Attachment 7

Analytic Guidelines

Attachment 7 - Analytic Guidelines

NHANES Analytic and Reporting Guidelines (available online at http://www.cdc.gov/nchs/about/major/nhanes/nhanes2003-2004/analytical_guidelines.htm)

Last Update: December, 2005
Last Correction: September, 2006

Introduction

This document presents analytic and reporting guidelines that should be used for NHANES data analyses and publications. It represents the latest information from the National Center for Health Statistics on recommended approaches for analysis of all NHANES data, but with a particular focus on data collected in the continuous NHANES (since 1999). Previous versions of NHANES analytic guidelines (the NHANES III Analytic Guidelines http://www.cdc.gov/nchs/about/major/nhanes/nhanes3/nh3gui.pdf and the NHANES 1999-2000 Addendum to the NHANES III Analytic Guidelines http://www.cdc.gov/nchs/data/nhanes/guidelines1.pdf) can still be used. These analytic guidelines will be modified and updated on a periodic basis as new information is acquired and as new statistical techniques for analysis of complex sample surveys are introduced. Users should regularly visit the NHANES website to see if a new version of these latest analytic guidelines has been released.

Summary recommendations

Following is the current list of analytic and reporting guidelines for NHANES public release data. Additional guidelines may be included on future updates as well as more detailed information and examples for some of the existing guidelines.

1. The first and over-riding analytic guideline is that the data user, prior to any analysis of the data, should read all relevant documentation for the survey and for the specific data items to be used in an analysis.

Many analytic problems and misinterpretation of the data can be avoided by reading the documentation, examining the data collection protocols and data collection instruments, and conducting preliminary descriptive evaluation of the data. The documentation will indicate how the data were collected, how the data are coded and the amount of missing data. The documentation will also indicate if a data item was collected on all or a sub-sample of sample persons, if it was collected on a limited age-range, or if exclusion criteria were applied for a specific examination component. Specific information on laboratory tests and quality control for these tests are available. For trend analysis, the current documentation can be compared with documentation from past NHANES surveys to determine if a specific data item is comparable with a similar data item collected in previous surveys.

Data collected in NHANES comes from interviews, examinations, and laboratory tests based on blood and urine samples. There may also be measures taken in the home, such as

dust or tap water collection. The source of a data item (interview, MEC, sera) is important for both assessment of quality of information and for determining the appropriate sampling weight to be used for producing statistical estimates.

As with any data set, NHANES data are subject to sampling and non-sampling errors (including measurement error). Interview (questionnaire) data are based on self-reports and are therefore subject to non-sampling errors such as recall problems, misunderstanding of the question, and a variety of other factors. Examination data and laboratory data are subject to measurement variation and possible examiner effects. The NHANES program maintains high standards to insure non-sampling and measurement errors are minimized. Prior to data collection, extensive protocols are developed and reviewed by the public health and scientific community. Prior to and during data collection, NHANES field staff participate in comprehensive training and annual refresher training for Interviewers and MEC staff. As data are processed, extensive quality control procedures are applied. Despite the rigorous quality control standards, estimates produced from any data set are subject to sampling and non-sampling variation and interpretation of analysis must proceed accordingly.

Data content and data collection protocols may change over time; this is another reason to read the documentation in order to understand any issues in comparability of data over time. Changes in methods may occur at any time and the user should not assume they have

2. NHANES has changed from a periodic survey to a continuous survey and the release of public use data files (and their format) has changed as well.

In the past, NHANES surveys were conducted on a periodic basis and the data were released as single, multiyear data sets. For example, NHANES III covered the 6 calendar years 1988-1994 and is generally analyzed as one, 6-year survey. In addition, previous NHANES public use data files tended to be large and few in number. Since 1999, NHANES has been planned and conducted as a continuous annual survey. For a variety of reasons, including disclosure issues, the continuous NHANES survey data is released on public use data files in two-year increments (e.g. NHANES 1999-2000, NHANES 2001-2002, NHANES 2003-2004, etc.). Since the inception of the continuous NHANES, public use data files are released on an ongoing basis as many smaller component-specific data files. For a two-year analysis, sample size is smaller and the number of geographic units in the sample is more limited than, for example NHANES III. Sample size and statistical power consideration should be used to determine if a two-year sample is sufficient for a particular analysis or if 4 (or even 6) years of the survey need to be combined to produce statistically reliable analysis. This is addressed more fully later in this document.

3. Be aware of the complex survey design and sample weighting methodology..

NHANES is a complex sample survey. The overall sample design and weighting methodology has been similar over the history of the survey. The sample design and weighting methodology for NHANES 1999-2004 is very similar to past NHANES data releases. Primary Sampling Units (PSUs) are generally single counties, although small counties are sometimes combined to meet a minimum population size. In the years 1999-2001, NHANES was based on a design linked to the National Health Interview Survey (NHIS). The NHANES PSUs were a subset of the PSUs previously selected for the NHIS. An independent set of PSU's was selected for 2002-2006; the sampling frame for this design was all counties in the United States.

The additional stages of selection in the probability design for NHANES 1999-2004

remain very similar to past NHANES designs. Clusters of households are selected and each person in a selected household is screened for demographic characteristics. One or more persons per household may be selected for the sample. For NHANES 1999-2000, there were 12,160 persons selected for the sample, 9,965 of those were interviewed (81.9 percent) and 9,282 (76.3 percent) were examined in the MEC. For NHANES 2001-2002, there were 13,156 persons selected for the sample, 11,039 of those were interviewed (83.9 percent), and 10,477 (79.6 percent) were examined in the MEC. For NHANES 2003-2004, there were 12,761 persons selected for the sample, 10,122 of those were interviewed (79.3 percent) and 9,643 (75.6 percent) were examined in the MEC.

As with any complex probability sample, the sample design information should be explicitly used when producing statistical estimates or undertaking statistical analysis of the NHANES data. In particular, sample weights and the first stage of the cluster design need to be considered. The sampling weights provided must be used to produce unbiased national estimates. The sample weights for NHANES 2003-2004 reflect the unequal probabilities of selection, non-response adjustments and adjustments to independent population controls. The proper sample weight must be used. If only data from the Interviewed sample is used, then the appropriate SAS variable is WTINT2YR. If data from the MEC examination is used, then the appropriate SAS variable is WTMEC2YR.

Because NHANES is a complex probability sample, analytic approaches based on data from simple random sample are usually inappropriate. Ignoring the complex design can lead to biased estimates and overstated significance levels. ***Sample weights and the stratification and clustering of the design must be incorporated into an analysis to get proper estimates and standard errors of estimates.***

Data are sometimes collected on sub-samples of the full design for any NHANES survey. These data are available but public release of these files may lag behind the main data release for any two-year period due to extra time needed for processing and quality assurance review. In addition, each subsample involves another stage of selection and separate sample weights that account for that stage of selection and additional non-response. For analysis of subsample data***, appropriate subsample weights must be used and they are included on any data file where relevant***.

4. Be aware of, and utilize, proper variance estimation procedures.

The procedure for variance estimation (sampling errors) is the same for 2003-2004 as for 2001-2002. This method creates Masked Variance Units (MVUs) which can be used as if they were stratified PSU's to estimate sampling errors (similar to past NHANES). The MVUs on the NHANES demographic data files are not the "true" design PSUs. They are a collection of secondary sampling units that are aggregated into groups called Masked Variance Units for the purpose of variance estimation. The MVUs produce variance estimates that closely approximate the variances that would have been estimated using the "true" design structure. These MVUs have been created for each two-year cycle of NHANES and can be used for any combination of two-year data cycles without recoding by the user.

For NHANES 2001-2002 and 2003-2004, the two-year weights and MVUs are included in the Demographics data file. The NHANES1999-2000 Demographic file was updated to include MVU's and four-year sample weights. ***Only the NHANES 1999-2002 data have special four year sample weights*** (as described in the NHANES Analytic Guidelines section on how and when to combine years of data). At this time, the preferred approach for calculating

sampling errors is to use the MVUs and to ignore the JK-1 technique that served as an interim approach for variance estimation when the NHANES 1999-2000 data were released.

The stratum variable is SDMVSTRA and the PSU variable is SDMVPSU. Software specific for survey data, such as SUDAAN, or software that has specific survey procedures, such as STATA and SAS, can be used to estimate sampling errors by the Taylor series (linearization) method. Typically, the data set should first be sorted by SDMVSTRA and SDMVPSU. For NHANES 1999-2000, SDMVSTRA is numbered 1-13; for NHANES 2001-2002, SDMVSTRA is numbered 14-28; and for NHANES 2003-2004 SDMVSTRA is numbered 29-43. Therefore, these files can be combined without any recoding of this variable. This procedure will also hold for combining NHANES 2001-2002 and 2003-2004 data files, as well as future two-year NHANES files. There are no replicate weights provided for NHANES 2003-2004. Replication techniques can still be used to estimate sampling errors if the software, such as WESVAR, computes its own set of replicate weights based on the nested MVU/PSU within stratum design.

Variance estimates for NHANES I, NHANES II, HHANES, and NHANES III utilized the true design PSUs. Pseudo strata and pseudo PSU variables were included on each public use data file for those surveys and the same software can be used to estimate sampling errors for each of those surveys.

5. Combining two or more 2-year cycles of the continuous NHANES is encouraged and strongly recommended in order to produce estimates with greater statistical reliability for demographic sub-domains and rare events,.

For two-year cycles, the sample size may be too small to produce statistically reliable estimates for very detailed demographic sub-domains (e.g. sex-age-race/ethnicity groups) or for relatively rare events. The sample design for NHANES makes it possible to combine two or more “cycles” to increase the sample size and analytic options. Each two-year cycle and any combination of those two years cycles is a nationally representative sample.

When combining cycles of data, it is extremely important that (1) the user verify that data items collected in all combined years were comparable in wording and methods and (2) use a proper sampling weight. Beginning in 2003, the survey content for each two year period is held as constant as possible to be consistent with the data release cycle. In the first four years of the continuous survey, this was not always the case, and some special data release and data access procedures had to be developed and used for selected survey content collected in “other than two-year” intervals (http://www.cdc.gov/nchs/data/nhanes/nhanes_release_policy.pdf) .

6. The decision on how many years of NHANES data are required for a particular analysis can be summarized by the concept of minimum sample size required.

The minimum sample size is determined by the statistic to be estimated (e.g. mean, total, proportion…), the reliability criteria (e.g. 20 or 30 percent relative standard error), the Design Effect for the statistics (DEFF defined as the variance inflation factor), and the degrees of freedom for the standard error estimate. For example, consider the minimum sample size to estimate a10 percent prevalence with relative standard error 30 percent or less, a survey DEFF of 1.5, and greater than 16 degrees of freedom for the standard error. The required minimum sample size is 150. Now consider the following simplified example (not real data).

Table1. Sample Size by Data Cycle and Sub-domain

	1999-2000	2001-2002	2003-2004	Combined 4 years	Combined 6 years
Total	210	210	210	420	630
Males	110	110	110	220	330
age < 40	60	60	60	120	180
age > 40	50	50	50	100	150
Females	100	100	100	200	300

In this example, one could estimate the proportion for the total population in each of the 2-year data cycles but none of the sub-domains meets the minimum sample size requirement. Combining the data from two cycles to produce a 4 year dataset (in this case, a 1999-2002 or a 2001-2004 dataset) allows the proportion to be reliably estimated for both males and females. For a more detailed domain however such as Males less than 40 years of age, 6 years of data are required.

Earlier NHANES surveys were conducted for four or more years and, thus, have larger samples than the two-year cycles of the continuous NHANES. However, in each of the NHANES conducted prior to 1999, many sub-domains did not meet minimum sample size requirements and in those cases, the above concerns were (and still are) relevant.

7. When combining two or more two-year cycles of continuous NHANES data, the user should use the following procedure for calculating the appropriate combined sample weights.

When two or more 2-year cycles of the continuous NHANES are combined, the user must calculate new sample weights before analyzing the data. NCHS does not calculate sample and release all possible combinations of multiple two-year cycles of the continuous survey because it would be impractical to produce them and include them on all public release files.

The sample weights for NHANES 1999-2000 were based on population estimates developed by the Bureau of the Census before the Year 2000 Decennial Census counts became available. The two-year sample weights for NHANES 2001-2002 were based on population estimates that incorporate the year 2000 Census counts. The two population estimates were not strictly comparable. Therefore, appropriate four-year sample weights (comparable to Census 2000 counts) were calculated and added to the demographic data files for both 1999-2000 and 2001-2002. The four-year sample weights have the same variable name in each file. For example, the four-year examination sample weight in both files is WTMEC4YR. Thus, users of the earlier release of the NHANES 1999-2000 demographic file must use the updated demographic file to appropriately analyze the combined four-year data 1999-2002. Because NHANES 2003-2004 uses the same year 2000 Census counts as were used for NHANES 2001-2002, there is no need to create special four-year weights for 2001-2004.

For a four year estimate for 2001-2004, one can create a new variable for a four year weight by assigning ½ of the 2 year weight for 2001-2002 if the person was sampled in 2001-2002 or assigning ½ of the 2 year weight for 2003-2004 if the person was sampled in 2003-

2004. This is possible because the 2 year weights for 2003-2004 are comparable to the 2001-2002 weights (in terms of a population basis). For an estimate for the 6-years 1999-2004, a 6-year weight variable can be created by assigning 2/3 of the 4 year weight for 1999-2002 if the person was sampled in 1999-2002 or assigning 1/3 of the 2 year weight for 2003-2004 if the person was sampled in 2003-04. This is possible because the 2003-2004 weights are also comparable (on a population basis) to the combined four-year weights specifically created for 1999-2002.

Summary comments and future additions to the NHANES Analytic Guidelines.

This document summarizes the most recent analytic and reporting guidelines that should be used for most NHANES analyses and publications. It is important for users to understand the entire document and to become familiar with statistical issues in the analysis of complex survey data.

These suggested guidelines provide a framework to users for producing estimates that conform to the analytic design of the survey. Because statistical methods for analyzing complex survey data are continually evolving, these recommendations may differ slightly from those used by analysts for previous NHANES surveys.

It is important to remember that the statistical guidelines in this document are not absolute. When conducting analyses, the analyst needs to use his/her subject matter knowledge (including methodological issues), as well as information about the survey design. ***The more one deviates from the original analytic categories and original analytic objectives defined in the planning documents, the more important it is to evaluate the results carefully and to interpret the findings cautiously.***

Future versions of this NHANES Analytic and Reporting Guidelines will include additional topics, such as sample sizes and response rates for each NHANES survey, hypothesis testing, multivariate analysis, and a discussion of the concept of statistical versus practical significance.

These are guidelines not standards. Depending upon the subject matter and statistical efficiency, specific analyses may depart from these guidelines. The burden of proof for statistical efficiency and for appropriate data interpretation is on the data analyst.

One final reminder for NHANES data users is that the NHANES data files, documentation, and Analytic Guidelines may be edited and/or updated to reflect new information and corrected or edited data. NHANES data users are encouraged to check the NHANES website periodically (available at: http://www.cdc.gov/nchs/about/major/nhanes/NHANES99_00.htm) to determine if new or revised data files and analytic guidelines have been released by NCHS for the data of interest. Data users are encouraged to subscribe to the NHANES listerv (available at: http://www.cdc.gov/nchs/about/major/nhanes/nhaneslist.htm) to receive information updates.