible to the public and decision makers. Complementary suppression decreases the amount of information that can be released.

On the other hand, if the number of events is reported (for example the number of pregnancies among women who used tobacco in a single year and small geography) then complementary suppression IS used to protect against a breach of confidentiality.

There are methods that an analyst can use to reduce the need for suppression when reporting a question with multiple responses. Multiple responses could include reporting categories of weight (healthy weight, overweight, obese, and extremely obese). For this example, assume that the percentage of people who are extremely obese is suppressed. The analyst could choose to collapse categories by grouping the two obesity categories into a single "obese" category. This would help avoid suppression, simplify reporting, and increase understanding of the information by a general audience.

Question 7: *What is the Poisson probability distribution and why is it used?*

The Poisson probability distribution is one of many ways to identify the spread (variability) for a population measure. It is an appropriate distribution when looking at discrete and independent events. Other probability distributions exist. For example, the most commonly taught probability distribution in statistics is the normal (or bell curve) distribution. However, it is outside the scope of this document to describe the theories behind the various probability distributions and the strengths and weakness of each.

HCPDP is using the Poisson distribution because it is the method for estimating RSE used by the National Center for Health Statistics (NCHS) and is appropriate for distributions based on rare events (small numbers). It is also easy to calculate and apply.

Question 8: *For full count data sources such as births and deaths, the National Center for Health Statistics (NCHS) uses a count of fewer than 20 to warn that data may be unreliable. Why is HPCDP using different counts?*

NCHS uses 20 as an arbitrary benchmark. A count of 20 events equates to a RSE of approximately 23%. NCHS does not suppress estimates but only warns that data may be unreliable.

HPCDP is using a different standard. Where counts equate to an RSE<30%, the estimates are reported. Where counts equate to an RSE >=30% but <50%, estimates are reported with a warning regarding reliability. Finally, where counts equate to an RSE >=50%, estimates are suppressed. HPCDP uses this same standard for survey data. Using these RSE levels makes HPCDP's standard less conservative with regard to warning for unreliability, but more conservative because estimates based on very small numbers are suppressed.

References

Martin, J.A., Hamilton, B.E., Sutton, P.D., Ventura, S.J., Menacker, F., Kirmeyer, S, and Mathews, T.J. 2009. National Vital Statistics Reports: Births: Final Data for 2006: 57(1). Available online at www.cdc.gov/nchs/data/nvsr/nvsr57/nvsr57_07.pdf. Accessed February 2010.

Heron, M., Hoyert, D.L., Murphy, S.L., Xu, J., Kochanek, K.D., and Tejada-Vera, B. 2009. National Vital Statistics Reports: Deaths: Final Data for 2006: 57(14). Available online at www.cdc.gov/nchs/data/nvsr/nvsr57/nvsr57_14.pdf. Accessed February 2010.

Klein, R.J., Proctor, S.E., Manon, M.S., Boudreault, M.A., and Turczyn, K.M. 2002. Healthy People 2010 Criteria for Data Suppression. Healthy People 2010 Statistical Notes: 24.