NHANES 1999-2000 Addendum to the NHANES III Analytic Guidelines

Last Update 8/30/02

For the most part, the statistical principles and reliability considerations stated in (http://www.cdc.gov/nchs/data/nhanes/nhanes3/nh3gui.pdf), the NHANES III Analytic Guidelines, can be used for NHANES 1999-2000 data sets. With only 2 years of data in NHANES 1999-2000, instead of the 6-years for NHANES III, there are some additional considerations. First, sample size is smaller and the number of geographic units in the sample is more limited. This increases the potential for inadvertent disclosure of individual participants and results in some limitations to the data release. Second, the sample design, weighting, and variance estimation methodology differ for NHANES 1999-2000. Finally, NCHS has adopted a new standard population for use in age-adjustment procedures.

As users gain experience with the NHANES 1999-2000 data, and as ongoing statistical research efforts become available, additional analytic guidance will be forthcoming. Some current recommendations may also change. Once additional data years are released, for example the 2001-2002 data, additional statistical issues may need to be addressed. The following text provides an addendum to NHANES III Analytic Guidelines and should be used until modified guidelines are published for NHANES 1999-2000.

The organization of the addendum is as follows:

1. Summary of Key Issues
2. Data Limitations and Descriptions of Key Variables
3. Survey, Sample Weights and Variance Estimation
4. Exploratory Data Analysis
5. Age-adjustment and Trend Analysis

1. Summary of Key Issues

NHANES 1999-2000 data files are being released on a periodic basis. Both the data files and associated documentation, including any analytic guidelines, may be edited and/or updated to reflect new data release files. Users should check periodically at (http://www.cdc.gov/nchs/about/major/nhanes/NHANES99_00.htm) to determine if any new or revised data files have been released.

.

The user should read the data file documentation before undertaking any analysis. The documentation will indicate how the data were collected and how the data are coded. For NHANES 1999-2000, 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. 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.

Sample weights should be used for the most appropriate design-based estimation. For NHANES 1999-2000 the use of sampling weights is recommended for all analyses because the sample design incorporates differential probabilities of selection. See section on sample weights for more details.

For complex sample surveys, exact mathematical formulas for sampling errors (variance of estimates) are usually not available. Variance approximation procedures are required to provide reasonable estimates of sampling error. These estimated sampling errors should be produced for all survey estimates to aid in determining statistical reliability. For NHANES 1999-2000, a Jackknife "leave-one-out" (JK1) procedure is recommended and several sets of replicate weights are provided. See section on variance estimation for details.

Recommended sample sizes are the same as those in NHANES III analytic guidelines. But because the overall sample size is smaller, collapsing of demographic subdomains may be required to meet sample size criteria.

Moving from a six-year (that is NHANES III) to a two-year data release, the sample size for the survey is smaller for both number of sampled persons and number of geographic areas (Primary Sampling Units or Stands) sampled. Not only are the data subject to larger sampling variation, but also analytic issues such as the effect of influential sample weights and influential observations become more problematic.

Due to smaller sample sizes in NHANES 99-00, standard errors for a variable in NHANES 1999-2000 will be approximately 70% greater than for the corresponding variable in NHANES III.

Also due to the smaller sample size, survey weights are larger than in past NHANES. Data users should first employ exploratory data analysis on their variables of interest, including examining the data for outliers due to influential values and/or influential weights.

As with any survey data set, care must be used to select the proper survey weight to be used in any specific analysis.   See section on sample weights for specific details.

The sample was designed to provide estimates for the Mexican American population of the United States. Other Hispanics (that is, non Mexican American Hispanics) enter into the sample with different selection probabilities. The sample is not specifically designed to give a nationally representative sample for total Hispanics. Estimates for "All Hispanics" could be problematic and users should interpret such estimates very carefully.

The standard population for age-adjusted rates has been changed. For comparisons of estimates based only on 1999-2000 data, the new year 2000 standard should be used. For comparisons with past data, say NHANES III, any comparison of age-adjusted rates requires that the SAME standard population be used. It is inappropriate to compare an age-adjusted rate from NHANES III based on the 1990 standard to an age-adjusted rate from NHANES 1999-2000 based on the 2000 standard.

2. Data Limitations and Description of Key Variables

The survey data, both interview and examination, are collected from sample persons under a strict promise of confidentiality. The CDC/NCHS Institutional Review Board approves NHANES as a "minimal risk" research protocol on the basis of the strict assurance of confidentiality and the strict review of all data prior to release in order to protect confidentiality.

In order to ensure adequate protection, all NCHS data file releases must be reviewed and approved by the NCHS Disclosure Review Board (DRB).

Any release of a Public Use Micro-data Set (PUMS) creates potential problems with disclosure of confidential information. The goal of NCHS is to release as much data, in the greatest level of detail, to as many data users as possible without compromising confidentiality. The disclosure review includes consideration of the identifiable nature of each data item. Data items that are determined to be direct identifiers (such as name, social security number, home address) cannot be released in any form. Indirect identifiers must be reviewed in the context of survey design. For NHANES 1999-2000, indirect identifiers include age, sex, race, ethnicity, geography, household characteristics, income, occupation and industry classification, survey weights, and any variables that may be linked to an external data source.

A disclosure avoidance review may recommend that a particular data item be released but in limited detail (top or bottom coding) or not released at all (data suppression). For analysis variables, each potential data item is reviewed to determine its impact on the potential for disclosure. For NHANES 1999-2000 data files, due in part to the limited sample size of a two-year release, many indirect identifiers have been deleted or top-coded.

The categories and descriptions for the following selected variables are consistent with the survey design and are recommended for use in analysis, publication, and presentation of the NHANES 1999-2000 data. These categories may be collapsed further for selected analyses, especially when three or more variables are used simultaneously. Any exceptions to these guidelines must be considered very carefully and there should be substantive reasons for choosing other categories. The following list includes SAS labels and variable names from the NHANES 1999-2000 data file documentation.

Age (age in months): SAS variable name RIDAGEMN. This variable gives age in months from infant (2 months) to 234 months of age. This variable is used mostly for applications using the CDC growth charts.

Age (single years of age): SAS variable name: RIDAGEYR. This variable gives age in single years of age. Although single year of age is provided on the data file, the sample sizes for such a detailed age classification are too small and some form of age grouping is required. The following age categories are consistent with the NHANES 1999-2000 sample design age groups and should be used in most analyses.

1 month to 5 years
6-11 years
12-19 years
20-39 years
40-59 years
60 years and older

Age groups used in an analysis should be determined by what is most appropriate for the specific analysis, the sex and/or race-ethnic classification required, the relative magnitude of an estimate (e.g.. a prevalence of 20 percent versus 10 percent), and the relative sampling error of the estimates.

In considering age groups to be used in an analysis, note that some questionnaire items and some examinations are done on a limited age range that may not correspond exactly to the sample design age groups. For adolescents, the household Youth Questionnaire ends at age 16 years and the household Adult Questionnaire begins at 17 years of age. This is another example of why the data file documentation should be consulted before starting any analysis of the data.

Analytic results should be presented in the most meaningful age groups, in the most detail that may be used, and in conjunction with established statistical reliability criteria (such as each estimate must have a relative standard error of 30 percent or less). Note that collapsing of age groups may be necessary to meet statistical guidelines for reliability and precision.

Gender: SAS Variable name RIAGENDR

Male (code 1)
Female (code 2)

Race-ethnicity. Only combined race-ethnicity variables/codes are currently available on the NHANES 1999-2000 data file.

SAS Variable name RIDRETH1

Mexican Americans (code 1)
Other Hispanics (code 2)
Non-Hispanic white (code 3)
Non-Hispanic black (code 4)
Other races including multiracial (code 5)

When using RIDRETH1, the Other Hispanic group (code 2) is available but it should not be used to provide a specific estimate for non-Mexican American Hispanics. Because of the limited number of geographic areas sampled, additional data years of NHANES may be required to provide a representative sample of non-Mexican American Hispanics. Also, the "other" category (code 5) forms an "all other" group that includes other (non-Mexican American) Hispanics, Asians, Native Americans, and those reporting more than one race.

SAS Variable name RIDRETH2

Non-Hispanic white (Code 1)
Non-Hispanic black (Code 2)
Mexican American (Code 3)
All other (code 4)

RIDRETH2 categories are the categories that are most nearly comparable to those recommended for NHANES III.

Race: No variable in current NHANES 1999-2000 public data release. At the present time, there are still unresolved issues in (1) the possible assignment of "missing" race, and (2) the coding structure/disclosure problems for multiple race categories.

Ethnicity: No variable in current NHANES 1999-2000 public data release. At the present time, there are unresolved data disclosure issues in the level of detail that can be released.

Education: SAS variable name DMD140

Never attended or did not complete High school (code 1)
High school or GED (code 2)
Greater than high school (code 3)

Income: No variable in current NHANES 1999-2000 public data release. At the present time, there are data disclosure issues that limit the level of detail that can be released for this variable.

Poverty index (poverty income ratio):

This calculated variable is based on family income and family size using tables published each year by the Bureau of the Census in a series Current Population reports on poverty in the United States. This is the best income variable to use when comparing data over time because it is standardized for inflation and other factors. However, the method of calculation has changed slightly over time. For NHANES 1999-2000, as with all surveys, there are a significant number of persons for whom this variable cannot be calculated because family income was not reported. Two SAS variables for PIR are available at this time.

SAS Variable name: INDPIR1. This variable is the PIR ratio as a continuous variable.

For some analyses, such as the use of the USDA food assistance program (WIC, Food Stamps, School Lunch) an eligibility cut point of 1.850 is used. A categorized SAS variable is provided as follows:

SAS variable name: INDPIR

0.000	(code 1)
0.001-1.000	(code 2)
1.001-1.850	(code 3)
1.851 and above	(code 4)
Refused	(code 7)
Unknown	(code 9)

Geographic (Census) Region: No variable in current NHANES 1999-2000 public data release. At the present time, there are data disclosure issues that prohibit the level of detail that can be released for this variable.

Metropolitan status (MSA): No variable in current NHANES 1999-2000 public data release. At the present time, there are data disclosure issues that prohibit the level of detail that can be released for this variable.

ANALYTIC NOTE ON RECOMMENDED GENDER, AGE, RACE-ETHIC GROUPS.

For descriptive analysis using NHANES 1999-2000 data, the sample size should be sufficient to estimate a 10 percent statistic with a relative standard error not exceeding 30 percent. For most analyses, the following age, gender, and race-ethnic subdomains are recommended in order to meet this sample size criterion.

Age and Gender

Boys <=5 Years
Girls<=5 Years
Boys 6-11 Years
Girls 6-11 Years
Boys 12-19 Years
Girls 12-19 Years
Men 20-39 Years
Women 20-39 Years
Men 40-59 Years
Women 40-59 Years
Men 60 Years And Older
Women 60 Years And Older

Race/Ethnicity, Gender and Age

Non-Hispanic Boys And Girls <=5 Years
Non-Hispanic Black BOYS 6-19 Years
Non-Hispanic Black Men 20 Years And Older
Non-Hispanic Black Girls 6-19 Years
Non-Hispanic Black Women 20 Years And Older
Mexican American Boys And Girls <=5 Years
Mexican American Boys 6-19 Years
Mexican American Men 20 Years And Older
Mexican American Girls 6-19 Years
Mexican American Women 20 Years And Older
Non-Hispanic White And Other Boys And Girls <=5 Years
Non-Hispanic White And Other Boys 6-19 Years
Non-Hispanic White And Other Men 20 Years And Older
Non-Hispanic White And Other Girls 6-19 Years
Non-Hispanic White And Other Women 20 Years And Older

3. Survey Design, Sample Weights and Variance Estimation for HANES 1999-2000

As with previous NHANES surveys, the NHANES 1999-2000 is a complex, multistage probability sample of the civilian non-institutionalized population of the United States. In-home personal interview and Mobile Examination Center (MEC) data are collected on individuals. While the NHANES III survey is designed to be nationally representative for either 3 or 6 years of data collection, the NHANES 1999-2004 survey is designed to give an annual sample that is nationally representative. Note that the current NHANES survey is nationally representative but it is subject to the limits of increased sampling error due to (1) the smaller number of individuals sampled in the annual sample and (2) the smaller number of PSUs available for each annual sample.

For NHANES 1999-2000, the first stage of selection was the PSU-level. The Primary Sampling Units (PSUs) were defined as single counties. For a few PSUs, the county population was too small and those counties were combined with geographically contiguous counties for form a PSU. The sample frame for the NHANES PSUs was the list of PSUs selected for the current design of the National Health Interview Survey (NHIS). The DHHS Survey Integration Plan (Hunter and Arnett, 1996) called for the efficient design of health-related population based surveys through integration with the National Health Interview Survey.

For the current NHIS design, there are 358 PSUs in the annual sample (Botman, et al, 2000). These PSUs are divided into 4 panels with each of the 4 panels comprising a nationally representative sample. In forming the 4 panels, large PSUs are split and the remaining PSUs are stratified according to population size, geographic region, and demographic characteristics. The National Medical Expenditure Survey (NMEP), conducted by the Agency for Health Research and Quality (AHRQ), uses two of the four panels. The remaining two panels are available for use by the NHANES. By splitting the large NHIS PSUs, there are approximately 200 PSUs available in the two national panels for the first stage-sampling frame for the NHANES.

In order to create six annual national samples, 120 of the 200 NHIS PSUS were selected using a measure of size related to 1990 Census county-specific information on the percent Mexican American, percent Black, and the NHIS PSU-selection probability. 20 PSUs were randomly assigned to each year in 1999-2004. For each year, a subset of 15 PSUs was selected with the remaining 5 PSUs held in reserve.

For 1999, due to a delay in the start of data collection, there were only 12 stands (the terminology "stand" refers to the data collection in the Mobile Examination Center within a PSU) for 12 distinct PSUs. Data year 2000 had 15 stands of data collection. For the combined 1999-2000 survey there were 27 stands, but one large PSU was in the survey for both years. For the purpose of variance estimation, the 1999-2000 survey is considered to have 26 PSUs.

Once a PSU was selected, the most current Census information, in this case the 1990 Census, was used to define segments of households. Within PSUs, the percent Mexican American population was used to form four density strata: (1) less than 10 percent, (2) 10-25 percent, (3) 25-60 percent and (4) over 60 percent. In order to achieve a sufficient sample size of non-Hispanic Black and Mexican Americans, within PSU selection probabilities for these domains were adjusted and extensive screening at the household level was required.

For 1999 and the first four stands of 2000, a separate sampling frame for new construction segments (housing built after 1990) was used. Because field experience indicated that use of 1990 Census information (in conjunction with the new construction frame) was inefficient and costly, the last 11 stands in 2000 used a double sampling procedure for the area frame.

For the 26 PSUs (27 stands) in 1999-2000, the final sample consisted of 681 segments. Once a segment was selected, field representatives visited all households in the segment and a screener questionnaire was used to determine sample person eligibility. Individuals were selected into the sample according to fixed sampling fractions. The sampling fractions were adjusted for each of the four segment density strata by the factors 1.0, 1.9, 2.5, and 3.0 (for the corresponding density strata).

Individual sampling fractions were set that distribute the sample into 53 age-sex-race-ethnicity domains for 1999. In order to meet survey objectives related to nutrition, the 2000 sample individual selection probabilities were modified to increase the number of sampled persons in low-income non-Hispanic White population domains. The addition of the low-income sub-domains for non-Hispanic White domains gave 76 age-sex-race-ethnicity-income domains for 2000. Table AG-1 lists the 76 sample domains used in 2000; combing the low-income and non-low income "other" group yields the 53 domains used in 1999.

With 15 PSUs per year, approximately 5,000 sample persons can be examined. Because both the actual duration of data collection as well as the response rates varies by PSU, the actual range for the number of examined sample persons per PSU was from approximately 250 to 400. In sample selection for NHANES 1999-2000, there were 22,839 households screened. Of these, 6,005 households had at least one eligible sample person identified for interviewing. There were a total of 12,160 eligible sample persons identified. Of these 9,965 were interviewed and 9,282 were examined. The overall response rate for those interviewed in 81.9 percent (9,965 out of 12,160) and the response rate for those examined was 76.3 percent (9,282 out of 12,160). Due to confidentiality considerations, the data year is not available on the current NHANES 1999-2000 public use data files.

Constructing Sample Weights.

Because differential probabilities of selection were used in NHANES 1999-2000, it is highly recommended that any statistical inference based on the survey data use the sample weights that are provided on the data file.

The sample weights were calculated from the base probabilities of selection, adjusted for non-response, and post-stratified to match population control totals. For NHANES 1999-2000, weighting adjustments involved multiple levels. Due to the nested levels of data collection (screener, household interview, examination) and to keep the weights from being too variable, a non-response adjustment was applied at each level of data collection, that is, for the screener interview, the household interview and the MEC examination. Post-stratification was applied at each nested level as well.

For the NHANES 1999-2000 file, both the final interview and final examination weights are provided. The interview weight should be used when an analysis uses only data from the household interview. If an analysis uses data from the MEC (MEC interview, examination, or laboratory data on the full MEC sample) exclusively, or in conjunction with the household interview data, the examination weight should be used.

Some MEC components and laboratory results were done only on a subsample of the sampled persons. There was a morning versus afternoon subsample and one-half samples for Balance, Mental Health and Audiometry components. Certain laboratory tests were done only on one-third or one-fourth subsamples. At this time, only full sample data have been released. No subsample data files are on the current data release. When subsamples are released, special survey weights will need to be constructed.

Construction of the full sample weights for NHANES 1999-2000 involved the following steps:

Screener Weight:

(1) For each sample person, the base weight was constructed as the reciprocal of the product of the probabilities of selection at each stage. This involved the initial NHIS PSU selection probability, the annual PSU selection, and the sampling fractions for the 53 (in 1999) or 76 (in 2000) domains of interest.

(2) The base weight was ratio adjusted for new construction, any sub-sampling, and deselection of unused household segments.

(3) A ratio adjustment was applied for screener non-response at the household level.

(4) A post-stratification adjustment was applied based on Census population control totals (shown in Table AG-3) to get the final "screener" weight.

Interview Weight

(5) For sample persons who completed the household interview, the final screener weight was adjusted for household non-response within the adjustment cells formed by the variables race-ethnicity, age, gender and household size. Excessive large ratios were trimmed before application to the weights. The maximum allowable adjustment ratio was 1.35.

(6) After non-response adjustment, a post-stratification adjustment, using the same Census population control totals, was applied to get the final interview weight. For survey estimates based only on household interview data, this weight should be used.

Examination Weight

(7) For sample persons who completed the MEC examinations, the final interview weight was adjusted for MEC non-response within the adjustment cells formed by race-ethnicity, age, sex, household size, household education, self-reported health status, and length of stay at current residence. The adjustment cells were collapsed to ensure at least 25 examined sample persons per cell, and the maximum adjustment ratio was 1.35.

(8) Excessive weights were then trimmed. Only 4 cases required trimming at this stage.

(9) A post stratification adjustment, using the same Census population control totals, was then applied to get final examination weight.

Variance Estimation for NHANES 1999-2000

Past NHANES data files have included stratification and PSU variables that could be used to calculate estimates of sampling error. These variables typically define a two-PSU per stratum classification and various estimation methods (BRR, Jackknife, Taylor series) and survey-specific software procedures (for example, see references for STATA, SAS, WESVAR, and SUDAAN) can be used to compute sampling errors. For NHANES 1999-2000, PSU variables cannot be released due to disclosure protection. Lack of design information on the data file creates a problem for calculating sampling errors.

To estimate sampling errors for NHANES 1999-2000, a Jackknife procedure can be used, specifically the “leave-one out” or JK-1 procedure (Wolter, 1985 Chapter 4; see also Rust, 1985 and Rust and Rao 1996). The methodology employed created groups of sampled individuals in such a manner as to preserve the basic design structure without disclosing geographic identity. NHANES 1999-2000 sampled individuals were aggregated into 52 groups. Deleting one of the 52 groups and re-weighting the remaining 51 groups formed replicate weights. For any particular set of replicate weights, sampled persons in the “left-out” group have zero weight. The re-weighting used the same methods for non-response and post-stratification adjustments that were used to create the original estimation weights. The order of the replicates on the data file has been randomized such that data year and original PSU structure are obscured.

The 52 replicate weights were constructed for both the final interview weights (SAS variable names WTIREP1 to WTIREP52) and for the final examination weights (SAS variables names WTMREP1 to WTMREP52). At this time only the software packages WESVAR and SUDAAN can use a JK-1 technique.

For SUDAAN, for variables based on the examination data, the two design statements required are:

```
WEIGHT WTMEC2YR;

JACKWGTS WTMREP01-WTMREP52/ADJJACK=.980769;
```

Here, SUDAAN requires the factor 51/52 = 0.980769 for the JK-1 procedure.   Two sample SUDAAN programs for using the Jackknife procedure are attached at the end of this document.

Preliminary research has indicated that the jackknife method with the 52 replicates yields estimates that are slightly smaller compared with the traditional sampling error estimates. In particular, a variable with a large design effect and a relatively large between-PSU component of the sampling error may be underestimated. Research efforts are ongoing that will determine if a correction factor is needed for the jackknife sample error estimates. At this time, the user is cautioned that the 52 replicate jackknife method may underestimate some sampling errors. Data users are cautioned to be careful when publishing results that are marginally significant based on the JK-1 variance estimates.

Sample Size Considerations

The basic statistical considerations discussed in the NHANES III Analytic Guidelines also apply to NHANES 1999-2000.   For the same or comparable variables in both surveys, design effects (DEFF) for NHANES 1999-2000 are very similar to the DEFFs for NHANES III because the level of clustering, the intra-class correlation coefficients, and the heterogeneity among the sample weights are similar. However, the sample sizes for the two years of NHANES 1999-2000 are smaller than sample sizes for the 6 years of NHANES III. A quick, but rough approximation, is that the relative standard errors are typically inflated by a factor equal to the square root of three, or about 1.7 (that is, the standard errors for a variable in NHANES 1999-2000 should be roughly 70 percent greater than the corresponding variable in NHANES III).

Within the NHANES survey, DEFF can be very different for different variables due to differences in variation by geography, by household intra class correlation, and by demographic heterogeneity. Because DEFFs are highly variable for NHANES 1999-2000 estimates, it is difficult to set a single minimum sample size for analysis. The general statistical consideration is that an estimate should have a relative standard error of 30 percent or less. The NHANES III Analytic Guidelines contain sample sizes required for reliable estimates and for testing differences between subdomain. The required sample size depends on the DEFF for the variable of interest. These sample size tables provide guidance, but at this time it is best to compute an estimate for the sampling error of a statistic and use a reliability cut-point such as 30 percent relative standard error.

4. Exploratory Data Analysis

Before analyzing the NHANES data, users should perform simple exploratory analyses to evaluate frequency distributions of the observed data, identify potential outliers, and evaluate the extent of missing data. Occasionally, extremely large measurement values (which may be valid values) with very large sampling weights can have significant effect on estimates and conclusions. As a general practice, such outliers should be reported. Analysts should use their subject-matter knowledge to decide whether to include, trim, or exclude these outliers in their analyses. When evaluating the extent of missing data, if a large proportion of data is found to be missing, analysts should decide if further adjustments are needed to compensate for missing information (Kalton and Kasprzyk, 1986; Little and Rubin, 1987, and Groves, et al 2002).

Preliminary steps in data analysis

1. Read all relevant data file documentation, examining the questionnaires, examination protocols, and the data file codebook. This includes determining which variables were obtained in the home interview and which were obtained in the MEC exam, determining for which subdomains the variable of interest was collected (for example serum folate was obtained for persons 3 years and older), and determining skip patterns in the questionnaire data.
2. Run frequency distributions for discrete variables. This will provide a check of the use of valid codes as well as determining the extent of missing data for each particular data item.
3. Run simple statistics (mean, standard error, range) and a normal probability plot on continuous variables to check for skewness and kurtosis.
4. For continuous variables plot the sample weight against the variable of interest for the subgroups of interest to check for influential observations. For binary variables run frequency distributions on the sample weights for those with versus those without the characteristic.
5. If possible, compare findings from NHANES 1999-2000 with NHANES III or with data from other sources.

5. Age-adjustment and trend analyses

Age-adjustment is important for trend analyses between NHANES surveys and for comparisons between subgroups within NHANES 1999-2000. It is also important to include, when possible, age-specific estimates along with age-adjusted estimates in any publication. If it is not possible to report both sets of data in a publication, the choice of crude (or age-specific) versus age-adjusted data should be made based upon the primary focus of the analysis. If a statistic of interest varies substantially by age within race-ethnic categories, the age-standardized estimates will be more appropriate. For comparison of age-adjusted statistics within and between NCHS surveys, the 2000 Census population should be used as the standard population (Klein and Schoenborn, 2001).

The following standard proportions are based on the 2000 standard population and should be used in NHANES 1999-2000 analyses when using 20 year age groups for 20 years and older.

Age Group	Proportion
20-39	0.3966
40-59	0.3718
60 +	0.2316

Past NHANES surveys did not have sample persons at ages 75 years and over. To compare age-adjusted (ages 20-74 years only) statistics for NHANES 1999-2000 with past NHANES surveys, the following standard proportions should be used:

Age Group	Proportion
20-39	0.4332
40-59	0.4062
60-74	0.1606

In the SUDAAN software, these proportions are used with statements STDVAR and STDWGT, where STDVAR lists the name of the variable with age categories used in standardization and STDWGT lists the corresponding proportions from the year 2000 Census.

REFERENCES:

Hunter EL and Arnett R. (1996). Survey “Reinventing” at Health and Human Services. Chance 9:54-57.

Botman SL, Moore TF, Moriarity, CL, and Parsons, VL (2000) Design and Estimation for the National Health Interview Survey, 1995-2004. National Center for Health Statistics. Vital and Health Stat 2(130).

STATA (1996) Statistical Software Release 5.0 College Station TX. Stata Corporation. (www.stata.com)

Shah BV, Barnwell BG and Bieler GS (1995). SUDAAN User's Manual: Software for the Statistical Analysis of Correlated Data. Research Triangle Park, NC, Research Triangle Institute (www.rti.org/page.cfm?sec=4)

WESVAR 4.0 users guide (2000) Westat Inc (www.westat.com/wesvar)

SAS Institute Inc (1989) SAS/STAT Users Guide Version6, Fourth Edition, SAS Institute, Cary NC (http://www.sas.com/products/stat/index)

Wolter K. (1985). Introduction to Variance Estimation. Springer-Verlag, New York.

Rust KF (1985). Variance Estimation for Complex Estimators in Sample Surveys. Journal of Official Statistics. 1:381-397

Rust KF, and Rao JNK (1996) Variance estimation for complex survey data using replicate methods Statistics in Medical Research 5:283-310

Groves RM, Dillman DD, Eltinge JL, and Little RJA (2002) Survey Nonresponse Wiley series in Probability and Statistics, New York.

Kalton G and Kasprzyk D (1986). The treatment of missing survey data. Survey Methodology 12:1-16

Little RJA and Rubin DB, 1987 Statistical Analysis with Missing Data Wiley:New York

Klein RJ and CA Schoenborn 2001 Age adjustment using the 2000 projected US population. Healthy People 2010 Statistical Note Number 20 ,Jan 2001

Table AG-1. NHANES 1999-2000 sampling domains

Black	Mexican American	Low income other	Non-low income other
Males and females 0-11 months	Males and females 0-11 months	Males and females 0-11 months	Males and females 0-11 months
Males and females 1-2 years	Males and females 1-2 years	Males and females 1-2 years	Males and females 1-2 years
Males and females 3-5 years	Males and females 3-5 years	Males and females 3-5 years	Males and females 3-5 years
Males 6-11 years	Males 6-11 years	Males 6-11 years	Males 6-11 years
Males 12-15 years	Males 12-15 years	Males 12-15 years	Males 12-15 years
Males 16-19 years	Males 16-19 years	Males 16-19 years	Males 16-19 years
Males 20-39 years	Males 20-39 years	Males 20-29 years	Males 20-29 years
		Males 30-39 years	Males 30-39 years
Males 40-59 years	Males 40-59 years	Males 40-49 years	Males 40-49 years
		Males 50-59 years	Males 50-59 years
Males 60+ years	Males 60+ years	Males 60-69 years	Males 60-69 years
		Males 70-79 years	Males 70-79 years
		Males 80+ years	Males 80+ years
Females 6-11 years	Females 6-11 years	Females 6-11 years	Females 6-11 years
Females 12-15 years	Females 12-15 years	Females 12-15 years	Females 12-15 years
Females 16-19 years	Females 16-19 years	Females 16-19 years	Females 16-19 years
Females 20-39 years	Females 20-39 years	Females 20-29 years	Females 20-29 years
		Females 30-39 years	Females 30-39 years
Females 40-59 years	Females 40-59 years	Females 40-49 years	Females 40-49 years
		Females 50-59 years	Females 50-59 years
Females 60+ years	Females 60+ years	Females 60-69 years	Females 60-69 years
		Females 70-79 years	Females 70-79 years
		Females 80+ years	Females 80+ years

Table AG-2. Number of interviewed, and MEC-examined SPs in NHANES 1999-2000 by collapsed domain

Collapsed race/ethnicity-sex-age domain	Number of interviewed SPs	Number of MEC-examined SPs	
White/other			
Male/female < 6 years	593	551	
Male 6-19 years	504	473	
Male 20+ years	1,261	1,137	
Female 6-19 years	527	493	
Female 20+ years	1,413	1,246	
Black, non-Hispanic			
Male/female < 6 years	342	324	
Male 6-19 years	516	496	
Male 20+ years	418	394	
Female 6-19 years	492	479	
Female 20+ years	505	467	
Mexican American			
Male/female < 6 years	622	586	
Male 6-19 years	768	745	
Male 20+ years	590	543	
Female 6-19 years	721	689	
Female 20+ years	693	657	
Total	9,965	9,282	

Table AG-3. CPS Population Control Totals for NHANES 1999-2000 Sample Domains

Sex	Age	Black	Mexican American	Sex	Age	Other
Male/Female	Less than 1 Year	545,640	508,219	Male/Female	Less than 1 Year	2,784,812
	1-2 Years	1,182,576	1,023,460		1-2 Years	5,613,331
	3-5 Years	1,911,787	1,580,996		3-5 Years	8,471,605
Male	6-11 Years	2,062,092	1,381,641	Male	6-11 Years	9,163,511
	12-15 Years	1,271,153	819,511		12-15 Years	5,999,593
	16-19 Years	1,228,275	843,178		16-19 Years	6,146,171
	20-39 Years	4,783,380	3,813,243		20-29 Years	13,586,247
					30-39 Years	16,215,536
	40-59 Years	3,599,250	1,803,824		40-49 Years	16,981,601
					50-59 Years	12,105,600
	60 Years and over	1,523,583	598,025		60-69 Years	8,158,706
					70-79 Years	5,971,108
					80 years and over	2,533,493
Female	6-11 Years	1,997,694	1,310,619	Female	6-11 Years	8,752,266
	12-15 Years	1,238,128	708,334		12-15 Years	5,777,273
	16-19 Years	1,207,389	764,444		16-19 Years	5,859,491
	20-39 Years	5,872,837	3,485,603		20-29 Years	13,818,585
					30-39 Years	16,644,536
	40-59 Years	4,375,742	1,799,971		40-49 Years	17,410,477
					50-59 Years	12,786,768
	60 and over	2,245,019	704,754		60-69 Years	8,960,783
					70-79 Years	7,872,911
					80 years and over	4,352,060

Sample SUDAAN Program for Descriptive Statistics

```
****PRODUCES MEANS, STANDARD ERROR OF MEANS,***;
***95 PERCENT CONFIDENCE LIMITS AND DESIGN*****;
****STANDARD EEVIATIONS WHICH ACCOUNT FOR THE COMPLEX
SAMPLE*****************;
libname in 'c:\Documents and Settings\';
libname in1 'c:\Documents and Settings\userid\My Documents\';
***This portion of the program creats a file*********;
***for input into PROC DESCRIPT**********************;
***by merging the file containing the analytic*******;
***variable of interest (in this case serum total****;
*** cholesterol) with the demographic*********;
***file containing the design variables*************;
***i.e. the full sample weight and the 52 replicate**;
*** weights as well as age and gender******************;
data demo;
set in.demo;
proc sort;by seqn;
data CHOL;
set in1.LAB13;
proc sort;by seqn;
proc format;value s 1='Men' 2='Women';
data comb;
merge demo CHOL;by seqn;
label LBXTC='Serum total cholesterol'
RIAGENDR='Gender';
STRATUM=1;
age1=1+(ridageyr>39)+(ridageyr>59);
run;
proc descript data="comb" design=jackknife deft;
subpopn RIDAGEYR>=20 & WTMEC2YR>=1;
weight wtmec2yr; */Use wtint2yr for
 variables based on interview data;
jackwgts wtmrep01-wtmrep52/adjjack=.980769;
*/Use wtirep01-wtirep52 for variables based on interview
data;
subgroup age1 RIAGENDR;
levels      3    2;
var LBXTC;
table RIAGENDR*age1;
print nsum mean semean
deffmean/style=NCHS nsumfmt=f8.0 meanfmt=f8.0
semeanfmt=f8.1;
rtitle "Mean serum total cholesterol of adults 20 years
 of age and older: United States, 1999-2000";
output nsum mean semean deffmean/filename=cholmean;
rformat riagendr s.;
run;
proc sort data=comb;
by stratum sdj1repn;
proc descript atlevel1=1 atlevel2=2;
subpopn RIDAGEYR>=20 & WTMEC2YR>=1;
nest stratum sdj1repn;
weight wtmec2yr; */Use wtint2yr for
```

```
 variables based on interview data;
subgroup age1 RIAGENDR;
levels      3    2;
var LBXTC;
table RIAGENDR*age1;
print nsum mean semean atlev1 atlev2
/style=NCHS nsumfmt=f8.0 meanfmt=f8.0
semeanfmt=f8.1;
rtitle "Mean serum total cholesterol of adults 20 years
 of age and older: United States, 1999-2000";
output atlev1 atlev2/filename=choldf;
rformat riagendr s.;
run;
proc means data=comb;
where ridageyr>=20 and wtmec2yr>=1;
var lbxtc;freq wtmec2yr;
output out=tempsdt std=std;
data tempsdt;set tempsdt;riagendr=0;
PROC SORT DATA=COMB;BY RIAGENDR age1;
PROC MEANS;
where ridageyr>=20;
VAR lbxtC;BY RIAGENDR age1;freq wtmec2yr;
OUTPUT OUT=TEMPSD STD=STD;
data tempsd;set tempsdt tempsd;
DATA RESULT;MERGE CHOLMEAN CHOLDF TEMPSD; BY RIAGENDR;
n1=atlev2-1;
tlow=tinv(.025,n1);
tup=tinv(.975,n1);
lowcl=round(mean+tlow*semean);
upcl=round(mean+tup*semean);
mean=round(mean);
semean=round(semean,.1);
csstd=sqrt(semean**2+std**2);
csstd=round(csstd,.1);
proc print split='/';
var riagendr nsum mean semean n1 csstd lowcl upcl
deffmean;format riagendr s. nsum 6.0 mean 5.0 semean 4.1
deffmean 5.2 lowcl 7.0 upcl 6.0 csstd 5.1
;label semean='se'/'of the'/'mean'
csstd='Complex'/'sample'/'Std'/'Dev'
n1='df'
deffmean='Design'/'effect' riagendr='Gender'
lowcl='Lower'/'95 %'/'CL' upcl='Upper'/'95 %'/'CL';
title1 'Mean serum total cholesterol of adults 20 years';
title2 'of age and older by gender: United States, 1999-2000';
run;
```

Sample SUDAAN Program for Age Standardization

```
***SUDAAN Program to estimate age standardized**;
****means and prevalences based on NHANES 1999-2000**;
****data. Estimates are stanardized to the*****;
***year 2000 Census population estimates*******;
************************************************;
libname in1 'c:\documents and settings\';
proc format;value s 1='Men' 2='Women';
value a 1='20-39 years' 2='40-59 years'
  3='60-74 years';
data chol;
set in1.chol;
label LBXTC='Serum total cholesterol'
RIAGENDR='Gender' age1='Age group'
hightc='>=240 mg/dl';
age1=1+(ridageyr>39)+(ridageyr>59);
if lbxtc>=240 then hightc=100;
else if lbxtc^=. then hightc=0;
proc freq;
where ridageyr>=20 & ridageyr<=74 & wtmec2yr>=1;
table hightc riagendr age1 ridageyr
age1*ridageyr/list missing;
format age1 a.;
proc sort data=chol;by hightc;
proc univariate noprint;var lbxtc;by hightc;
output out=temp n=n  min=min max=max;
proc print label;var hightc min max;
proc descript data="CHOL" design=jackknife;
subpopn RIDAGEYR>=20 & RIDAGEYR<=74 & WTMEC2YR>=1;
weight wtmec2yr; */Use wtint2yr for
 variables based on interview data;
jackwgts wtmrep01-wtmrep52/adjjack=.980769;
*/Use wtirep01-wtirep52 for variables based on interview
data;
subgroup age1 RIAGENDR;
levels     3       2;
stdvar age1;
stdwgt .4332 .4062 .1606;
var hightc;
table RIAGENDR;
print nsum mean="Percent"
semean="se percent"/style=NCHS nsumfmt=f8.0
meanfmt=f8.1 semeanfmt=f8.2;
rtitle "Age standardized prevalence of high serum total
 cholesterol of adults 20-74 years of age:
 United States, 1999-2000";
rfootnote "Age adjusted by the direct method to the year
 2000 Census population projections using the age groups
 20-29 years, 30-39 years, 40-49 years, 50-59 years and
 60-74 years.
NOTE: High serum total cholesterol is defined as a value of at
     least 240 mg/dl.";
rformat riagendr s.;
run;
```