# COMPARABILITY OF DATA: BRFSS 2006

The BRFSS is a cross-sectional surveillance survey currently involving 54 reporting areas (1, 2). BRFSS questionnaires, data, and reports are available on the Internet at www.cdc.gov/brfss. It is important to note that any survey will have natural variation across sample sites; therefore, some variation between states is to be expected. The complex sample design and the multiple reporting areas complicate the analysis of the BRFSS. Although CDC works with the states to minimize deviations, in 2006 there were some deviations in sampling and weighting protocols, sample size, response rates, and collection or processing procedures. In addition, California's questionnaire had a few minor differences in wording of questions. The following section identifies other known variations for the 2006 data year.

## A. 2006 Data Anomalies and Deviations from Sampling Frame and Weighting Protocols

In 39 states, a portion of sample records intended for use during one month took more than one month to complete. In eight instances, states used their monthly sample over a period of several months. This deviation will disproportionately affect analyses based on monthly, rather than annual, data. Additionally, Michigan received its sample quarterly rather than monthly.  Alaska did not use the designated sample to collect data in November and December.  Kentucky did not use the designated sample to collect data in May.

Several states did not collect data for all 12 months of the year.  Louisiana did not collect data in January and February. Texas and U.S. Virgin Islands did not collect data in January.  Guam did not collect any data in 2006.

Several states were unable to close out the December sample in 2006 and data collection continued into early 2007.  Alaska, Colorado, Illinois, Kentucky, Louisiana, Michigan, Nevada, New Mexico, North Carolina, Ohio, South Dakota, and Utah had some completed interviews in January 2007.  When attempting to analyze the data by month, the variable FMONTH should be used to distinguish the monthly sample for a state.

More information about the quality of the survey data can be found in the 2006 *BRFSS Summary Data Quality Report*.

## B. Other Limitations of the 2006 Data

Telephone coverage by landline varies by state and also by subpopulation. Telephone coverage averages 93.1% for the United States as a whole, but noncoverage ranges from 3.1% in Utah and Washington to 12.1% in Arkansas. It is estimated that 23.8% of households in Puerto Rico are without telephone service. Data on telephone coverage in U.S. households are available at http://www.fcc.gov/wcb/iatd/trends.html.

California modified the wording and/or response categories of core questions addressing health plans, diabetes, Hispanic ethnicity, educational attainment, household income.  California also inserted additional questions into the core of the survey and reordered sections of the core. The data from these questions may therefore have limited comparability to those of other reporting areas.

The data from an optional module is included if asked of all eligible respondents within a state for the entire data collection year. A state may have indicated the use of an optional module in 2006, but the data may have been moved into the state-added questions if it does not represent all eligible respondents.

A change in 2002 to the final disposition codes has continued to present some inconsistencies in closing out the questionnaire. Prior to 2002, interviews that were terminated during or after the

demographics section were coded as complete interviews, and any remaining unanswered questions were coded as refused by the interviewer. In 2002, a revised procedure was implemented for handling partial completes. The revised procedure for partial completes is to stop coding questions at the point of interview termination to assign the appropriate disposition code. The missing and refused values should be taken into account when determining which records to include in an analysis. Records with a termination in the questionnaire followed by coded refusals for the remainder of the eligible responses have been dispositioned as 120 Partial Completes.

Another issue regarding partial completes is the inappropriate coding of the remaining questions. For example, some questions allow responses of 01-76, 77, 88, and 99 (with 99 as the refusal code). The problem occurs when an interviewer incorrectly codes the remaining questions as refused by entering '9' instead of a '99' for these questions. Where nine (9) is a valid response for these questions and should not have been used to indicate refusal.; When reviewing responses to a partial complete, data users should therefore be aware that a core section or module that follows the demographics section may contain questions incorrectly coded as refused ('9 filled'). Records are evaluated upon submission to identify core sections following the demographics section or optional modules with more than half of the responses coded as unknown, refused, or missing (7, 77, 777, 9, 99, 999, or blank).  Records meeting the criteria have their disposition code changed to reflect a partial complete (120) and the data for the corresponding variables set to missing to remove some of the confusion over miscoded refusals.  The Sexual Violence and Intimate Partner Violence optional modules are impacted by this change of disposition because of a large number of refusals to all questions in the modules.

Several states continue to ask the Diabetes module questions directly after the Diabetes questions in the core of the survey. In addition, several states ask the Adult Asthma module questions after the asthma questions in the core. Some states have also asked the Childhood Asthma module questions in the demographics section of the core survey after question 6, (CHILDREN) – number of children under age 18 in household.

More information about survey item nonresponse can be found in the 2006 BRFSS Summary Data Quality Report and in the respective states' Data Quality Reports.


## STATISTICAL AND ANALYTIC ISSUES

### Estimation Procedures

Unweighted data on the BRFSS represent the actual responses of each respondent, before any adjustment is made for variation in respondents' probability of selection, disproportionate selection of population subgroups relative to the state's population distribution, or nonresponse. Weighted BRFSS data represent results that have been adjusted to compensate for these issues. Irrespective of state sample design, use of the final weight in analysis is necessary if generalizations are to be made from the sample to the population.


### Statistical Issues

The procedures for estimating variances described in most statistical texts and used in most statistical software packages are based on the assumption of simple random sampling (SRS). However, the data collected in the BRFSS are obtained through a complex sample design; therefore, the direct application of standard statistical analysis methods for variance estimation and hypothesis testing may yield misleading results. There are computer programs available that take such complex sample designs into account. SAS Version 8's SURVEYMEANS and SURVEYREG procedures, SUDAAN, and Epi Info's C-Sample are among those suitable for analyzing BRFSS data (3,4,5). SAS and SUDAAN can be used for tabular and regression analyses (3,4); SUDAAN has these and additional options (4). Epi Info's C-sample can be used to

calculate simple frequencies and two-way cross-tabulations (5). When using these software products, users must know the stratum, the primary sampling units, and the record weight—all of which are on the public use data file. For more information on calculating variance estimations using SAS, see the *SAS/STAT Users Guide, Version 8* (3). For information about SUDAAN, see the SUDAAN Users Manual, Release 7.5 (4). For information about Epi Info, see *Epi Info, Version 6.0* (5).

Although the overall number of respondents in the BRFSS is more than sufficiently large for statistical inference purposes, subgroup analyses can lead to estimators that are unreliable. Consequently, users need to pay particular attention to the subgroup sample when analyzing subgroup data, especially within a single data year or geographic area. Small sample sizes may produce unstable estimates. Reliability of an estimate depends on the actual unweighted number of respondents in a category, not on the weighted number. Interpreting and reporting weighted numbers that are based on a small, unweighted number of respondents can mislead the reader into believing that a given finding is much more precise than it actually is. The BRFSS follows a rule of not reporting or interpreting percentages based upon a denominator of fewer than 50 respondents (unweighted sample). For this reason, the FIPS County code is removed from the data file for any county with less than 50 respondents.


## Analytic Issues

### Advantages and Disadvantages of Telephone Surveys

Compared with face-to-face interviewing techniques, telephone interviews are easy to conduct and monitor and are cost efficient. However, telephone interviews have limitations. Telephone surveys may have higher levels of noncoverage than face-to-face interviews because some U.S. households cannot be reached by telephone. As mentioned earlier, approximately 98% of households in the United States have telephones. A number of studies have shown that the telephone and non-telephone populations are different with respect to demographic, economic, and health characteristics (6, 7, 8). Although the estimates of characteristics for the total population are unlikely to be substantially affected by the omission of the households without telephones, some of the subpopulation estimates could be biased. Telephone coverage is lower for population subgroups such as blacks in the South, people with low incomes, people in rural areas, people with less than 12 years education, people in poor health, and heads of households under 25 years of age (9). However, poststratification adjustments for age, race, and sex, and other weighting adjustments used for the BRFSS data minimize the impact of differences in noncoverage, undercoverage, and nonresponse at the state level.

Despite the above limitations, prevalence estimates from the BRFSS correspond well with findings from surveys based on face-to-face interviews, including studies conducted by the National Institute on Alcohol Abuse and Alcoholism, CDC's National Center for Health Statistics, and the American Heart Association (10, 11). A summary of methodologic studies of BRFSS is provided in the publication section at www.cdc.gov/brfss.

Surveys based on self-reported information may be less accurate than those based on physical measurements. For example, respondents are known to underreport weight. Although this type of potential bias is an element of both telephone and face-to-face interviews, the underreporting should be taken into consideration when interpreting self-reported data. However, when measuring change over time, this type of bias is likely to be constant, and is therefore not a factor in trend analysis.

With ongoing changes in telephone technology, there are more and more households that have cellular telephones and no traditional telephone lines in their homes. These households are presently not in the sampling frame for the BRFSS, which may bias the survey results, especially if the percentage of cellular-telephone-only households increases in the coming years (12–14).

The BRFSS is continuing to study the impact of cellular phones on survey response and the feasibility of various methods for data collection to complement present survey methods (1).

**Aggregating Data Over Time**

When data from one time period are insufficient for estimating the prevalence of a risk factor, data from multiple periods can be combined as long as the prevalence of the risk factor of interest did not substantially change during one of the periods. One method that can be used to assess the stability of the prevalence estimates is as follows (10):

1. Compute the prevalence for the risk factor for each period.

2. Rank the estimates from low to high.

3. Identify a statistical test appropriate for comparing the lowest and the highest estimates at the 5% level of significance. For example, depending on the type of data, a t-test or the sign test might be appropriate.

4. Test the hypothesis that prevalence is not changing by using a two-sided test in which the null hypothesis is that the prevalences are equal.

5. Determine whether the resulting difference could be expected to occur by chance alone less than 5% of the time (i.e., test at the 95% confidence level).

**Analyzing Subgroups**

Provided that the prevalence of risk factors did not change rapidly over time, data combined for two or more years may provide a sufficient number of respondents for additional prevalence estimates for population groups (such as age/sex/race subgroups or county populations). Before combining data for subgroups, it is necessary to determine whether the total number of respondents will yield the precision needed, which depends upon the intended use of the estimate. For example, greater precision would be required to justify implementing expensive programs than that needed for general information only.

The table below shows the sample size required for each of several levels of precision, based on a calculation in which the estimated risk factor prevalence is 50% and the design effect is 1.5.

| Precision Desired | Sample Size Needed |
|---|---|
| 2% | 3600 |
| 4% | 900 |
| 6% | 400 |
| 8% | 225 |
| 10% | 144 |
| 15% | 64 |
| 20% | 36 |

Precision is indicated by the width of the 95% confidence interval around the prevalence estimate. For example, precision of 2% indicates that the 95% confidence interval is plus (+) or minus (-) 2% of 50%, or 48% to 52%. As shown in the table, to yield this high a level of precision, the sample size required is about 3,600 persons. When a lower level of precision is acceptable, the sample size can be considerably smaller.

The **design effect** is a measure of the complexity of the sampling design that indicates how the design differs from simple random sampling. It is defined as the variance for the actual sampling

design divided by the variance for a simple random sample of the same size (10,15). For most risk factors in most states, the design effect is less than 1.5. If it is more than 1.5, however, sample sizes may need to be larger than those shown in the table above.

The standard error of a percentage is largest at 50% and decreases as a percentage approaches 0% or 100%. From this perspective, the required sample sizes listed in the table above are conservative estimates. They should be reasonably valid for percentages between 20% and 80%, but may significantly overstate the required sample sizes for smaller or larger percentages.


**Creating Synthetic Estimates**

Even after combining data for several years, sample sizes may still be inadequate for risk factor estimates for some geographic areas (e.g., counties) or subpopulations (e.g., people with diabetes). In such situations, the analyst may wish to derive synthetic estimates by extrapolating from BRFSS data collected at the state level.

Synthetic estimates can be calculated using the population estimates for the subgroup of interest and the statewide BRFSS risk factor prevalences for that subgroup. This approach assumes that the risk factor prevalences for specific subgroups in each area are the same as the statewide risk factor prevalences for the same subgroups. For example, it assumes that the risk factor prevalences for black women in every county of a state are the same as those for black women in the entire state. The accuracy of the estimate depends on the validity of this assumption, which is often impossible to judge. However, a "ballpark" estimate may be sufficient for establishing broad goals and objectives for prevention strategies. For a discussion of the precision of such estimates, see Levy and Lemeshow, 1991 (16).

An example for estimating the number of people with hypertension in a hypothetical county, as well as the overall prevalence of hypertension in that county, is shown below. The sex and race distribution of the county's population differs from the statewide population, and these differences need to be taken into account. By developing a table like the one below, a synthetic estimate for the overall county prevalence of hypertension can be made.

Synthetic Estimates of Prevalence of Hypertension in a Hypothetical County, 2000

| Subgroup | Statewide Prevalence* 2000 | County Population 2000 | County Population With Hypertension 2000 |
|---|---|---|---|
| *Men* | | | |
| White | 15.6 | 10,000 | 1,560 |
| Black | 27.0 | 25,000 | 6,750 |
| | | | |
| *Women* | | | |
| White | 19.5 | 12,000 | 2,340 |
| Black | 26.5 | 28,000 | 7,420 |
| *Total* | | 75,000 | 18,070 |

*Per 100 persons

The statewide prevalence values, given as rates per 100 persons, are computed from the BRFSS data. The estimated number of persons with hypertension for each race-sex group in the county was obtained by multiplying the statewide prevalence for that group by the county population for the group. To determine the total county prevalence, the number of people with hypertension in each race-sex group in the county were summed and this sum (18,070) was divided by the county's total population (75,000) to yield an overall prevalence of 24.1 per 100 persons.

**Creating Direct Estimates**

Provided that the subpopulation sample size is sufficient, analysts may choose to produce direct estimates. SUDAAN or a similar program will be needed for direct estimates. If possible, it is desirable to re-adjust the poststratification weight (_POSTSTR) to the age-by-race-by-gender population distribution of the subarea (e.g., county). To locally post-stratify the CDC BRFSS weights used for the direct estimate, poststratify _WT2 to the population of interest. The equivalent local final weight is a product of _WT2 and the local poststratification factor.

**New Calculated Variables and Risk Factors**

Not all of the variables that appear on the public use data set are taken directly from the state files. CDC prepares a set of SAS programs that are used for end of year processing. These programs prepare the data for analysis and add weighting, sample design, calculated variables, and risk factors to the data set. The following calculated variables and risk factors, created for the user's convenience, are examples of results from this procedure:

MODCAT_, VIGCAT_, PACAT_, _RFHLTH, _RFNOPA, _RFHYPE4

The procedures for the variables vary in complexity; some only combine codes, while others require sorting and combining selected codes from multiple variables, which may result in the calculation of an intermediate variable (e.g., MODCAT_, VIGCAT_, PACAT_). For further details regarding the calculated variables and risk factors, refer to the document entitled "Calculated Variables and Risk Factors for the 2006 Behavioral Risk Factor Surveillance System," located at http://www.cdc.gov/brfss/technical_infodata/surveydata/2006.htm.

# REFERENCES

1.  Mokdad AH, Stroup DF, Giles WH. Public health surveillance for behavioral risk factors in a changing environment. Recommendations from the Behavioral Risk Factor Surveillance Team. *MMWR Recomm Rep* 2003;52:1–12.


2.  Holtzman D.  The Behavioral Risk Factor Surveillance System. In Blumenthal DS, DiClemente RJ, editors Community-based health research: issues and methods, New York, Springer Publishing Company Inc., 2004;115–131.


3.  SAS Institute Inc., SAS/STAT User's Guide, Version 8. Cary, NC: SAS Institute, Inc., 1999; 3181–3272.


4.  Shah BV, Barnwell BG, Bieler GS. SUDAAN User's Manuel, Release 7.5. Research Triangle Park, NC:  Research Triangle Institute, 1997.


5.  Dean AG, Dean JA, Coulombier D, Brendel KA, Smith DC, Burton AH, Dicker RC, Sullivan K, Fagan RF, Arner TG. Epi Info, Version 6.0: A word processing, database, and statistics program for public health on IBM-compatible microcomputers. Centers for Disease Control and Prevention, 1995.


6.  Groves RM, Kahn RL. Surveys by Telephone: A national comparison with personal interviews, New York, Academic Press, 1979.


7.  Banks MJ. Comparing health and medical care estimates of the phone and nonphone populations. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1983; 569–574.


8.  Thornberry OT, Massey JT. Trends in United States Telephone Coverage Across Time and Subgroups. In Groves, RM et al editors Telephone Survey Methodology. New York, John Wiley & Sons, 1988; 25–49.


9.  Massey JT, Botman SL. Weighting Adjustments for Random Digit Dialed Surveys. In Groves, RM et al editors Telephone Survey Methodology. New York, John Wiley & Sons, 1988; 143–160.


10. Frazier EL, Franks AL, Sanderson LM. Behavioral Risk Factor Data. In Using chronic disease data: a handbook for public health practitioners. Centers for Disease Control and Prevention. 1992; 4.1–1.17.


11. Nelson DE, Powell-Griner E, Town M, Kovar MG. A comparison of national estimates from the National Health Interview Survey and the Behavioral Risk Factor Surveillance System. *Am J Public Health* 2003; 93;1335–1341.

12. Link MW, Mokdad AH. Leaving answering machine messages: do they increase response rates for RDD surveys?  *Int J Public Opin Res* 2004; 482.


13. Link MW, Mokdad AH. Are web and mail feasible options for the Behavioral Risk Factor Surveillance System?  In Cohen SB, Lepkowski JM, editors Eighth Conference on Health Survey Research Methods. Hyattsville, MD: National Center for Health Statistics, 2004; 149–158.


14. Link MW, Mokdad AH, Town M, Roe D, Weiner J. Improving response rates for the Behavioral Risk Factor Surveillance system: use of lead letters and answering machine messages. Proceedings of the American Statistical Association, Survey Methodology Section (CD-ROM). Alexandria, VA, 2004; 141–148.


15. Groves RM. Survey errors and survey costs. New York: John Wiley and Sons, 1989; 265, 271–272.


16. Levy PS, Lemeshow S. Sampling of populations: methods and applications. New York: John Wiley and Sons, 1991; 347–350.