COMPARABILITY OF DATA: BRFSS 2002

The BRFSS is a cross-sectional surveillance survey currently involving 54 reporting areas. It is important to note that any survey will have natural variation over 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 2002 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 2002 data year.

A. 2002 Data Anomalies and Deviations from Sampling Frame and Weighting Protocols

In 75% of the states, a portion of sample records intended for use during one month took more than one month to complete, in some instances, several states used their monthly sample into several months. This deviation will disproportionately affect analyses based on monthly, rather than annual data.

Several states did not collect data for all 12 months of the year or completed interviews in calendar year 2003. Massachusetts and Puerto Rico did not report any interviews in January. The Virgin Islands did not complete any interviews in January or February. New Jersey did not complete any interviews in January, June, July, August, and September. Nevada completed over 800 interviews in January and February, 2003. Alabama, Arkansas, Colorado, Idaho, Illinois, Indiana, Kentucky, North Dakota, New Mexico, New York, Oklahoma, Oregon, Pennsylvania, South Carolina, Texas, Utah, Virginia, the Virgin Islands, and Wyoming had some completed interviews in January, 2003.

More information about the quality of the survey data can be found in the *2002 BRFSS Summary Data Quality Report*.

B. Other 2002 limitations of the data

Telephone coverage varies by state and also by sub-population. Telephone coverage averages 97.6% for U.S. states as a whole, but ranges from 1.1% noncoverage in Connecticut and New Hampshire, to 6.6% in Mississippi. It is estimated that 23.8% of households in Puerto Rico are without telephone service. Data on telephone coverage in United States households are available at *http://factfinder.census.gov* .

Pennsylvania asked all survey respondents the core questions on the 2002 survey, but five different sets of modules were asked based on the residence of the survey respondents. Individuals in Armstrong, Franklin, Chester, and Lancaster Counties were all asked different sets of modules and the remaining state residents were asked a different

set of module questions. This resulted in five different questionnaires being implemented in Pennsylvania and all module data collected was moved to the state-added questions section. Illinois used a dual questionnaire and collected data on core items involving health status, health care access, exercise, asthma, diabetes, immunization, tobacco use, prostate cancer screening, colorectal cancer screening, and HIV/AIDS knowledge and prevention, and demographics from all eligible respondents. The alcohol consumption core question on drinking and driving (variable DRINKDRI) was asked of about half the respondents while the rest of the questions (ALCDAY3, AVEDRNK, and DRNK2GE5) were asked of all respondents. For the firearms core questions, question 1 (FIREARM4) was asked to all respondents, while questions 2 and 3 (GUNLOAD and LOADULK2) were asked on only half of the respondents. Fruit and vegetable consumption, oral health, use of seatbelts, family planning, and women's health questions were asked of about half of eligible respondents. Modules on hypertension awareness, cholesterol awareness, physical activity, healthy days, and childhood asthma were asked of approximately half of all eligible respondents.

California modified the wording of the following core questions, firearms, health plan, diabetes, the frequency of alcohol consumption, Hispanic ethnicity question, level of education, and household income questions. The data from the firearms questions are not included in the data set. In addition, California used different age cut-offs for the colorectal cancer screening questions. These questions may have limited comparability to those of other reporting areas.

Some states had a problem with skip patterns in the Health Care Access section of the Core Survey. For January and part of February for some states, due to CATI programming issues, question 4 (MEDCARE) and question 5 (MEDREAS) had the same variable name. This produced a potential error in the data if someone answered the MEDCARE question "1. Yes" to MEDCARE and then answered 2, 7, or 9 to MEDREAS, their results were overwritten and may have been lost. Several states did call individuals a second time to clarify the MEDCARE and MEDREAS variable values. There were several issues in regards to the new disposition codes that were implemented this year, especially the new code for partial completes (coded 120) and completed interviews (coded 110). In the past, interviews terminated during or after the demographics section would have been coded as a complete interview and the remaining questions unanswered would be coded as refused by the interviewer. For 2002 states handled partial complete interviews differently, they were generally handled in one of three ways: 1) answered the remaining questions as refused and coded the record a 110 Complete, 2) answered the remaining questions as refused and coded the record a 120 Partial complete, or 3) did not ask the remaining questions (answers left as missing) and coded the record a 120 Partial complete. The differences in how the interviews are dispositioned and where in the survey the interview was terminated will have an impact on refusal rates for certain questions/modules and will also affect numbers of completes and partial completes. This will vary state to state. These factors should be taken into account when determining which records to include in any analyses. Changes are being made in 2003 to further clarify the difference between partial completes and complete interviews using just disposition codes.

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. There were a few states that asked the Childhood Asthma module questions of respondents that answered "Refused" to the demographics question assessing the number children in the household. Most states collecting module asked the questions only of respondents indicating there were children in the household.

More information about survey item nonresponse can be found in the *2002 BRFSS Summary Data Quality Report* and in the respective states' *Data Quality Reports*.

STATISTICAL AND ANALYTIC ISSUES

Estimation Procedures

Unweighted data on the BRFSS are the actual responses of each respondent. Unweighted data represent results before any adjustment is made for variation in respondents= probability of selection, for 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 such differences. 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 given in most statistical texts and the programs available in most statistical software packages are based on the assumption of simple random sampling (SRS). 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 EpiInfo=s C-Sample are among those suitable for analyzing BRFSS data. SAS and SUDAAN can be used for both tabular and regression analyses; SUDAAN has more available options. EpiInfo=s C-sample can be used to calculate simple frequencies and two-way cross-tabulations. These software products require knowing the stratum, the primary sampling units, and the record weight—all of which are on the master data file. For more information on calculating variance estimations using SAS, see SAS Institute, 1999 (10).

For information about SUDAAN, see Shah, Barnwell, Bieler, 1997 (1). For information about EpiInfo, see Dean, et al, 1995 (2).

Although the overall number of persons in the BRFSS is quite large for statistical inference purposes, subgroup analyses can lead to estimators that are unreliable. Consequently, analysis of subgroups, especially within a single data year or geographic area, requires that the user pay particular attention to the subgroup sample size. 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).

Analytic Issues

Advantages and Disadvantages of Telephone Surveys

Compared with in-person interviewing techniques, telephone interviews are easy to conduct and monitor, and cost efficient. However, telephone interviews have limitations. Telephone surveys may have higher levels of noncoverage than in-person interviews because a percentage of United States 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 (3,4,5). Although the estimates of characteristics for the total population are unlikely to be substantially affected by the omission of the non-telephone households, some of the subpopulation estimates could be biased due to the noncoverage of households without telephones. Telephone coverage is lower for population subgroups such as blacks in the South, persons with low incomes, persons in rural areas, persons with less than 12 years education, persons in poor health, and heads of households under 25 years of age (6). However, post-stratification 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 in-person 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 (7). A summary of methodologic studies of BRFSS is provided in the publication section on *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 under report weight. Although this type of potential bias is an element of both telephone and in-person interviews, the under reporting should be considered by the analyst interpreting self-reported data. When measuring change over time, this type of bias is likely to be constant, and 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 and may bias the survey results, especially if the percentage of cellular phone only households increase in the coming years. The BRFSS has plans to study the impact of cellular phones on survey response and also the feasibility of various methods for data collection to complement present survey methods.

Aggregating Data Over Time

When data from one time period are insufficient for estimating the prevalence of a risk factor, data may be combined for several periods as long as the periods being combined are not times during which the prevalence of the risk factor of interest has been substantially changing. One method that can be used to assess the stability of the prevalence estimates is discussed below (7).

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

When the prevalence of risk factors does not change rapidly over time, data combined for two or more years may provide a sufficient number of respondents so that additional prevalence estimates can be made for population

groups (such as age/sex/race subgroups or county populations). Before combining data for subgroups, determine whether the total number of respondents will yield the precision needed. The level of precision needed depends upon the intended use of the estimate. For example, greater precision would be required to justify implementing expensive programs than that 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, a desired precision of 2% means that the 95% confidence interval is + or - 2% of 50%, or 48–52%. 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 and 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 (7, 8). 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 here.

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 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.

As a cautionary note, users should remember that the 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 CDC strongly urges all users to follow the general rule of not reporting or interpreting percentages based upon a denominator with fewer than 50 unweighted respondents.**

Creating Synthetic Estimates

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

Synthetic estimates can be calculated by using the population estimates for the subgroup of interest and the state 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 Aballpark@ 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 (9).

An example for estimating the number of persons 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, 1990

State Subgroup	Prevalence* 1990	County Population 1990	County Population with Hypertension 1990
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 state 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

If the subpopulation sample size is sufficient to do so, analysts may choose to produce direct estimates. SUDAAN or a similar program will be needed for direct estimates. The subarea (i.e., county) is treated as a population domain for which the risk estimate is sought, and will be defined as a SUBGROUP variable in SUDAAN. Temporal and spatial stratification must be incorporated into the estimates of variable, by inclusion in the NEST statement in SUDAAN. If possible, it is desirable to re-adjust the poststratification weight (_POSTSTR) to the age-by-race-by gender population distribution of the small area (i.e., county).

To locally post-stratify the CDC BRFSS weights used for the direct estimate, post-stratify _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, intermediate, and calculated variables, and risk factors to the data set. The following calculated variables and risk factors are examples of results from this procedure, and are created for the user's convenience:

_LTASTHM, _CASTHMA, _ASTHMST, _EXTEETH, _ALTEETH, _DENTVST, _FLUSHOT, _PNEUMOC, _RFMAM2Y, _RFPAP3Y, _RFPSA2Y, _RFBLDST, _RFSIGMD, _AIDSTST, _HIGHRSK, _STDCNDM, _RFFRARM, _RFFRAR2, _MCNTRAC, _FCNTRAC

Creation of the variables vary in complexity; some only combine codes, while others require sorting and combining selected codes from multiple variables and may result in the calculation of an intermediate variable. For further details regarding the calculated variables and risk factors, refer to the document entitled "Calculated Variables and Risk Factors for the 2002 Behavioral Risk Factor Surveillance System."

REFERENCES

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

2. 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.

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

4. 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, pp 569–574.

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

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

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

8. Groves RM. Survey Errors and Survey Costs. New York: John Wiley and Sons, 1989; 265, 271–272.

9. Levy PS, Lemeshow S. Sampling of Populations: Methods and Applications. New York: John Wiley and Sons, 1991; 347–350.

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