Analyzing the NHANES III Multiply Imputed Data Set: Methods and Examples

Prepared for:

National Center for Health Statistics
Hyattsville, Maryland

Prepared by:

Joseph L. Schafer

June, 2001

Author's academic affiliation: Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802. Direct comments and queries to jls@stat.psu.edu.

Summary

The NHANES III Multiply Imputed Data Set provides a new method for handling missing data in many analyses of NHANES III. Missing values for key variables were imputed five times, producing five simulated complete data files distributed on CD-ROM. This document describes recommended methods for analyzing the NHANES III Multiply Imputed Data Set. Estimates and standard errors must be computed five times, once for each of the completed data files, using techniques that take into account the NHANES III complex sample design. The five sets of estimates and standard errors are then combined in a straightforward manner to produce a single set which accounts for missing-data uncertainty in addition to ordinary sampling variability. Example analyses are provided in SAS and SAS-callable SUDAAN. Results from this new procedure are compared to those from conventional analyses of previously released NHANES III data sets (DHHS, CD-ROM, Series 11, Number 1A, 1997; Number 2A, 1998).

1 Introduction

1.1 Nonresponse in NHANES III

The third National Health and Nutrition Examination Survey (NHANES III) experienced moderate rates of nonresponse at each stage of the data collection process. Previously released data sets from NHANES III (DHHS, CD-ROM, Series 11, Number 1A, 1997; Number 2A, 1998) provided sample weights that include adjustment factors for different types of nonresponse. One adjustment corrects for biases arising from differential rates of participation by sampled persons in the household interview. A second adjustment corrects for biases arising from different rates of participation in the physical examination in the Mobile Examination Center (MEC). Methods for weighted estimation and procedures for calculating standard errors have been described in *NHANES III Reference Manuals and Reports* (DHHS, 1996). Details of NHANES III data collection procedures are available in *Plan and Operation of the Third National Health and Nutrition Examination Survey*, 1988–94, (DHHS, 1994, 1996).

Weighting methodology for nonresponse is convenient because it does not require the data user to perform any sophisticated computations beyond those already needed for weighted estimation. However, weighting methods are limited in a number of respects. First, these methods were designed primarily for *unit nonresponse*, which occurs when the sampled person does not respond to any items on the survey instrument (e.g. by refusing to participate in the survey). Weighting methods are not effective for handling *item nonresponse*, the intermittent missing values that arise when sampled persons respond to some but not all of the survey items. Despite the weighting adjustments that were made, the NHANES III public-use data files still contain non-trivial amounts of missing values on many items. Data analysts typically ignore the missing values, calculating estimates from a reduced set of individuals by various 'complete case' or 'available case' procedures (Little and Rubin, 1987). Case-deletion methods may introduce bias if the response rates for individual items vary across subgroups. Case deletion may also make it difficult for one data analyst to precisely replicate the results published by another analyst, because different rules for discarding incomplete cases often create ambiguity about which portion of the sample should be used for a particular analysis.

Another shortcoming of weighting methods is that they may ignore valuable information contained in inter-variable relationships which could be used to make accurate predictions of the missing data values. Weighting adjustments were designed to correct for biases arising when rates of unit nonresponse vary by subgroups. But the weights were not designed to produce optimal estimates of population characteristics for any particular survey variable. In many cases, auxiliary information from observed variables correlated with the missing items is not taken into account when the weights are adjusted, making the resulting estimates inefficient (Little, 1986).

Finally, unless special corrective measures are taken, variance estimates obtained from adjusted weights may not reflect the extra degree of uncertainty introduced by the uncontrolled nonresponse process. This understatement of uncertainty may lead to standard errors that are downwardly biased and interval estimates that cover their population targets with lower-than-nominal rates of coverage. Techniques for variance estimation from the NHANES III public-use files have been described in the reports 'Weighting and estimation methodology,' and 'Analytic and reporting guidelines,' both available in *NHANES III Reference Manuals and Reports* (DHHS, 1996). Those documents describe two methods for calculating standard errors from NHANES III data: a linearization approach and a method based on replicate weights. The former method makes no allowance for the effect of weighting adjustments for nonresponse, whereas the latter does. Neither of these techniques, however, accounts for the effects of uncontrolled item nonresponse either in estimating a population quantity or in calculating a standard error for the estimate. This is not a shortcoming of these estimation procedures *per se*, because the procedures were not designed to handle incomplete survey data. Rather, it is an artifact of applying these procedures to a deficient data set containing non-trivial rates of missing values.

1.2 The NHANES III Multiply Imputed Data Set

Responding to new developments in statistical methods for missing data, the National Center for Health Statistics assembled a team of researchers to investigate alternatives to the conventional weighting methods used in NHANES III. This research effort evolved into the NHANES

III Multiple Imputation Research Project, culminating in the release of the NHANES III Multiply Imputed Data Set on CD-ROM. Multiple imputation is a simulation-based approach to missing data in which each missing value is replaced by $M > 1$ plausible values generated by a statistical model, resulting in M different but equally plausible versions of the complete data set (Rubin, 1987, 1996). Each version is analyzed separately in the same manner, and the results from the M analyses are combined by simple rules to produce estimates, standard errors and confidence intervals that incorporate missing-data uncertainty. Five versions of the complete data are distributed in the NHANES III Multiply Imputed Data Set CD-ROM. The decision to create and distribute $M = 5$ versions was made based on pilot studies and exploration of the rates and patterns of missing information on important survey items.

Details of the statistical models and computational methods used to create the multiple imputations are described in the companion report 'Multiple imputation models and procedures for NHANES III.' These procedures were designed to impute values with distributional characteristics similar to the data actually observed for each variable, both overall and within important demographic subgroups. The imputation procedures were also designed to preserve important relationships among NHANES III variables, so that more complicated analyses (e.g. regression modeling) involving groups of variables could accurately estimate these relationships. Finally, the imputation procedures were designed to reflect appropriate levels of missing-data uncertainty in the individual survey items on a case-by-case basis. For example, consider an examined person with a missing value for a single body measurement (e.g. waist circumference) but recorded values for all other body measurements. Because the various body measurements in NHANES III are highly correlated, the recorded values for the individual's other measurements can be used to predict the missing measurement quite accurately; as a result, the imputed values for the missing measurement will exhibit relatively little variation across the $M = 5$ data sets. On the other hand, if an individual has no recorded values for any body measurements, then the imputed values will exhibit greater variation across the $M = 5$ data sets, roughly comparable to five sets of measurements randomly sampled from a population of persons of similar age, race/ethnicity and gender.

1.3 Uses of the multiply imputed data

The NHANES III Multiply Imputed Data Set provides an improved method for handling missing values in many but not all analyses of NHANES III. It is intended as a companion to, but not a replacement for, the previously released NHANES III public-use data files (DHHS, CD-ROM, Series 11, Number 1A, 1997; Number 2A, 1998). Users of NHANES III data are encouraged to analyze the new multiply imputed files using the methods described in this document. National estimates and standard errors calculated by these new procedures may differ somewhat from those obtained from previously released public-use files because a different treatment has been applied to missing values.

The statistical theory underlying multiple imputation (Rubin, 1987) and a large simulation study (Little et al., 1995) suggest that the procedures used to create the NHANES III Multiply Imputed Data Set produce high-quality population estimates and accurate standard errors over repeated application. The new method is thought to have significant advantages over reweighting in adjusting for nonresponse at the MEC examination stage. Gains in precision are apparent particularly in some examination variables for persons who were interviewed but not examined (Little and Rubin, 1992). The standard errors from the multiple-imputation method explicitly account for the additional uncertainty introduced by missing values.

The imputation procedures used to create the NHANES III Multiply Imputed Data Set were designed to be compatible with many common analytic techniques including the estimation of prevalences, means, quantiles, linear and logistic regression coefficients. No imputation procedure, however, can effectively solve the missing-data problem for all potential future analyses by all data users. Users of the NHANES III Multiply Imputed Data Set should be aware of the basic properties of the imputation models and their primary strengths and limitations.

One key feature of the imputation models is that they are based upon an assumption of multivariate normality; that is, they assume that the variables whose missing values are to be imputed are jointly normally distributed within demographic subgroups defined by age, sex, and race/ethnicity, and primary sampling unit. Some variables that consist of discrete

categories (e.g. self-reported health status, which takes values from 1 = excellent to 5 = poor) were modeled and as if they were normally distributed, and the continuous imputed values were rounded off to the nearest category. Other variables whose distributions were skewed were transformed by standard power functions such as the logarithm, square root, or reciprocal square root; modeling and imputation were carried out on the transformed data, and after imputation they were transformed back to the original scale. Some variables whose distributions were especially problematic were transformed by a method based on the empirical cumulative distribution function (cdf), forcing them to approximate normality. This empirical cdf method preserves distributional shape quite well in an overall sense, but tends to produce duplication of extreme values rather than a smooth continuum in the tails. Any of these transformation methods may fail to accurately describe the extreme tail behavior for some variables. For this reason, the NHANES III Multiply Imputed Data Set should not be used for statistical analyses that are sensitive to extreme values, e.g. estimation of a 98th percentile. For analyses that are less sensitive to tail behavior—e.g. the estimation of means, medians, quartiles, or 10th and 90th percentiles—the imputation procedure is expected to perform well.

Data users should also understand that a multivariate normal imputation model is capable of preserving fairly simple relationships among variables including simple correlations and partial correlations, but more complicated relationships (e.g. curvilinear relationships and three-way associations) are not supported. As a result, some complex associations among variables may have been dampened by the imputation procedure, which may adversely affect certain types of statistical analyses. For example, in regression modeling, one may be interested in measuring interactions. An interaction occurs when the influence of a predictor on the response varies by the levels of another predictor. The normal model underlying the imputation procedure does not preserve interactions among most variables, so power to detect interactions (i.e. the probability that an interaction will be deemed 'statistically significant') may be substantially reduced, particularly in regions of the data where nonresponse rates are high. Users should be aware that some interactions in the completed data may be smaller than they otherwise would have been if no data had been missing. One notable exception,

however, is that the imputation models were designed to preserve two and three-way interactions among crucial demographic variables (gender and race/ethnicity). Moreover, because separate imputation models were fit to classes defined by age, interactions between age and other variables will be preserved as well.

Finally, only a modest number of NHANES III variables could be included in the imputation models. The largest of these models involved about 100 variables from the NHANES III screener, household interview, and examination. Imputations produced under a statistical model will not reflect potential relationships with variables excluded from that model. For this reason, users are advised not to use the NHANES III Multiply Imputed Data Set to analyze relationships between variables in this data set and non-imputed variables extracted from other NHANES III public-use data files; doing so could result in underestimation of the strength of these relationships.

1.4 Comparing results with those of previous methods

We encourage users of the NHANES III data not only to analyze the NHANES III Multiply Imputed Data Set by to the methods described in this document, but to compare the results to those obtained by conventional analyses of the previously released NHANES III public-use files (DHHS, CD-ROM, Series 11, Number 1A, 1997; Number 2A, 1998). Examples of such comparisons are provided in Section 4. In some analyses, the estimates and standard errors from the two methods may appear to be similar. In other analyses, differences may arise, particularly in subgroups where rates of missing values are high.

Similarities among results from the two methods are inevitable because the non-missing values which make up the major part of both CD-ROM data sets are identical. But even though the two methods may lead to similar results in a particular application, the methods are not equivalent and they do have different statistical properties over repeated use. In some applications, the conventional estimates will be less efficient (i.e. have greater variability) than the multiple-imputation estimates because the adjusted weights are not using covariate information as effectively as the imputation methods are. For this reason, a standard error which would accurately reflect the variability of the conventional estimate ought

to be somewhat larger than a standard error obtained from the multiple-imputation estimate. However—depending on the variance estimation method being used—the standard error actually computed for the conventional estimate may tend to understate that estimate's true variability because missing-data uncertainty may not be accounted for properly. These two effects—a less efficient population estimate combined with a downwardly biased variance estimation procedure—may sometimes appear to cancel out, causing the standard errors for the conventional estimates to resemble the standard errors from the multiple imputation method in a single analysis. But over repeated application, the two methods would not have identical properties. Under the conventional method, confidence intervals would miss their population target values more often than they should, and conventional decision rules for hypothesis testing would produce Type I errors (false rejections of null hypotheses) more often than they should.

In addition to improved statistical properties, multiple imputation offers the user some operational advantages as well. Imputation helps to remove ambiguity about which subset of the sample ought to be used for any particular analysis. It is no longer necessary to discard cases which are missing one or more variables needed for an analysis; one simply uses the entire sample each time. Finally, imputation also helps to reduce confusion over which survey weight to use for a particular analysis. With the previously released NHANES III public-use files, data users were advised to use one weight for analyses involving items from the household questionnaires, and another weight for analyses involving items from the physical examination. Users of the NHANES III Multiply Imputed Data Set, however, should simply use the 'final interview weight' (`WTPFQX6`) for all estimation procedures.

1.5 Scope of the rest of this document

Section 2 describes the recommended procedures for analyzing the NHANES III Multiply Imputed Data Set. Some examples of typical analyses are provided in Section 3 in SAS and SAS-callable SUDAAN (Research Triangle Institute, 1998). These examples illustrate the estimation of means, percentages, medians, and percentiles for the entire population of adults and for subclasses defined by sex and age. A sample program is also provided for

estimating logistic regression coefficients (log odds ratios). Finally, Section 4 presents some comparisons between results obtained from the NHANES III Multiply Imputed Data Set and from conventional analysis of the previously released public-use data files.

2 Recommended procedures for analysis

2.1 Overview of analysis procedures

Analyzing a multiply imputed data set is similar to analyzing a conventional data set with no missing values. Most statistical procedures that would be appropriate for the full NHANES III data will be appropriate for the NHANES III Multiply Imputed Data Set, subject to the limitations discussed in Section 1.3. The only major difference is that any estimation procedure must be carried out five times, once for each version of the completed data. As the speed, memory, and data storage capacity of modern computers continue to rapidly expand, performing an identical analysis five times rather than once is not expected to impose an undue burden on most data users.

Because of the complex survey design used in NHANES III, traditional methods of statistical analysis based on the assumption of a simple random sample may not be reliable. Sample weights are needed to produce correct estimates of population quantities. Other aspects of the sample design (e.g. PSU pairings) should be taken into account to obtain correct standard errors and significance levels for hypothesis tests. Use of special computer programs for data from complex samples, such as SUDAAN (Research Triangle Institute, 1998) or WesVarPC (Westat, 1996) is strongly recommended. Appropriate methods for the analysis of data from NHANES III are described in *NHANES III Reference Manuals and Reports* (DHHS, 1996). Users of the NHANES III Multiply Imputed Data Set should, for the most part, follow the guidelines given for analysis of those public-use files. The only differences pertain to the handling of missing values (which are no longer an issue because they have been imputed) and the choice of survey weight. Procedures for weighted estimation and the calculation of standard errors from each of the five completed data sets are described in Section 2.2 below. Methods for combining the five sets of results are given in Section 2.3.

In rare instances, users may need to merge the NHANES III Multiply Imputed Data Set with information from other NHANES III public-use files. Any merging of records across files should make use of the common sequence identification number variable (`SEQN`). Joint analyses involving variables in the NHANES III Multiply Imputed Data Set and other variables may not be valid, as described in Section 1.3.

2.2 Obtaining estimates and standard errors from each completed data set

NHANES III is a stratified, multistage area sample of the noninstitutionalized civilian U.S. population, with oversampling of young children (under 5), the elderly (60+), Mexican Americans and African Americans. Data were collected in two 3-year phases known as Phase 1 (1988-1991) and Phase 2 (1991-1994). Each of these Phases individually is a national probability sample, but analysts are encouraged to combine them and use all six years of survey data. Because of the complex design, unweighted summary statistics will not in general produce estimates representative of the national population. Users are strongly encouraged to apply weighted estimation procedures using the sample weights provided in the NHANES III data files. For example, if y_i represents a measurement of a numeric variable for subject i, the mean for all individuals in a given domain $\mathcal{D}$ would be estimated by

$$\bar{y} = \frac{\sum_{i \in \mathcal{D}} w_i y_i}{\sum_{i \in \mathcal{D}} w_i}, \tag{1}$$

where w_i denotes the weight given to subject i. The value of the weight indicates how many 'population persons' are represented by the sampled person. If NHANES III had been a simple random sample of one out of every 6,000 Americans, then each sample weight would have been $w_i = 6,000$ and (1) would reduce to an ordinary sample mean. But because of the oversampling used in NHANES III, the sample weights do vary considerably and hence should be taken into account.

The idea of weighted estimation can be extended to many types of population quantities. For example, suppose one needed to estimate the percentage of persons exhibiting a particular characteristic (e.g. hypertension). A weighted estimate of this percentage could be expressed in the form (1) by letting $y_i = 100$ if person i exhibits the characteristic and $y_i = 0$ if he or

she does not. A weighted median of a numeric variable could be found by finding the value Y^* for which the sum of the weights of all persons having observed values less than or equal to Y^* is approximately one half of the total weight,

$$\sum_{i:y_i \leq Y^*} w_i \approx 0.5 \sum_i w_i.$$

Weighted estimation procedures for more complicated quantities, e.g. coefficients from linear or logistic regression models, are also possible, although in some cases these estimates cannot be written in closed form and must be calculated by iterative procedures. Computational routines for calculating estimates from weighted survey data are available in several commercial statistical software packages, including SUDAAN (Research Triangle Institute, 1998), WesvarPC (Westat, 1996) and Stata (Stata Corp., 1997).

Which weight should be used when analyzing the NHANES III Multiply Imputed Data Set? Users of previously released NHANES III public-use files will recall that those files contained a variety of weights for different types of analyses. With those files, users were advised to use the 'final interview weight' (variable `WTPFQX6`) for analyses involving items from the household questionnaires, and the 'final examination weight' (variable `WTPFEX6`) for analyses involving items from the physical examination or joint analyses involving household questionnaire and examination items. The latter weight differs from the former in that it includes a nonresponse adjustment for subjects who were interviewed but not examined. In the NHANES III Multiply Imputed Data Set, however, missing examination items for all interviewed persons have been imputed, so the additional stage of nonresponse adjustment has become unnecessary. Therefore, the only weight needed for estimation in the NHANES III Multiply Imputed Data Set is the 'final interview weight' (`WTPFQX6`). This is the weight recommended for analyses of the NHANES III Multiply Imputed Data Set.

Standard errors for weighted estimates should be calculated in a manner that reflects the survey's complex design. Two methods have been recommended for variance estimation from NHANES III. The first method, known as *Taylor linearization*, takes advantage of the fact that many estimators of interest (e.g. ratios and regression coefficients) can be expressed as analytic functions of weighted sums $\hat{Y} = \sum_i w_i y_i$ for suitably defined survey variables y_i.

If an unbiased estimate for the variance $V(\hat{Y})$ can be found, then the initial (first-order) term of the Taylor series expansion of the function $\hat{Z} = g(\hat{Y})$ can be used to obtain an approximately unbiased variance estimate for $\hat{Z}$. The NHANES III sample was drawn by a two-PSU-per-stratum design. Under this design, the linearized variance estimate is obtained by summing over all the strata the squared differences between the estimates for the two PSUs within each stratum (Wolter, 1985). Indicators of the PSU (variable `SPPPSU6`) and stratum (variable `SDPSTRA6`) are provided in the NHANES III Multiply Imputed Data Set, allowing software packages for the analysis of survey data to calculate linearization-based standard errors. Depending on the program being used, the sample may need to be sorted by `SPPPSU6` and `SDPSTRA6` prior to running the estimation procedure. The degrees of freedom associated with the linearization variance estimate is the number of PSUs minus the number of strata, which in NHANES III is 49.

A second recommended technique for calculating standard errors for NHANES III estimates, called the *replicate method*, uses multiple sets of sample weights. Each set of weights is obtained by discarding some PSUs from the sample and reweighting the remaining units to resemble a full sample. Fifty-two replicates of the 'final interview weight' for NHANES III (variables `WTPQRP1` through `WTPQRP52`) were created by Fay's method, which is a modification of balanced repeated replication (BRR) for a two-PSU-per-stratum design (Wolter, 1985; Judkins, 1990). BRR discards one PSU from each stratum in creating each replicate, multiplying the sample weights by 0 and 2, respectively, for the deleted and retained PSUs. Fay's method perturbs the weights in a less extreme manner, multiplying them instead by factors of k and $2-k$ for some value of k between 0 and 1. Variance estimation by the replicate method proceeds as follows. Let $\hat{Z}$ denote the weighted estimate of a quantity calculated using the full-sample weight `WTPFQX6`. Let $\hat{Z}_{(j)}$ denote the same estimate calculated using the jth replicate weight `WTPQRP`j, $j = 1, \ldots, 52$. The estimated variance for $\hat{Z}$ is

$$\hat{V}(\hat{Z}) = \frac{1}{52\,(1-k)^2} \sum_{j=1}^{52} \left(\hat{Z}_{(j)} - \hat{Z}\right)^2, \tag{2}$$

The NHANES III replicate weights `WTPQRP1`–`WTPQRP52` were created using $k = 0.3$, so users should substitute $k = 0.3$ into the (2) when calculating variance estimates. The degrees of

freedom associated with this estimate is equal to the number of replicates, which in this case is 52.

One advantage of the replicate method is that it may be applied to many different kinds of estimators, including quantities that may be very complicated functions of the data. As long as a weighted estimation procedure is available, that procedure is simply repeated for each replicate weight, and the variation among the resulting estimates is used to obtain a standard error. Another advantage of the replicate method is that variability due to postratification and adjustments for unit nonresponse can be built into the replicate weights. In many cases, the two methods for calculating standard errors—Taylor linearization and the replicate method—will tend to produce similar results when applied to NHANES III data.

2.3 Combining the results across versions

When analyzing the NHANES III Multiply Imputed Data Set, the procedures described above—weighted estimation and calculation of standard errors by the linearization or replicate method—must be carried out five times, once for each of the five versions of the completed data. The five sets of estimates and standard errors must be temporarily stored and then combined using Rubin's rules for repeated-imputation inference (Rubin and Schenker, 1986; Rubin, 1987).

Rubin's rules require only simple arithmetic. Let Q denote a population quantity to be estimated, such as a prevalence rate, mean, quantile, or regression coefficient. Let $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_5$ denote the five estimates of Q obtained from the five imputed data files, and let $U_1, U_2, \ldots, U_5$ denote the corresponding variance estimates (squared standard errors) obtained by the linearization or replicate method. Let ν_{com} denote the complete-data degrees of freedom, i.e. the degrees of freedom associated with each of the variance estimates U_j. The overall estimate of Q is simply the average of the five estimates,

$$\bar{Q} = \frac{1}{5} \sum_{j=1}^{5} \hat{Q}_j. \tag{3}$$

The overall variance estimate associated with $\bar{Q}$ is

$$T = \bar{U} + \left(1 + \frac{1}{5}\right) B, \tag{4}$$

where $\bar{U} = \frac{1}{5}\sum_{j=1}^{5} U_j$ is the within-imputation variance and $B = \frac{1}{5-1}\sum_{j=1}^{5}(\hat{Q}_j - \bar{Q})^2$ is the between-imputation variance. The degrees of freedom associated T are obtained in the following manner. When the complete-data degrees of freedom are large ($\nu_{com} = \infty$), Rubin (1987) recommends the use of

$$\nu_m = (5-1)\left[1 + \frac{\bar{U}}{\left(1+\frac{1}{5}\right)B}\right]^2 .$$

If ν_{com} is not large, a more appropriate value is

$$\nu = \left[\frac{1}{\nu_m} + \frac{1}{\nu_{obs}}\right]^{-1}, \tag{5}$$

where

$$\nu_{obs} = \left(\frac{\nu_{com}+1}{\nu_{com}+3}\right)\nu_{com}\,\frac{\bar{U}}{T}$$

(Barnard and Rubin, 1999). For analyzing the NHANES III Multiply Imputed Data Set, we recommend (5) with $\nu_{com} = 49$ when $U_1, \ldots, U_5$ are obtained by linearization and $\nu_{con} = 52$ if $U_1, \ldots, U_5$ are obtained by the replicate method. Interval estimates may be calculated as $\bar{Q} \pm t_\nu\sqrt{T}$, where t_ν is a quantile of Student's t-distribution. For diagnostic purposes, it is useful to calculate the estimated percent rate of missing information for Q, which is given by

$$100 \times \left[1 - \left(\frac{\nu+1}{\nu+3}\right)\left(\frac{\nu_{com}+3}{\nu_{com}+1}\right)\frac{\bar{U}}{T}\right] \tag{6}$$

(Barnard and Rubin, 1999). Even for very simple estimands (e.g. population means), the estimated percent rate of missing information may differ considerably from the percentage of missing values for the variable in question. In many cases the rate of missing information will be lower, because the multiple imputation procedures utilize information contained in inter-variable relationships to predict the missing data values.

The computations described above can be easily implemented on a computer. If the software used to calculate weighted estimates and standard errors is also able to perform basic arithmetic on variables and arrays, then the entire analysis can be carried out within a single program. Procedures for analyzing survey data are available in SUDAAN (Research Triangle Institute, 1997). The SAS-callable version of SUDAAN is especially convenient because the SAS language makes it possible to automate the process of calculating five sets

of estimates and standard errors and combining the results. Several examples of analyses in SAS-callable SUDAAN are provided in Section 3 below. The statistical software package Stata (Stata Corp., 1997) also has a large number of commands (those whose names begin with the prefix **svy**) for the analysis of data from complex surveys. These commands can be executed repeatedly within Stata to calculate and store five sets of estimates and standard errors, and the results may be combined within Stata by the methods described above. Implementations of Rubin's (1987) rules in Stata are available from Dr. John Barnard of the Harvard University Department of Statistics (`barnard@stat.harvard.edu`).

3 Analysis examples

3.1 National estimates for means, prevalences, and quantiles

One of the most common uses of data from NHANES III is the estimation of means and prevalences for various characteristics within demographic subgroups of the population. Our first example analysis, shown in Figure 1, estimates means or prevalences for seven variables from the NHANES III examination for adults by categories of age (20–39, 40–59, 60+) and sex. The seven variables are bone mineral density of the femur neck, waist circumference, body mass index (derived from weight and height), overweight status (derived from body mass index), systolic blood pressure, serum iron and serum total cholesterol. In this example, the SUDAAN procedure PROC DESCRIPT is used to calculate weighted estimates and standard errors by the Taylor linearization approach. After applying PROC DESCRIPT to each of the five completed data files, the results are combined by the methods of Rubin (1987) and Barnard and Rubin (1999) with $\nu_{com} = 49$. This program executed without errors using SUDAAN Release 7.5.3 and SAS Version 6.12.

Two details about this program should be noted. First, it assumes that five permanent SAS data sets `NH3MI1`, `NH3MI2`, ..., `NH3MI5` which contain the five versions of the completed NHANES III data already exist. To create these SAS data sets, one must run the input statements contained in the files `CORE.SAS`, `IMP1.SAS`, `IMP2.SAS`, ..., `IMP5.SAS` and `NH3MI.SAS` distributed with the NHANES III Multiply Imputed Data Set. Second, when the Taylor

linearization method is used, SUDAAN requires that the data be sorted by stratum and PSU identifiers (variables `SDPSTRA6` and `SDPPSU6`) prior to estimation.

The results from this example, presented in a slightly re-formatted form, are shown in Table 1. This columns of this table contain the estimates, standard errors, degrees of freedom, lower and upper endpoints of 95% intervals, and estimated percent rate of missing information calculated according to (6). The rate of missing information is particularly interesting because it reveals the extent to which the standard errors are affected by the variability of imputed values across the five data sets. For comparison, the actual percentage rates of imputed values for each estimand are shown along the right-hand side of Table 1. For the estimands that are functions of a more than one NHANES III variable (e.g. body mass index), the number shown is the percentage of subjects in the given domain for which at least one of the required variables was imputed. Notice that in most cases, the estimated percent rate of missing information is substantially lower than the actual percentage of imputed values, indicating that the imputation procedure is effectively making use of other information to predict the missing data. For a few estimands (e.g. mean serum iron for females age 60+), the percent rate of missing information is higher than the actual imputation rate. This phenomenon suggests that in those particular domains, the imputed values exert a higher-than-average degree of influence over the estimand in question.

This example is easily modified to use the replicate method of variance estimation rather than Taylor linearization. A modified version of the program using the replicate method is shown in Figure 2. Note that with the replicate method, it is no longer necessary to sort the data by `SDPSTRA6` and `SDPPSU6` prior to estimation. Because the replicate weights `WTPQRP1`, ..., `WTPQRP52` were created by Fay's method with $k = 0.3$, it is essential to specify a Fay adjustment factor when calling the variance estimation routines. In SUDAAN, this factor is expressed as $1/(1-k)^2 = 2.0408$ in the `ADJFAY` option to the `REPWGT` command. The results from the replicate method are displayed in Table 2. The point estimates are identical to those displayed in Table 1, but the standard errors are somewhat different. The estimated percent rates of missing information are also somewhat different. These discrepancies in the rates of missing information between Table 2 and Table 1 are entirely due to the different

variance estimation procedures, because the component of variance due to imputation—the between-imputation variance described in Section 2.3—is identical under the two methods.

The SUDAAN procedure PROC DESCRIPT may also be used to estimate population percentiles. A sample program illustrating the estimation of medians and 90th percentiles for two examination variables—systolic blood pressure and total cholesterol—is provided in Figure 3. The results from this program are displayed in Table 3. This example uses the Taylor linearization method of variance estimation but could be easily modified as in the previous example to use the replicate method.

```
*****************************************************************;
*                                                                ;
*   MI Analysis - Example A                                      ;
*                                                                ;
*   Analysis of NHANES III Multiply Imputed Data Set using       ;
*   SAS-callable SUDAAN, PROC DESCRIPT                           ;
*                                                                ;
*   Estimates means for                                          ;
*      BDPFNDMI = Bone mineral density femur neck (gm/cm sq)     ;
*      BMPWSTMI = Waist circumference (cm)                       ;
*      BMI      = Body mass index (derived from weight & height) ;
*      OVERWT   = 100 if overweight, 0 else (derived from BMI)   ;
*      SYSTOLIC = exam systolic BP (avg of three measurements)   ;
*      FEPMI    = serum iron (ug/dl)                             ;
*      TCPMI    = serum total cholesterol (mg/dl)                ;
*   for adults by categories of sex and age (20-39, 40-59, 60+)  ;
*                                                                ;
*   Variance estimation by Taylor linearization (WR) method      ;
*                                                                ;
*   Notes:                                                       ;
*                                                                ;
*     (1) This program is only an example. You may need to       ;
*     modify it to suit your needs.                              ;
*                                                                ;
*     (2) This program assumes that the SAS programs             ;
*     CORE.SAS, IMP1.SAS, ..., IMP5.SAS, and NH3MI.SAS have      ;
*     already been run to create the SAS data sets               ;
*     NH3MI1, ..., NH3MI5. These programs are provided on the    ;
*     NHANES III Multiply Imputed Data Set CD-ROM                ;
*                                                                ;
*****************************************************************;

*****************************************************************;
*   Specify the directory where SAS datasets NH3MI1, ..., NH3MI5 ;
*   have been stored. You may need to modify the line below.     ;
*****************************************************************;
LIBNAME NH3MI 'C:\MyDir';

*****************************************************************;
*   This macro cycles through the five imputed data sets,        ;
*   preparing the variables for use by SUDAAN's PROC DESCRIPT.   ;
*   It then calls PROC DESCRIPT to calculate estimates and       ;
*   standard errors for each imputed data set, storing the       ;
*   results in five temporary SAS data sets called               ;
*   SUDNOUT1, ..., SUDNOUT5                                      ;
*****************************************************************;
%MACRO ANALYZE;

   %DO IMPNO = 1 %TO 5;

   DATA FORSUDN (REPLACE=YES);
      SET NH3MI.NH3MI&IMPNO;
      ************************************;
      * categorize age                   ;
      ************************************;
      AGE = HSAGEIR;
      IF HSAGEU = 1 THEN AGE = AGE / 12;
      IF AGE GE 20 AND AGE LE 39 THEN AGEGRP = 1;
      ELSE IF AGE GE 40 AND AGE LE 59 THEN AGEGRP = 2;
      ELSE IF AGE GE 60 THEN AGEGRP = 3;
```

Figure 1: Program in SAS and SAS-callable SUDAAN for analysis of population means and prevalence rates, with standard errors calculated by Taylor linearization

```
   ELSE AGEGRP = 0;
   ************************************;
   * calculate body mass index        ;
   * and overweight indicator         ;
   ************************************;
   BMI = BMPWTMI / (BMPHTMI/100)**2;
   OVERWT = 0;
   IF HSSEX = 1 AND BMI GE 27.8 THEN OVERWT = 100;
   IF HSSEX = 2 AND BMI GE 27.3 THEN OVERWT = 100;
   ************************************;
   * average systolic blood pressure  ;
   ************************************;
   SYSTOLIC = (PEP6G1MI + PEP6H1MI + PEP6I1MI) / 3;
   ************************************;
   * output select variables for adults  ;
   ************************************;
   KEEP SDPSTRA6 SDPPSU6 WTPFQX6 AGE AGEGRP HSSEX
      BDPFNDMI BMPWSTMI BMI OVERWT SYSTOLIC FEPMI TCPMI;
   IF AGE GE 20 THEN OUTPUT;
   RUN;

************************************;
* sort data by pseudo-stratum and PSU ;
* in preparation for SUDAAN        ;
* linearization (WR) method        ;
************************************;
PROC SORT DATA=FORSUDN;
   BY SDPSTRA6 SDPPSU6;
   RUN;

************************************;
* call SUDAAN Proc Descript using the ;
* linearization (WR) method, storing  ;
* the results in a temporary data set ;
* called SUDNOUT                   ;
************************************;
PROC DESCRIPT DATA=FORSUDN FILETYPE=SAS DESIGN=WR MEANS;
   NEST SDPSTRA6 SDPPSU6 / MISSUNIT;
   WEIGHT WTPFQX6;
   VAR BDPFNDMI BMPWSTMI BMI OVERWT SYSTOLIC FEPMI TCPMI;
   SUBGROUP HSSEX AGEGRP;
   LEVELS 2 3;
   TABLES HSSEX*AGEGRP;
   OUTPUT MEAN SEMEAN / FILENAME=SUDNOUT FILETYPE=SAS REPLACE;
   RUN;

************************************;
* read in SUDNOUT, renaming the    ;
* estimates and standard errors    ;
* to create SUDNOUTi (i=1,2,3,4,5) ;
************************************;
DATA SUDNOUT&IMPNO;
   SET SUDNOUT;
   EST&IMPNO = MEAN;
   SE&IMPNO = SEMEAN;
   KEEP VARIABLE HSSEX AGEGRP EST&IMPNO SE&IMPNO;
   RUN;

************************************;
* sort SUDNOUTi (i=1,2,3,4,5) in   ;
```

Figure 1 (continued)

```
   * preparation for final merge         ;
   ************************************;
   PROC SORT DATA=SUDNOUT&IMPNO;
      BY VARIABLE HSSEX AGEGRP;
      RUN;

   %END;
%MEND ANALYZE;

%ANALYZE;

*****************************************************************;
*   Combine the estimates and standard errors stored in         ;
*   SUDNOUTi (i=1,2,3,4,5) using Rubin's rules                  ;
*****************************************************************;
DATA COMBINED;
   MERGE SUDNOUT1 SUDNOUT2 SUDNOUT3 SUDNOUT4 SUDNOUT5;
   BY VARIABLE HSSEX AGEGRP;
   ************************************;
   * labels for the estimands          ;
   ************************************;
   LENGTH QTYLABEL $25.;
   IF      VARIABLE = 1 THEN QTYLABEL = 'Mean BMD femur neck';
   ELSE IF VARIABLE = 2 THEN QTYLABEL = 'Mean waist circumference';
   ELSE IF VARIABLE = 3 THEN QTYLABEL = 'Mean body mass index';
   ELSE IF VARIABLE = 4 THEN QTYLABEL = 'Pct overweight';
   ELSE IF VARIABLE = 5 THEN QTYLABEL = 'Mean systolic BP';
   ELSE IF VARIABLE = 6 THEN QTYLABEL = 'Mean serum iron';
   ELSE IF VARIABLE = 7 THEN QTYLABEL = 'Mean serum cholesterol';
   ************************************;
   * labels for the demographic groups   ;
   ************************************;
   LENGTH GRPLABEL $25.;
   IF HSSEX = 0 THEN DO;
     IF      AGEGRP=0 THEN GRPLABEL = 'All Adults (20+ years)';
     ELSE IF AGEGRP=1 THEN GRPLABEL = 'M&F 20-39 years';
     ELSE IF AGEGRP=2 THEN GRPLABEL = 'M&F 40-59 years';
     ELSE IF AGEGRP=3 THEN GRPLABEL = 'M&F 60+ years';
     END;
   ELSE IF HSSEX = 1 THEN DO;
     IF      AGEGRP=0 THEN GRPLABEL = 'All Males (20+ years)';
     ELSE IF AGEGRP=1 THEN GRPLABEL = 'Males 20-39 years';
     ELSE IF AGEGRP=2 THEN GRPLABEL = 'Males 40-59 years';
     ELSE IF AGEGRP=3 THEN GRPLABEL = 'Males 60+ years';
     END;
   ELSE IF HSSEX = 2 THEN DO;
     IF      AGEGRP=0 THEN GRPLABEL = 'All Females (20+ years)';
     ELSE IF AGEGRP=1 THEN GRPLABEL = 'Females 20-39 years';
     ELSE IF AGEGRP=2 THEN GRPLABEL = 'Females 40-59 years';
     ELSE IF AGEGRP=3 THEN GRPLABEL = 'Females 60+ years';
     END;
   ************************************;
   * combine results by Rubin's rules    ;
   ************************************;
   EST = MEAN( EST1, EST2, EST3, EST4, EST5);
   WITHNVAR = MEAN(SE1**2, SE2**2, SE3**2, SE4**2, SE5**2);
   BETWNVAR = VAR( EST1, EST2, EST3, EST4, EST5);
   TOTVAR = WITHNVAR + (1 + 1/5)*BETWNVAR;
   SE = TOTVAR**.5;
   ************************************;
```

Figure 1 (continued)

```
      * calculate degrees of freedom for    ;
      * the t-approximation, endpoints of   ;
      * 95% interval, and percent rate of   ;
      * missing information.                ;
      * Degrees of freedom are found        ;
      * by the method of Barnard and Rubin  ;
      * (1999) assuming df=49 for complete  ;
      * data.                               ;
      ************************************;
      DFCOM = 49;
      IF BETWNVAR GT 0 THEN DO;
         ***********************************;
         * usual case                       ;
         ***********************************;
         DFM = (5-1) * (1 + (5*WITHNVAR)/((5+1)*BETWNVAR))**2;
         DFOBS = ((DFCOM+1)/(DFCOM+3)) * DFCOM * WITHNVAR/TOTVAR;
         DF = 1 / ( 1/DFM + 1/DFOBS );
         LOWER95 = EST - TINV(.975,DF)*SE;
         UPPER95 = EST + TINV(.975,DF)*SE;
         RATIO = ((DF+1)*(DFCOM+3))/((DF+3)*(DFCOM+1));
         PCTMIS = 100*( 1 - RATIO*WITHNVAR/TOTVAR );
         END;
      ELSE IF BETWNVAR EQ 0 THEN DO;
         ***********************************;
         * special case to avoid division by ;
         * zero if between-imputation        ;
         * variance happens to be zero       ;
         ***********************************;
         DF = DFCOM;
         LOWER95 = EST - TINV(.975,DF)*SE;
         UPPER95 = EST + TINV(.975,DF)*SE;
         PCTMIS = 0;
         END;
      FORMAT EST SE LOWER95 UPPER95 8.4 DF 10.1 PCTMIS 5.1;
      RUN;

*****************************************************************;
*   print results                                                ;
*****************************************************************;
OPTIONS LINESIZE=132;
PROC PRINT DATA=COMBINED;
   VAR QTYLABEL GRPLABEL EST SE DF LOWER95 UPPER95 PCTMIS;
   RUN;
```

Figure 1 (continued)

Table 1: Results from MI Analysis Example A—estimates, standard errors, degrees of freedom, lower and upper endpoints of the 95% interval estimates, and estimated percent rate of missing information—with percent rate of imputed values shown for comparison

	Est.	SE	df	lower	upper	% mis	% imputed
Mean BMD femur neck							
All Adults (20+ years)	0.8237	0.0027	46.7	0.8183	0.8291	1.0	22.2
M/F 20–39 years	0.8937	0.0025	32.3	0.8886	0.8988	17.0	20.3
M/F 40–59 years	0.8113	0.0027	40.0	0.8058	0.8168	9.3	16.1
M/F 60+ years	0.6978	0.0038	28.4	0.6902	0.7055	21.1	28.8
All Males (20+ years)	0.8710	0.0029	37.0	0.8652	0.8769	12.3	19.3
Males 20–39 years	0.9355	0.0038	21.9	0.9276	0.9434	29.3	15.7
Males 40–59 years	0.8401	0.0038	39.1	0.8324	0.8478	10.2	15.5
Males 60+ years	0.7679	0.0041	40.7	0.7596	0.7763	8.5	26.1
All Females (20+ years)	0.7807	0.0033	43.8	0.7740	0.7873	5.1	24.7
Females 20–39 years	0.8531	0.0032	23.9	0.8466	0.8596	26.5	24.4
Females 40–59 years	0.7839	0.0035	37.6	0.7769	0.7910	11.7	16.5
Females 60+ years	0.6454	0.0041	16.1	0.6367	0.6542	39.0	31.2
Mean waist circumference							
All Adults (20+ years)	91.8776	0.2444	46.2	91.3857	92.3696	1.8	16.4
M/F 20–39 years	87.3867	0.3351	45.9	86.7122	88.0612	2.4	10.5
M/F 40–59 years	95.0592	0.3239	46.1	94.4073	95.7111	2.1	12.7
M/F 60+ years	96.6704	0.2431	40.2	96.1791	97.1617	9.1	25.8
All Males (20+ years)	95.3156	0.2773	44.5	94.7569	95.8743	4.3	16.0
Males 20–39 years	90.8838	0.3995	45.2	90.0793	91.6882	3.3	12.1
Males 40–59 years	98.7542	0.3989	43.6	97.9501	99.5582	5.4	12.8
Males 60+ years	100.3508	0.3342	44.2	99.6774	101.0241	4.6	22.7
All Females (20+ years)	88.7527	0.3486	46.0	88.0511	89.4544	2.1	16.8
Females 20–39 years	83.9921	0.4671	46.8	83.0524	84.9319	0.7	9.1
Females 40–59 years	91.5530	0.4706	44.4	90.6048	92.5011	4.5	12.6
Females 60+ years	93.9187	0.2967	41.5	93.3197	94.5176	7.7	28.6
Mean body mass index							
All Adults (20+ years)	26.4785	0.1082	46.7	26.2607	26.6963	1.1	9.9
M/F 20–39 years	25.5951	0.1400	46.1	25.3133	25.8768	2.0	7.3
M/F 40–59 years	27.5026	0.1411	45.9	27.2185	27.7867	2.3	8.3
M/F 60+ years	26.8734	0.1085	46.4	26.6551	27.0916	1.5	13.8
All Males (20+ years)	26.5790	0.1069	45.8	26.3637	26.7942	2.5	10.0
Males 20–39 years	25.8686	0.1494	44.7	25.5677	26.1695	4.0	9.1
Males 40–59 years	27.4878	0.1585	44.9	27.1686	27.8070	3.8	8.7
Males 60+ years	26.8257	0.1390	45.3	26.5458	27.1057	3.2	12.0
All Females (20+ years)	26.3872	0.1512	46.8	26.0831	26.6913	0.8	9.7
Females 20–39 years	25.3295	0.2019	46.9	24.9233	25.7358	0.6	5.8
Females 40–59 years	27.5166	0.2051	46.6	27.1040	27.9292	1.2	8.0
Females 60+ years	26.9090	0.1287	45.7	26.6499	27.1681	2.7	15.5
Percent overweight							
All Adults (20+ years)	34.1990	0.6865	46.8	32.8178	35.5803	0.8	9.9
M/F 20–39 years	26.9193	0.9225	45.1	25.0614	28.7772	3.5	7.3
M/F 40–59 years	41.1406	1.1091	45.6	38.9075	43.3737	2.9	8.3
M/F 60+ years	39.5125	1.0747	42.0	37.3439	41.6812	7.1	13.8
All Males (20+ years)	32.6846	0.8994	45.0	30.8732	34.4959	3.6	10.0
Males 20–39 years	25.6281	1.2274	41.3	23.1498	28.1063	7.9	9.1
Males 40–59 years	39.5681	1.5370	46.0	36.4743	42.6619	2.2	8.7
Males 60+ years	38.4953	1.3889	42.7	35.6937	41.2969	6.4	12.0
All Females (20+ years)	35.5755	0.9450	46.4	33.6738	37.4773	1.6	9.7
Females 20–39 years	28.1727	1.4102	47.0	25.3356	31.0098	0.4	5.8
Females 40–59 years	42.6328	1.3679	44.8	39.8774	45.3882	4.0	8.0
Females 60+ years	40.2731	1.4804	43.0	37.2876	43.2585	6.1	15.5

Table 1 (continued)

	Est.	SE	df	lower	upper	% mis	% imputed
Mean systolic blood pressure							
All Adults (20+ years)	121.1273	0.3734	46.2	120.3757	121.8788	1.9	15.6
M/F 20–39 years	112.6892	0.2956	38.0	112.0908	113.2876	11.3	10.5
M/F 40–59 years	121.7014	0.3616	40.0	120.9706	122.4322	9.2	12.7
M/F 60+ years	137.5690	0.4819	32.6	136.5881	138.5498	16.6	23.6
All Males (20+ years)	123.4720	0.4234	43.0	122.6181	124.3258	6.1	15.4
Males 20–39 years	117.5362	0.4247	26.0	116.6633	118.4091	23.9	12.0
Males 40–59 years	124.3262	0.5681	33.7	123.1713	125.4812	15.5	12.4
Males 60+ years	136.0907	0.6575	39.3	134.7613	137.4202	9.9	21.3
All Females (20+ years)	118.9961	0.4769	46.6	118.0365	119.9557	1.1	15.9
Females 20–39 years	107.9842	0.3213	46.6	107.3378	108.6307	1.2	9.2
Females 40–59 years	119.2107	0.4733	40.7	118.2546	120.1667	8.6	13.0
Females 60+ years	138.6742	0.6560	28.1	137.3307	140.0176	21.4	25.7
Mean serum iron							
All Adults (20+ years)	91.2341	0.6997	44.1	89.8240	92.6442	4.8	14.2
M/F 20–39 years	96.8120	1.0470	36.6	94.6898	98.9341	12.7	12.1
M/F 40–59 years	87.9444	0.9035	42.4	86.1216	89.7673	6.8	11.8
M/F 60+ years	84.3704	0.8222	31.3	82.6942	86.0466	18.0	18.3
All Males (20+ years)	98.6262	0.8028	33.6	96.9939	100.2584	15.6	14.2
Males 20–39 years	104.2946	1.3692	28.6	101.4925	107.0967	20.9	13.6
Males 40–59 years	95.5580	1.2037	32.1	93.1065	98.0095	17.2	12.1
Males 60+ years	90.1035	1.2441	40.5	87.5900	92.6169	8.8	16.4
All Females (20+ years)	84.5153	0.8540	44.4	82.7947	86.2360	4.4	14.3
Females 20–39 years	89.5485	1.4165	34.8	86.6722	92.4248	14.4	10.9
Females 40–59 years	80.7198	1.1555	45.6	78.3933	83.0463	2.8	11.6
Females 60+ years	80.0840	0.9039	15.8	78.1659	82.0022	39.6	20.0
Mean serum cholesterol							
All Adults (20+ years)	204.0814	0.7839	43.2	202.5009	205.6620	5.8	14.7
M/F 20–39 years	188.3679	0.9700	39.3	186.4064	190.3293	9.9	12.6
M/F 40–59 years	213.0604	1.1104	36.9	210.8103	215.3104	12.3	12.2
M/F 60+ years	223.8145	1.1029	40.4	221.5862	226.0429	8.8	18.9
All Males (20+ years)	202.0179	0.8801	45.2	200.2454	203.7903	3.3	14.6
Males 20–39 years	190.5982	1.2186	35.2	188.1248	193.0716	14.0	14.1
Males 40–59 years	212.7610	1.2710	33.9	210.1778	215.3442	15.3	12.3
Males 60+ years	212.0425	1.2546	46.0	209.5170	214.5679	2.2	16.8
All Females (20+ years)	205.9571	0.9786	39.2	203.9780	207.9362	10.1	14.8
Females 20–39 years	186.2028	1.0492	41.6	184.0848	188.3209	7.6	11.3
Females 40–59 years	213.3444	1.3890	37.1	210.5304	216.1584	12.1	12.1
Females 60+ years	232.6161	1.5527	33.4	229.4586	235.7736	15.8	20.7

```
*****************************************************************;
*                                                                ;
*   MI Analysis - Example B                                      ;
*                                                                ;
*   Identical to Analysis A, except that variance estimates      ;
*   are calculated by replicate method (BRR) with Fay adjustment.;
*   Now there is no need to sort data by PSU and stratum prior   ;
*   to running SUDAAN's PROC DESCRIPT.                           ;
*                                                                ;
*   Notes:                                                       ;
*                                                                ;
*     (1) This program is only an example. You may need to       ;
*     modify it to suit your needs.                              ;
*                                                                ;
*     (2) This program assumes that the SAS programs             ;
*     CORE.SAS, IMP1.SAS, ..., IMP5.SAS, and NH3MI.SAS have      ;
*     already been run to create the SAS data sets               ;
*     NH3MI1, ..., NH3MI5. These programs are provided on the    ;
*     NHANES III Multiply Imputed Data Set CD-ROM                ;
*                                                                ;
*****************************************************************;

*****************************************************************;
*   Specify the directory where SAS datasets NH3MI1, ..., NH3MI5 ;
*   have been stored. You may need to modify the line below.     ;
*****************************************************************;
LIBNAME NH3MI 'C:\MyDir';

*****************************************************************;
*   This macro cycles through the five imputed data sets,        ;
*   preparing the variables for use by SUDAAN's PROC DESCRIPT.   ;
*   It then calls PROC DESCRIPT to calculate estimates and       ;
*   standard errors for each imputed data set, storing the       ;
*   results in five temporary SAS data sets called               ;
*   SUDNOUT1, ..., SUDNOUT5                                      ;
*****************************************************************;
%MACRO ANALYZE;

   %DO IMPNO = 1 %TO 5;

   DATA FORSUDN (REPLACE=YES);
      SET NH3MI.NH3MI&IMPNO;
      ************************************;
      * categorize age                   ;
      ************************************;
      AGE = HSAGEIR;
      IF HSAGEU = 1 THEN AGE = AGE / 12;
      IF AGE GE 20 AND AGE LE 39 THEN AGEGRP = 1;
      ELSE IF AGE GE 40 AND AGE LE 59 THEN AGEGRP = 2;
      ELSE IF AGE GE 60 THEN AGEGRP = 3;
      ELSE AGEGRP = 0;
      ************************************;
      * calculate body mass index        ;
      * and overweight indicator         ;
      ************************************;
      BMI = BMPWTMI / (BMPHTMI/100)**2;
      OVERWT = 0;
      IF HSSEX = 1 AND BMI GE 27.8 THEN OVERWT = 100;
      IF HSSEX = 2 AND BMI GE 27.3 THEN OVERWT = 100;
      ************************************;
```

Figure 2: Program in SAS and SAS-callable SUDAAN for analysis of population means and prevalence rates, with standard errors calculated by replicate method

```
      * average systolic blood pressure     ;
      ************************************;
      SYSTOLIC = (PEP6G1MI + PEP6H1MI + PEP6I1MI) / 3;
      ************************************;
      * output select variables for adults  ;
      ************************************;
      KEEP WTPFQX6 WTPQRP1-WTPQRP52 AGE AGEGRP HSSEX
         BDPFNDMI BMPWSTMI BMI OVERWT SYSTOLIC FEPMI TCPMI;
      IF AGE GE 20 THEN OUTPUT;
      RUN;

   ************************************;
   * call SUDAAN Proc Descript using the ;
   * replicate (BRR) method, storing     ;
   * the results in a temporary data set ;
   * called SUDNOUT                      ;
   * Use Fay method with adjustment      ;
   * ADJFAY  = 2.0408                    ;
   ************************************;
   PROC DESCRIPT DATA=FORSUDN FILETYPE=SAS DESIGN=BRR MEANS;
      WEIGHT WTPFQX6;
      REPWGT WTPQRP1-WTPQRP52 / ADJFAY=2.0408;
      VAR BDPFNDMI BMPWSTMI BMI OVERWT SYSTOLIC FEPMI TCPMI;
      SUBGROUP HSSEX AGEGRP;
      LEVELS 2 3;
      TABLES HSSEX*AGEGRP;
      OUTPUT MEAN SEMEAN / FILENAME=SUDNOUT FILETYPE=SAS REPLACE;
      RUN;

   ************************************;
   * read in SUDNOUT, renaming the       ;
   * estimates and standard errors       ;
   * to create SUDNOUTi (i=1,2,3,4,5)    ;
   ************************************;
   DATA SUDNOUT&IMPNO;
      SET SUDNOUT;
      EST&IMPNO = MEAN;
      SE&IMPNO = SEMEAN;
      KEEP VARIABLE HSSEX AGEGRP EST&IMPNO SE&IMPNO;
      RUN;

   ************************************;
   * sort SUDNOUTi (i=1,2,3,4,5) in      ;
   * preparation for final merge         ;
   ************************************;
   PROC SORT DATA=SUDNOUT&IMPNO;
      BY VARIABLE HSSEX AGEGRP;
      RUN;

   %END;
%MEND ANALYZE;

%ANALYZE;

*****************************************************************;
*   Combine the estimates and standard errors stored in          ;
*   SUDNOUTi (i=1,2,3,4,5) using Rubin's rules                   ;
*****************************************************************;
DATA COMBINED;
   MERGE SUDNOUT1 SUDNOUT2 SUDNOUT3 SUDNOUT4 SUDNOUT5;
```

Figure 2 (continued)

```
BY VARIABLE HSSEX AGEGRP;
************************************;
* labels for the estimands          ;
************************************;
LENGTH QTYLABEL $25.;
IF      VARIABLE = 1 THEN QTYLABEL = 'Mean BMD femur neck';
ELSE IF VARIABLE = 2 THEN QTYLABEL = 'Mean waist circumference';
ELSE IF VARIABLE = 3 THEN QTYLABEL = 'Mean body mass index';
ELSE IF VARIABLE = 4 THEN QTYLABEL = 'Pct overweight';
ELSE IF VARIABLE = 5 THEN QTYLABEL = 'Mean systolic BP';
ELSE IF VARIABLE = 6 THEN QTYLABEL = 'Mean serum iron';
ELSE IF VARIABLE = 7 THEN QTYLABEL = 'Mean serum cholesterol';
************************************;
* labels for the demographic groups  ;
************************************;
LENGTH GRPLABEL $25.;
IF HSSEX = 0 THEN DO;
  IF      AGEGRP=0 THEN GRPLABEL = 'All Adults (20+ years)';
  ELSE IF AGEGRP=1 THEN GRPLABEL = 'M&F 20-39 years';
  ELSE IF AGEGRP=2 THEN GRPLABEL = 'M&F 40-59 years';
  ELSE IF AGEGRP=3 THEN GRPLABEL = 'M&F 60+ years';
  END;
ELSE IF HSSEX = 1 THEN DO;
  IF      AGEGRP=0 THEN GRPLABEL = 'All Males (20+ years)';
  ELSE IF AGEGRP=1 THEN GRPLABEL = 'Males 20-39 years';
  ELSE IF AGEGRP=2 THEN GRPLABEL = 'Males 40-59 years';
  ELSE IF AGEGRP=3 THEN GRPLABEL = 'Males 60+ years';
  END;
ELSE IF HSSEX = 2 THEN DO;
  IF      AGEGRP=0 THEN GRPLABEL = 'All Females (20+ years)';
  ELSE IF AGEGRP=1 THEN GRPLABEL = 'Females 20-39 years';
  ELSE IF AGEGRP=2 THEN GRPLABEL = 'Females 40-59 years';
  ELSE IF AGEGRP=3 THEN GRPLABEL = 'Females 60+ years';
  END;
************************************;
* combine results by Rubin's rules   ;
************************************;
EST = MEAN( EST1, EST2, EST3, EST4, EST5);
WITHNVAR = MEAN(SE1**2, SE2**2, SE3**2, SE4**2, SE5**2);
BETWNVAR = VAR( EST1, EST2, EST3, EST4, EST5);
TOTVAR = WITHNVAR + (1 + 1/5)*BETWNVAR;
SE = TOTVAR**.5;
************************************;
* calculate degrees of freedom for   ;
* the t-approximation, endpoints of  ;
* 95% interval, and percent rate of  ;
* missing information.               ;
* Degrees of freedom are found       ;
* by the method of Barnard and Rubin ;
* (1999) assuming df=52 for complete ;
* data.                              ;
************************************;
DFCOM = 52;
IF BETWNVAR GT 0 THEN DO;
   ************************************;
   * usual case                       ;
   ************************************;
   DFM = (5-1) * (1 + (5*WITHNVAR)/((5+1)*BETWNVAR))**2;
   DFOBS = ((DFCOM+1)/(DFCOM+3)) * DFCOM * WITHNVAR/TOTVAR;
   DF = 1 / ( 1/DFM + 1/DFOBS );
```

Figure 2 (continued)

```
      LOWER95 = EST - TINV(.975,DF)*SE;
      UPPER95 = EST + TINV(.975,DF)*SE;
      RATIO = ((DF+1)*(DFCOM+3))/((DF+3)*(DFCOM+1));
      PCTMIS = 100*( 1 - RATIO*WITHNVAR/TOTVAR );
      END;
   ELSE IF BETWNVAR EQ 0 THEN DO;
      ***********************************;
      * special case to avoid division by ;
      * zero if between-imputation        ;
      * variance happens to be zero       ;
      ***********************************;
      DF = DFCOM;
      LOWER95 = EST - TINV(.975,DF)*SE;
      UPPER95 = EST + TINV(.975,DF)*SE;
      PCTMIS = 0;
      END;
   FORMAT EST SE LOWER95 UPPER95 8.4 DF 10.1 PCTMIS 5.1;
   RUN;

*****************************************************************;
*   print results                                                ;
*****************************************************************;
OPTIONS LINESIZE=132;
PROC PRINT DATA=COMBINED;
   VAR QTYLABEL GRPLABEL EST SE DF LOWER95 UPPER95 PCTMIS;
   RUN;
```

Figure 2 (continued)

Table 2: Results from MI Analysis Example B—estimates, standard errors, degrees of freedom, lower and upper endpoints of the 95% interval estimates, and estimated percent rate of missing information—with percent rate of imputed values shown for comparison

	Est.	SE	df	lower	upper	% mis	% imputed
Mean BMD femur neck							
All Adults (20+ years)	0.8237	0.0013	47.5	0.8211	0.8263	3.9	22.2
M/F 20–39 years	0.8937	0.0022	28.1	0.8892	0.8982	22.8	20.3
M/F 40–59 years	0.8113	0.0028	43.2	0.8055	0.8170	8.4	16.1
M/F 60+ years	0.6978	0.0029	18.4	0.6917	0.7039	35.7	28.8
All Males (20+ years)	0.8710	0.0021	27.5	0.8667	0.8753	23.5	19.3
Males 20–39 years	0.9355	0.0036	20.4	0.9280	0.9430	32.6	15.7
Males 40–59 years	0.8401	0.0036	40.1	0.8328	0.8474	11.3	15.5
Males 60+ years	0.7679	0.0038	41.4	0.7603	0.7756	10.1	26.1
All Females (20+ years)	0.7807	0.0018	33.8	0.7770	0.7844	17.1	24.7
Females 20–39 years	0.8531	0.0027	17.2	0.8475	0.8587	37.9	24.4
Females 40–59 years	0.7839	0.0035	40.0	0.7769	0.7910	11.4	16.5
Females 60+ years	0.6454	0.0036	11.7	0.6375	0.6534	50.3	31.2
Mean waist circumference							
All Adults (20+ years)	91.8776	0.1963	48.5	91.4831	92.2721	2.8	16.4
M/F 20–39 years	87.3867	0.2953	48.3	86.7930	87.9804	3.0	10.5
M/F 40–59 years	95.0592	0.2694	48.4	94.5175	95.6008	2.9	12.7
M/F 60+ years	96.6704	0.2375	42.0	96.1911	97.1496	9.5	25.8
All Males (20+ years)	95.3156	0.2052	43.8	94.9020	95.7293	7.8	16.0
Males 20–39 years	90.8838	0.3313	46.8	90.2173	91.5502	4.7	12.1
Males 40–59 years	98.7542	0.3397	44.3	98.0698	99.4386	7.3	12.8
Males 60+ years	100.3508	0.3007	46.0	99.7454	100.9561	5.7	22.7
All Females (20+ years)	88.7527	0.3106	48.6	88.1285	89.3770	2.6	16.8
Females 20–39 years	83.9921	0.4337	49.8	83.1210	84.8633	0.8	9.1
Females 40–59 years	91.5530	0.4303	46.3	90.6871	92.4188	5.3	12.6
Females 60+ years	93.9187	0.3259	45.3	93.2624	94.5749	6.4	28.6
Mean body mass index							
All Adults (20+ years)	26.4785	0.0833	49.2	26.3111	26.6460	1.7	9.9
M/F 20–39 years	25.5951	0.1191	48.5	25.3557	25.8344	2.7	7.3
M/F 40–59 years	27.5026	0.1120	47.8	27.2773	27.7279	3.6	8.3
M/F 60+ years	26.8734	0.1034	49.3	26.6656	27.0811	1.6	13.8
All Males (20+ years)	26.5790	0.0798	47.2	26.4185	26.7395	4.3	10.0
Males 20–39 years	25.8686	0.1232	45.8	25.6205	26.1167	5.9	9.1
Males 40–59 years	27.4878	0.1319	46.3	27.2225	27.7532	5.4	8.7
Males 60+ years	26.8257	0.1122	46.8	26.5999	27.0515	4.8	12.0
All Females (20+ years)	26.3872	0.1263	49.6	26.1335	26.6410	1.1	9.7
Females 20–39 years	25.3295	0.1753	49.8	24.9775	25.6816	0.7	5.8
Females 40–59 years	27.5166	0.1704	49.3	27.1742	27.8591	1.6	8.0
Females 60+ years	26.9090	0.1397	48.8	26.6282	27.1897	2.3	15.5
Percent overweight							
All Adults (20+ years)	34.1990	0.5686	49.6	33.0566	35.3414	1.1	9.9
M/F 20–39 years	26.9193	0.9017	47.8	25.1061	28.7325	3.6	7.3
M/F 40–59 years	41.1406	0.8689	47.0	39.3927	42.8885	4.6	8.3
M/F 60+ years	39.5125	1.0144	43.7	37.4678	41.5573	8.0	13.8
All Males (20+ years)	32.6846	0.7435	46.5	31.1883	34.1808	5.2	10.0
Males 20–39 years	25.6281	1.1678	42.9	23.2729	27.9833	8.7	9.1
Males 40–59 years	39.5681	1.3203	48.4	36.9139	42.2223	2.9	8.7
Males 60+ years	38.4953	1.2986	44.4	35.8790	41.1116	7.2	12.0
All Females (20+ years)	35.5755	0.8278	49.0	33.9120	37.2390	2.0	9.7
Females 20–39 years	28.1727	1.2525	49.9	25.6568	30.6886	0.5	5.8
Females 40–59 years	42.6328	1.1104	45.7	40.3974	44.8682	5.9	8.0
Females 60+ years	40.2731	1.3838	44.8	37.4855	43.0607	6.9	15.5

Table 2 (continued)

	Est.	SE	df	lower	upper	% mis	% imputed
Mean systolic blood pressure							
All Adults (20+ years)	121.1273	0.1933	44.9	120.7380	121.5165	6.8	15.6
M/F 20–39 years	112.6892	0.2606	36.5	112.1609	113.2175	14.6	10.5
M/F 40–59 years	121.7014	0.3624	42.4	120.9702	122.4326	9.2	12.7
M/F 60+ years	137.5690	0.3444	20.0	136.8505	138.2874	33.2	23.6
All Males (20+ years)	123.4720	0.2784	36.9	122.9078	124.0361	14.2	15.4
Males 20–39 years	117.5362	0.3732	21.4	116.7608	118.3116	31.2	12.0
Males 40–59 years	124.3262	0.5226	32.4	123.2622	125.3903	18.4	12.4
Males 60+ years	136.0907	0.5649	37.6	134.9469	137.2346	13.5	21.3
All Females (20+ years)	118.9961	0.2397	47.5	118.5140	119.4783	4.0	15.9
Females 20–39 years	107.9842	0.3065	49.5	107.3685	108.5999	1.3	9.2
Females 40–59 years	119.2107	0.4430	41.8	118.3165	120.1048	9.8	13.0
Females 60+ years	138.6742	0.5344	20.3	137.5606	139.7877	32.6	25.7
Mean serum iron							
All Adults (20+ years)	91.2341	0.5421	43.7	90.1413	92.3269	8.0	14.2
M/F 20–39 years	96.8120	0.9744	36.4	94.8366	98.7873	14.6	12.1
M/F 40–59 years	87.9444	0.6498	38.0	86.6290	89.2599	13.2	11.8
M/F 60+ years	84.3704	0.6476	22.6	83.0294	85.7114	29.5	18.3
All Males (20+ years)	98.6262	0.7558	33.1	97.0887	100.1637	17.7	14.2
Males 20–39 years	104.2946	1.3742	30.1	101.4884	107.1009	20.7	13.6
Males 40–59 years	95.5580	1.0109	26.5	93.4820	97.6340	24.6	12.1
Males 60+ years	90.1035	1.0471	38.9	87.9854	92.2216	12.4	16.4
All Females (20+ years)	84.5153	0.7053	45.3	83.0951	85.9356	6.4	14.3
Females 20–39 years	89.5485	1.3456	34.9	86.8166	92.2804	16.0	10.9
Females 40–59 years	80.7198	0.8978	47.0	78.9138	82.5259	4.6	11.6
Females 60+ years	80.0840	0.7502	9.5	78.4007	81.7673	57.1	20.0
Mean serum cholesterol							
All Adults (20+ years)	204.0814	0.7034	44.5	202.6644	205.4985	7.2	14.7
M/F 20–39 years	188.3679	0.8718	38.9	186.6044	190.1313	12.3	12.6
M/F 40–59 years	213.0604	1.1585	40.1	210.7191	215.4016	11.3	12.2
M/F 60+ years	223.8145	1.0055	40.9	221.7838	225.8453	10.6	18.9
All Males (20+ years)	202.0179	0.7476	47.0	200.5139	203.5218	4.6	14.6
Males 20–39 years	190.5982	1.1209	34.3	188.3208	192.8756	16.6	14.1
Males 40–59 years	212.7610	1.4515	39.7	209.8268	215.6952	11.6	12.3
Males 60+ years	212.0425	1.0617	48.3	209.9081	214.1768	2.9	16.8
All Females (20+ years)	205.9571	0.8715	38.5	204.1935	207.7206	12.7	14.8
Females 20–39 years	186.2028	0.9972	43.2	184.1920	188.2136	8.4	11.3
Females 40–59 years	213.3444	1.3552	38.5	210.6022	216.0866	12.7	12.1
Females 60+ years	232.6161	1.5175	34.3	229.5332	235.6990	16.6	20.7

```
*****************************************************************;
*                                                                ;
*   MI Analysis - Example C                                      ;
*                                                                ;
*   Analysis of NHANES III Multiply Imputed Data Set using       ;
*   SAS-callable SUDAAN, PROC DESCRIPT                           ;
*                                                                ;
*   Estimates median and 90th percentile for                     ;
*      SYSTOLIC = exam systolic BP (avg of three measurements)   ;
*      TCPMI    = serum total cholesterol (mg/dl)                ;
*   for adults by categories of sex and age (20-39, 40-59, 60+)  ;
*                                                                ;
*   Variance estimation by Taylor linearization (WR) method      ;
*                                                                ;
*   Notes:                                                       ;
*                                                                ;
*     (1) This program is only an example. You may need to       ;
*     modify it to suit your needs.                              ;
*                                                                ;
*     (2) This program assumes that the SAS programs             ;
*     CORE.SAS, IMP1.SAS, ..., IMP5.SAS, and NH3MI.SAS have      ;
*     already been run to create the SAS data sets               ;
*     NH3MI1, ..., NH3MI5. These programs are provided on the    ;
*     NHANES III Multiply Imputed Data Set CD-ROM                ;
*                                                                ;
*****************************************************************;

*****************************************************************;
*   Specify the directory where SAS datasets NH3MI1, ..., NH3MI5 ;
*   have been stored. You may need to modify the line below.     ;
*****************************************************************;
LIBNAME NH3MI 'C:\MyDir';

*****************************************************************;
*   This macro cycles through the five imputed data sets,        ;
*   preparing the variables for use by SUDAAN's PROC DESCRIPT.   ;
*   It then calls PROC DESCRIPT to calculate estimates and       ;
*   standard errors for each imputed data set, storing the       ;
*   results in five temporary SAS data sets called               ;
*   SUDNOUT1, ..., SUDNOUT5                                      ;
*****************************************************************;
%MACRO ANALYZE;

   %DO IMPNO = 1 %TO 5;

   DATA FORSUDN (REPLACE=YES);
      SET NH3MI.NH3MI&IMPNO;
      ************************************;
      * categorize age                   ;
      ************************************;
      AGE = HSAGEIR;
      IF HSAGEU = 1 THEN AGE = AGE / 12;
      IF AGE GE 20 AND AGE LE 39 THEN AGEGRP = 1;
      ELSE IF AGE GE 40 AND AGE LE 59 THEN AGEGRP = 2;
      ELSE IF AGE GE 60 THEN AGEGRP = 3;
      ************************************;
      * average systolic blood pressure  ;
      ************************************;
      SYSTOLIC = (PEP6G1MI + PEP6H1MI + PEP6I1MI) / 3;
      ************************************;
```

Figure 3: Program in SAS and SAS-callable SUDAAN for analysis of population medians and percentiles, with standard errors calculated by Taylor linearization

```
      * output select variables for adults  ;
      ************************************;
      KEEP SDPSTRA6 SDPPSU6 WTPFQX6 AGE AGEGRP HSSEX SYSTOLIC TCPMI;
      IF AGE GE 20 THEN OUTPUT;
      RUN;

   ************************************;
   * sort data by pseudo-stratum and PSU ;
   * in preparation for SUDAAN          ;
   * linearization (WR) method          ;
   ************************************;
   PROC SORT DATA=FORSUDN;
      BY SDPSTRA6 SDPPSU6;
      RUN;

   ************************************;
   * call SUDAAN Proc Descript using the ;
   * linearization (WR) method, storing  ;
   * the results in a temporary data set ;
   * called SUDNOUT                      ;
   ************************************;
   PROC DESCRIPT DATA=FORSUDN FILETYPE=SAS DESIGN=WR;
      NEST SDPSTRA6 SDPPSU6 / MISSUNIT;
      WEIGHT WTPFQX6;
      PERCENTILE 90 / MEDIAN;
      VAR SYSTOLIC TCPMI;
      SUBGROUP HSSEX AGEGRP;
      LEVELS 2 3;
      TABLES HSSEX*AGEGRP;
      OUTPUT / PERCENTILE=ALL FILENAME=SUDNOUT FILETYPE=SAS REPLACE;
      RUN;

   ************************************;
   * read in SUDNOUT, renaming the       ;
   * estimates and standard errors       ;
   * to create SUDNOUTi (i=1,2,3,4,5)    ;
   ************************************;
   DATA SUDNOUT&IMPNO;
      SET SUDNOUT;
      EST&IMPNO = QTILE;
      SE&IMPNO = SEQTILE;
      KEEP VARIABLE PCTILES HSSEX AGEGRP EST&IMPNO SE&IMPNO;
      RUN;

   ************************************;
   * sort SUDNOUTi (i=1,2,3,4,5) in      ;
   * preparation for final merge         ;
   ************************************;
   PROC SORT DATA=SUDNOUT&IMPNO;
      BY VARIABLE PCTILES HSSEX AGEGRP;
      RUN;

   %END;
%MEND ANALYZE;

%ANALYZE;

*****************************************************************;
*   Combine the estimates and standard errors stored in          ;
*   SUDNOUTi (i=1,2,3,4,5) using Rubin's rules                   ;
```

Figure 3 (continued)

```
*************************************************************;
DATA COMBINED;
   MERGE SUDNOUT1 SUDNOUT2 SUDNOUT3 SUDNOUT4 SUDNOUT5;
   BY VARIABLE PCTILES HSSEX AGEGRP;
   ************************************;
   * labels for the estimands          ;
   ************************************;
   LENGTH QTYLABEL $25.;
   IF      VARIABLE = 1 THEN DO;
      IF      PCTILES = 1 THEN QTYLABEL = 'Systolic BP: median';
      ELSE IF PCTILES = 2 THEN QTYLABEL = 'Systolic BP: 90th';
      END;
   ELSE IF VARIABLE = 2 THEN DO;
      IF      PCTILES = 1 THEN QTYLABEL = 'Serum cholesterol: median';
      ELSE IF PCTILES = 2 THEN QTYLABEL = 'Serum cholesterol: 90th';
      END;
   ************************************;
   * labels for the demographic groups  ;
   ************************************;
   LENGTH GRPLABEL $25.;
   IF HSSEX = 0 THEN DO;
     IF      AGEGRP=0 THEN GRPLABEL = 'All Adults (20+ years)';
     ELSE IF AGEGRP=1 THEN GRPLABEL = 'M&F 20-39 years';
     ELSE IF AGEGRP=2 THEN GRPLABEL = 'M&F 40-59 years';
     ELSE IF AGEGRP=3 THEN GRPLABEL = 'M&F 60+ years';
     END;
   ELSE IF HSSEX = 1 THEN DO;
     IF      AGEGRP=0 THEN GRPLABEL = 'All Males (20+ years)';
     ELSE IF AGEGRP=1 THEN GRPLABEL = 'Males 20-39 years';
     ELSE IF AGEGRP=2 THEN GRPLABEL = 'Males 40-59 years';
     ELSE IF AGEGRP=3 THEN GRPLABEL = 'Males 60+ years';
     END;
   ELSE IF HSSEX = 2 THEN DO;
     IF      AGEGRP=0 THEN GRPLABEL = 'All Females (20+ years)';
     ELSE IF AGEGRP=1 THEN GRPLABEL = 'Females 20-39 years';
     ELSE IF AGEGRP=2 THEN GRPLABEL = 'Females 40-59 years';
     ELSE IF AGEGRP=3 THEN GRPLABEL = 'Females 60+ years';
     END;
   ************************************;
   * combine results by Rubin's rules   ;
   ************************************;
   EST = MEAN( EST1, EST2, EST3, EST4, EST5);
   WITHNVAR = MEAN(SE1**2, SE2**2, SE3**2, SE4**2, SE5**2);
   BETWNVAR = VAR( EST1, EST2, EST3, EST4, EST5);
   TOTVAR = WITHNVAR + (1 + 1/5)*BETWNVAR;
   SE = TOTVAR**.5;
   ************************************;
   * calculate degrees of freedom for   ;
   * the t-approximation, endpoints of  ;
   * 95% interval, and percent rate of  ;
   * missing information.               ;
   * Degrees of freedom are found       ;
   * by the method of Barnard and Rubin ;
   * (1999) assuming df=49 for complete ;
   * data.                              ;
   ************************************;
   DFCOM = 49;
   IF BETWNVAR GT 0 THEN DO;
      ***********************************;
      * usual case                       ;
```

Figure 3 (continued)

```
      ***********************************;
      DFM = (5-1) * (1 + (5*WITHNVAR)/((5+1)*BETWNVAR))**2;
      DFOBS = ((DFCOM+1)/(DFCOM+3)) * DFCOM * WITHNVAR/TOTVAR;
      DF = 1 / ( 1/DFM + 1/DFOBS );
      LOWER95 = EST - TINV(.975,DF)*SE;
      UPPER95 = EST + TINV(.975,DF)*SE;
      RATIO = ((DF+1)*(DFCOM+3))/((DF+3)*(DFCOM+1));
      PCTMIS = 100*( 1 - RATIO*WITHNVAR/TOTVAR );
      END;
   ELSE IF BETWNVAR EQ 0 THEN DO;
      ***********************************;
      * special case to avoid division by ;
      * zero if between-imputation        ;
      * variance happens to be zero       ;
      ***********************************;
      DF = DFCOM;
      LOWER95 = EST - TINV(.975,DF)*SE;
      UPPER95 = EST + TINV(.975,DF)*SE;
      PCTMIS = 0;
      END;
   FORMAT EST SE LOWER95 UPPER95 8.4 DF 10.1 PCTMIS 5.1;
   RUN;

*****************************************************************;
*   print results                                                ;
*****************************************************************;
OPTIONS LINESIZE=132;
PROC PRINT DATA=COMBINED;
   VAR QTYLABEL GRPLABEL EST SE DF LOWER95 UPPER95 PCTMIS;
   RUN;
```

Figure 3 (continued)

Table 3: Results from MI Analysis Example C—estimates, standard errors, degrees of freedom, lower and upper endpoints of the 95% interval estimates, and estimated percent rate of missing information—with percent rate of imputed values shown for comparison

	Est.	SE	df	lower	upper	% mis	% imputed
Median systolic blood pressure							
All Adults (20+ years)	117.1898	0.3458	44.8	116.4933	117.8863	3.9	15.6
M/F 20–39 years	111.2142	0.3361	41.2	110.5356	111.8928	8.1	10.5
M/F 40–59 years	119.3322	0.3750	42.6	118.5756	120.0887	6.5	12.7
M/F 60+ years	135.2051	0.5915	35.1	134.0044	136.4058	14.1	23.6
All Males (20+ years)	120.3699	0.4279	44.6	119.5079	121.2319	4.2	15.4
Males 20–39 years	116.1555	0.3841	39.1	115.3786	116.9323	10.2	12.0
Males 40–59 years	121.7327	0.4960	40.9	120.7309	122.7345	8.3	12.4
Males 60+ years	133.9107	0.8175	41.9	132.2608	135.5607	7.3	21.3
All Females (20+ years)	113.5631	0.4973	45.2	112.5617	114.5645	3.4	15.9
Females 20–39 years	106.4024	0.4023	43.2	105.5912	107.2135	5.9	9.2
Females 40–59 years	116.7217	0.5923	45.7	115.5293	117.9142	2.7	13.0
Females 60+ years	136.1986	0.8517	34.5	134.4689	137.9284	14.7	25.7
90th percentile systolic blood pressure							
All Adults (20+ years)	146.3037	0.7799	45.7	144.7337	147.8737	2.6	15.6
M/F 20–39 years	127.5741	0.5482	20.7	126.4332	128.7150	31.0	10.5
M/F 40–59 years	143.4287	0.7322	42.9	141.9520	144.9053	6.2	12.7
M/F 60+ years	166.6055	0.8190	25.6	164.9208	168.2903	24.4	23.6
All Males (20+ years)	145.1587	0.9108	44.1	143.3233	146.9942	4.8	15.4
Males 20–39 years	131.3423	0.8008	9.0	129.5293	133.1553	58.4	12.0
Males 40–59 years	144.5345	1.0965	37.9	142.3147	146.7544	11.4	12.4
Males 60+ years	163.1835	1.0776	46.1	161.0146	165.3524	2.0	21.3
All Females (20+ years)	147.5593	1.1422	34.5	145.2394	149.8791	14.7	15.9
Females 20–39 years	121.1338	0.6517	46.1	119.8221	122.4454	2.1	9.2
Females 40–59 years	140.9636	1.1033	27.1	138.7003	143.2269	22.6	13.0
Females 60+ years	168.9626	1.1883	22.3	166.5004	171.4248	28.6	25.7
Median serum cholesterol							
All Adults (20+ years)	199.9297	0.9793	38.5	197.9481	201.9114	10.8	14.7
M/F 20–39 years	184.4993	1.2743	30.7	181.8994	187.0992	18.6	12.6
M/F 40–59 years	208.6696	1.0984	42.4	206.4536	210.8857	6.8	12.2
M/F 60+ years	220.7035	1.2216	34.9	218.2233	223.1838	14.3	18.9
All Males (20+ years)	199.1715	1.1327	39.3	196.8809	201.4621	10.0	14.6
Males 20–39 years	186.9771	1.8699	17.7	183.0446	190.9096	35.9	14.1
Males 40–59 years	209.8829	1.4201	42.2	207.0174	212.7484	7.0	12.3
Males 60+ years	209.3389	1.3064	41.7	206.7018	211.9760	7.5	16.8
All Females (20+ years)	200.6942	1.1890	30.0	198.2660	203.1224	19.3	14.8
Females 20–39 years	182.5324	1.2605	41.6	179.9878	185.0769	7.6	11.3
Females 40–59 years	207.6423	1.5300	30.6	204.5200	210.7645	18.8	12.1
Females 60+ years	229.2949	1.6057	34.2	226.0324	232.5574	15.0	20.7
90th percentile serum cholesterol							
All Adults (20+ years)	259.9572	1.3808	39.5	257.1653	262.7491	9.8	14.7
M/F 20–39 years	237.7845	1.6307	39.4	234.4872	241.0819	9.9	12.6
M/F 40–59 years	266.6022	2.2766	39.9	262.0006	271.2037	9.4	12.2
M/F 60+ years	279.1802	1.8583	38.5	275.4198	282.9407	10.8	18.9
All Males (20+ years)	254.5999	1.5858	37.2	251.3874	257.8123	12.0	14.6
Males 20–39 years	242.7376	2.1240	42.8	238.4535	247.0216	6.3	14.1
Males 40–59 years	263.4706	2.7295	46.1	257.9766	268.9645	2.1	12.3
Males 60+ years	264.4632	2.9354	46.6	258.5565	270.3700	1.2	16.8
All Females (20+ years)	264.5131	1.8355	39.1	260.8007	268.2256	10.2	14.8
Females 20–39 years	232.8625	2.0431	36.9	228.7222	237.0027	12.4	11.3
Females 40–59 years	270.8136	2.4724	46.2	265.8375	275.7896	1.9	12.1
Females 60+ years	287.8816	2.8469	29.5	282.0637	293.6996	19.9	20.7

3.2 Logistic regression example

Many analyses of NHANES III data involve exploration of relationships among variables by techniques such as linear and logistic regression. The NHANES III Multiply Imputed Data Set is well suited for analyses of this type. Rubin's rules for combining point estimates and standard errors apply not only to descriptive statistics such as means, prevalences, and quantiles, but to regression coefficients and other complicated estimates. The approach is no different; the estimates of interest and their standard errors are computed five times, once for each of the completed data files, and the results are combined to yield a single set of estimates and standard errors as described in Section 2.3. The method for obtaining estimates and standard errors from each of the completed data files should take into account the sample weights and the NHANES III sample design.

An example program for performing regression analysis is shown in Figure 4. This example, which uses the SUDAAN logistic regression procedure PROC LOGISTIC, models the probability of being classified as overweight by weighted logistic regression with standard errors obtained by the replicate method. The covariates in this model include sex, age group (20–39, 40–59, 60+), a race/ethnicity classification, poverty status, and responses to two key questions on the NHANES III adult questionnaire: self-reported health status (excellent, very good, good, fair, poor) and self-reported activity level compared to others (more active, less active, about the same). Each of these covariates is categorical and is included in the model via a set of dummy indicators. The reference level for each covariate corresponding to the omitted dummy variable is specified by the `REFLEVEL` command in PROC LOGISTIC.

The results from this example are displayed in Table 4. The columns of this table contain the estimated logistic regression coefficients, standard errors, degrees of freedom, T-ratios (the estimated coefficients divided by their standard errors), p-values for testing whether the population coefficients are zero, and estimated percent rates of missing information. Note that the rates of missing information vary considerably among the coefficients. This is to be expected, because the percentages of missing values for the covariates also vary considerably.

This example nicely illustrates one of the practical advantages of multiple imputation. If

a similar logistic regression analysis were performed with the previously released NHANES III public use files, one would have to omit from the procedure any individual who had a missing value for examination height or weight (from which the response indicator is derived), poverty status, self-reported health status, or self-reported activity level. These restrictions would remove from the procedure any individual who was interviewed but not examined, and any individual who failed to respond to one or more of the interview questions pertaining to household income, health status or activity level. With the NHANES III Multiply Imputed Data Set, however, the analysis proceeds very simply using all 18,825 interviewed adults.

The method for combining estimates and standard errors described in Section 2.3 can be extended to permit joint inferences about groups of estimands. This is helpful, for example, for addressing the joint significance of a group of covariates in a logistic regression model. Joint inferences involve combining vectors of estimates and their associated covariance matrices across the five completed data files. The rules are relatively simple extensions of those described above; for details, refer to Barnard and Rubin (1999).

```
*****************************************************************;
*                                                                ;
*   MI Analysis - Example D                                      ;
*                                                                ;
*   Logistic regression analysis of the NHANES III Multiply      ;
*   Imputed Data Set using SAS-callable SUDAAN, PROC LOGISTIC    ;
*                                                                ;
*   Models the log-odds of being classified as overweight        ;
*   (adults only) given the following covariates:                ;
*                                                                ;
*       sex (1=Male, 2=Female)                                   ;
*       age group (1=20-39, 2=40-59, 3=60+)                      ;
*       race-ethnicity (1=non-Hispanic white/other, non-         ;
*          Hispanic black, Mexican-American)                     ;
*       self-rating of health status (1=excellent, 2=very good,  ;
*          3=good, 4=fair, 5=poor)                               ;
*       compare own activity level to others (1=more active,     ;
*          2=less active, 3=about the same)                      ;
*       poverty status (1=at or below poverty line, 2=above)     ;
*                                                                ;
*   Variance estimation by replicate (BRR) method with Fay       ;
*   adjustment.                                                  ;
*                                                                ;
*   Notes:                                                       ;
*                                                                ;
*     (1) This program is only an example. You may need to       ;
*     modify it to suit your needs.                              ;
*                                                                ;
*     (2) This program assumes that the SAS programs             ;
*     CORE.SAS, IMP1.SAS, ..., IMP5.SAS, and NH3MI.SAS have      ;
*     already been run to create the SAS data sets               ;
*     NH3MI1, ..., NH3MI5. These programs are provided on the    ;
*     NHANES III Multiply Imputed Data Set CD-ROM                ;
*                                                                ;
*****************************************************************;

*****************************************************************;
*   Specify the directory where SAS datasets NH3MI1, ..., NH3MI5 ;
*   have been stored. You may need to modify the line below.     ;
*****************************************************************;
LIBNAME NH3MI 'C:\MyDir';

*****************************************************************;
*   This macro cycles through the five imputed data sets,        ;
*   preparing the variables for use by SUDAAN's PROC LOGISTIC.   ;
*   It then calls PROC LOGISTIC to calculate estimates and       ;
*   standard errors for each imputed data set, storing the       ;
*   results in five temporary SAS data sets called               ;
*   SUDNOUT1, ..., SUDNOUT5                                      ;
*****************************************************************;
%MACRO ANALYZE;

   %DO IMPNO = 1 %TO 5;

   DATA FORSUDN (REPLACE=YES);
      SET NH3MI.NH3MI&IMPNO;
      ************************************;
      * calculate body mass index        ;
      * and overweight indicator         ;
      ************************************;
```

Figure 4: Program in SAS and SAS-callable SUDAAN for logistic regression analysis, with standard errors calculated by replicate method

```
   BMI = BMPWTMI / (BMPHTMI/100)**2;
   OVERWT = 0;
   IF HSSEX = 1 AND BMI GE 27.8 THEN OVERWT = 1;
   IF HSSEX = 2 AND BMI GE 27.3 THEN OVERWT = 1;
   ************************************;
   * categorize age                   ;
   ************************************;
   AGE = HSAGEIR;
   IF HSAGEU = 1 THEN AGE = AGE / 12;
   IF AGE GE 20 AND AGE LE 39 THEN AGEGRP = 1;
   ELSE IF AGE GE 40 AND AGE LE 59 THEN AGEGRP = 2;
   ELSE IF AGE GE 60 THEN AGEGRP = 3;
   ************************************;
   * race-ethnicity classification    ;
   ************************************;
   RACEETHN = DMARETHN;
   IF DMARETHN = 4 THEN RACEETHN = 1;
   ************************************;
   * poverty status classification    ;
   ************************************;
   IF DMPPIRMI LE 1.0 THEN POVERTY = 1;
   ELSE IF DMPPIRMI GT 1.0 THEN POVERTY = 2;
   ************************************;
   * output select variables for adults  ;
   ************************************;
   KEEP WTPFQX6 WTPQRP1-WTPQRP52 AGE AGEGRP HSSEX RACEETHN
      HAB1MI HAT28MI POVERTY OVERWT;
   IF AGE GE 20 THEN OUTPUT;
   RUN;

*******************************************************;
* Because this is the SAS-callable version, SUDAAN     ;
* PROC LOGISTIC is invoked as PROC RLOGIST             ;
*******************************************************;
PROC RLOGIST SUDDATA=FORSUDN FILETYPE=SAS DESIGN=BRR;
   WEIGHT WTPFQX6;
   REPWGT WTPQRP1--WTPQRP52 / ADJFAY=2.0408;
   SUBGROUP HSSEX AGEGRP RACEETHN HAB1MI HAT28MI POVERTY;
   LEVELS 2 3 3 5 3 2;
   REFLEVEL HSSEX=1 AGEGRP=1 RACEETHN=1 HAB1MI=1 HAT28MI=3 POVERTY=2;
   MODEL OVERWT = HSSEX AGEGRP RACEETHN HAB1MI HAT28MI POVERTY;
   OUTPUT / BETAS=DEFAULT FILENAME=SUDNOUT FILETYPE=SAS REPLACE;
   RUN;

************************************;
* read in SUDNOUT, renaming the       ;
* estimates and standard errors       ;
* to create SUDNOUTi (i=1,2,3,4,5)    ;
************************************;
DATA SUDNOUT&IMPNO;
   SET SUDNOUT;
   EST&IMPNO = BETA;
   SE&IMPNO = SEBETA;
   KEEP MODELRHS EST&IMPNO SE&IMPNO;
   RUN;

************************************;
* sort SUDNOUTi (i=1,2,3,4,5) in      ;
* preparation for final merge         ;
************************************;
```

Figure 4 (continued)

```
   PROC SORT DATA=SUDNOUT&IMPNO;
      BY MODELRHS;
      RUN;

   %END;
%MEND ANALYZE;

%ANALYZE;

*****************************************************************;
*   Combine the estimates and standard errors stored in         ;
*   SUDNOUTi (i=1,2,3,4,5) using Rubin's rules                  ;
*****************************************************************;
DATA COMBINED;
   MERGE SUDNOUT1 SUDNOUT2 SUDNOUT3 SUDNOUT4 SUDNOUT5;
   BY MODELRHS;
   ************************************;
   * labels for the coefficients      ;
   ************************************;
   LENGTH QTYLABEL $25.;
   IF      MODELRHS =  1 THEN QTYLABEL = 'Intercept';
   ELSE IF MODELRHS =  2 THEN QTYLABEL = 'Male';
   ELSE IF MODELRHS =  3 THEN QTYLABEL = 'Female';
   ELSE IF MODELRHS =  4 THEN QTYLABEL = 'Age 20-39';
   ELSE IF MODELRHS =  5 THEN QTYLABEL = 'Age 40-59';
   ELSE IF MODELRHS =  6 THEN QTYLABEL = 'Age 60+';
   ELSE IF MODELRHS =  7 THEN QTYLABEL = 'Non-Hispanic white/other';
   ELSE IF MODELRHS =  8 THEN QTYLABEL = 'Non-Hispanic black';
   ELSE IF MODELRHS =  9 THEN QTYLABEL = 'Mexican-American';
   ELSE IF MODELRHS = 10 THEN QTYLABEL = 'Health excellent';
   ELSE IF MODELRHS = 11 THEN QTYLABEL = 'Health very good';
   ELSE IF MODELRHS = 12 THEN QTYLABEL = 'Health good';
   ELSE IF MODELRHS = 13 THEN QTYLABEL = 'Health fair';
   ELSE IF MODELRHS = 14 THEN QTYLABEL = 'Health poor';
   ELSE IF MODELRHS = 15 THEN QTYLABEL = 'More active than others';
   ELSE IF MODELRHS = 16 THEN QTYLABEL = 'Less active than others';
   ELSE IF MODELRHS = 17 THEN QTYLABEL = 'About the same';
   ELSE IF MODELRHS = 18 THEN QTYLABEL = 'At or below poverty line';
   ELSE IF MODELRHS = 19 THEN QTYLABEL = 'Above poverty line';
   ************************************;
   * combine results by Rubin's rules  ;
   ************************************;
   EST = MEAN( EST1, EST2, EST3, EST4, EST5);
   WITHNVAR = MEAN(SE1**2, SE2**2, SE3**2, SE4**2, SE5**2);
   BETWNVAR = VAR( EST1, EST2, EST3, EST4, EST5);
   TOTVAR = WITHNVAR + (1 + 1/5)*BETWNVAR;
   SE = TOTVAR**.5;
   TRATIO = EST/SE;
   ************************************;
   * calculate degrees of freedom for  ;
   * the t-approximation, endpoints of ;
   * 95% interval, and percent rate of ;
   * missing information.              ;
   * Degrees of freedom are found      ;
   * by the method of Barnard and Rubin ;
   * (1999) assuming df=52 for complete ;
   * data.                             ;
   ************************************;
   DFCOM = 52;
   IF BETWNVAR GT 0 THEN DO;
```

Figure 4 (continued)

```
      ***********************************;
      * usual case                       ;
      ***********************************;
      DFM = (5-1) * (1 + (5*WITHNVAR)/((5+1)*BETWNVAR))**2;
      DFOBS = ((DFCOM+1)/(DFCOM+3)) * DFCOM * WITHNVAR/TOTVAR;
      DF = 1 / ( 1/DFM + 1/DFOBS );
      PVALUE = 2 * (1 -PROBT( ABS(TRATIO),DF ) );
      RATIO = ((DF+1)*(DFCOM+3))/((DF+3)*(DFCOM+1));
      PCTMIS = 100*( 1 - RATIO*WITHNVAR/TOTVAR );
      END;
   ELSE IF BETWNVAR EQ 0 THEN DO;
      ***********************************;
      * special case to avoid division by ;
      * zero if between-imputation       ;
      * variance happens to be zero      ;
      ***********************************;
      DF = DFCOM;
      PVALUE = 2 * (1 -PROBT( ABS(TRATIO),DF ) );
      PCTMIS = 0;
      END;
   FORMAT EST SE 8.4 TRATIO 6.2 PVALUE 6.4 DF 4.1 PCTMIS 5.1;
   RUN;

*****************************************************************;
*   print results                                                ;
*****************************************************************;
OPTIONS LINESIZE=132;
PROC PRINT DATA=COMBINED;
   VAR QTYLABEL EST SE DF TRATIO PVALUE PCTMIS;
   RUN;
```

Figure 4 (continued)

Table 4: Results from MI Analysis Example D—estimated coefficients, standard errors, degrees of freedom, T-ratios, p-values, and estimated percent rate of missing information

	Est.	SE	df	T-ratio	p-value	% mis
Intercept	−1.5411	0.0814	41.8	−18.94	0.0000	9.7
Male	0.0000	0.0000	52.0	—	—	0.0
Female	0.0491	0.0508	46.4	0.97	0.3386	5.2
Age 20–39	0.0000	0.0000	52.0	—	—	0.0
Age 40–59	0.6964	0.0554	41.9	12.57	0.0000	9.7
Age 60+	0.6065	0.0761	47.7	7.97	0.0000	3.7
Non-Hispanic white/other	0.0000	0.0000	52.0	—	—	0.0
Non-Hispanic black	0.4481	0.0591	48.8	7.58	0.0000	2.4
Mexican-American	0.4103	0.0660	48.5	6.22	0.0000	2.7
Health excellent	0.0000	0.0000	52.0	—	—	0.0
Health very good	0.4281	0.0722	47.3	5.93	0.0000	4.3
Health good	0.6852	0.0778	46.1	8.81	0.0000	5.6
Health fair	0.7669	0.0827	47.9	9.27	0.0000	3.5
Health poor	0.3757	0.1208	46.1	3.11	0.0032	5.5
More active than others	−0.4014	0.0647	37.7	−6.21	0.0000	13.4
Less active than others	0.2732	0.0572	48.2	4.77	0.0000	3.1
About the same	0.0000	0.0000	52.0	—	—	0.0
At or below poverty line	−0.0160	0.0638	31.7	−0.25	0.8035	19.1
Above poverty line	0.0000	0.0000	52.0	—	—	0.0

4 Comparisons with analyses of previously released NHANES III files

4.1 Estimates of means and prevalences

In this section we compare the results of some of our example analyses of the NHANES III Multiply Imputed Data Set to those of conventional analyses of the previously released NHANES III public use files (DHHS, CD-ROM, Series 11, Number 1A, 1997; Number 2A, 1998). In those files, adjustment factors for unit nonresponse were incorporated into the sample weights, but no adjustments were provided for item nonresponse. Subjects whose data values were missing because of refusal, responses of 'Don't know,' etc. have traditionally been omitted from analyses of these files.

A conventional analysis of population means and prevalence rates, which corresponds to Example A of the previous section (Figure 1 and Table 1), is shown in Figure 5. In this example, the relevant variables are extracted from the NHANES III examination and laboratory results data files and merged into a single data set. Because the variables in this analysis were collected during the NHANES III examination, weighted estimates are calculated using 'final examination weight' (variable `WTPFEX6`) which includes adjustments for unit nonresponse at the examination stage. Nevertheless, the variables in question still contain some missing items (denoted in the files by 8-fills) for which no adjustments were made; these 8-fills are changed to the SAS missing value code '`.`' and subsequently ignored. Weighted estimates and standard errors are computed using SUDAAN's PROC DESCRIPT by the Taylor linearization method.

The results from this analysis are displayed in Table 5. Comparing these results to those of the multiple-imputation analysis shown in Table 1, we see that in some respects they are quite similar. The point estimates in Table 5 agree with those from Table 1 to within 4% of their values. In all cases, the discrepancy between the two estimates is no more than one-third of the size of the standard error reported in Table 1. The differences in standard errors are somewhat more substantial. The standard errors reported in Table 5 are on average about 3% wider than those of Table 1, but the discrepancies vary considerably; in one case the standard error in Table 5 is 13% smaller than the corresponding value in Table 1, and in

another case it is 27% larger.

How should one interpret the discrepancies in results from the two methods? On average, the conventional analysis seems to produce point estimates that are similar to, and interval estimates that are slightly wider than, those from the multiple-imputation analysis. At first glance, this might suggest that the two methods have essentially similar properties, except that the multiple-imputation analysis might on average be slighly more precise than the conventional analysis. But this interpretation is not entirely correct, because the validity of a procedure is based not on the results of a single application but on its performance in repeated use over many applications.

The operating characteristics of a statistical procedure arise from the subtle interplay between (a) the actual bias and variability of the estimation method, and (b) the accuracy of the method for calculating the standard errors and confidence intervals, over repeated application in many samples. It is desirable to have an estimation method which on average produces an estimate close to the true population value. It is also desirable to have an interval estimation procedure which on average covers the true population value with the advertised probability (e.g. 95%) and which also produces an interval that is as narrow as possible. There is no way to tell, merely by examining the results from Table 5 or Table 1, whether the interval-estimation procedures are performing as they should or whether one method is superior to the other. The only way to assess performance is by analytic arguments and by empirical simulation studies. For discussion on the theoretical properties of multiple imputation and its advantages over conventional methods, see Rubin (1987, 1996), Meng (1994) and their references. An extensive simulation study demonstrating the good performance of multiple imputation in NHANES-style surveys is described Little et al. (1995).

```
*****************************************************************;
*                                                                ;
*   Conventional Analysis - Example A                            ;
*                                                                ;
*   Analysis of NHANES III public-use data files using           ;
*   SAS-callable SUDAAN, PROC DESCRIPT                           ;
*                                                                ;
*   Estimates means for                                          ;
*      BDPFNBMD = Bone mineral density femur neck (gm/cm sq)     ;
*      BMPWAIST = Waist circumference (cm)                       ;
*      BMPBMI   = Body mass index                                ;
*      OVERWT   = 100 if overweight, 0 else (from BMPBMI)        ;
*      SYSTOLIC = exam systolic BP (avg of three measurements)   ;
*      FEP      = serum iron (ug/dl)                             ;
*      TCP      = serum total cholesterol (mg/dl)                ;
*   for adults by categories of sex and age (20-39, 40-59, 60+)  ;
*                                                                ;
*   Variance estimation by Taylor linearization (WR) method      ;
*                                                                ;
*****************************************************************;

*****************************************************************;
*   Specify the paths for NHANES III public-use ASCII data       ;
*   on CD-ROM, EXAM and LAB files. You will need to modify       ;
*   these paths if your CD-ROM is not drive E:                   ;
*****************************************************************;
FILENAME EXAM "E:\EXAM\EXAM.DAT" LRECL=6235;
FILENAME LAB "E:\LAB\LAB.DAT" LRECL=1979;

*****************************************************************;
*   Read in select variables from the NHANES III public-use      ;
*   file EXAM.DAT                                                ;
*****************************************************************;
DATA EXAMVARS;
   INFILE EXAM MISSOVER;
   INPUT
      SEQN     1-5
      HSSEX    15
      HSAGEIR  16-17
      HSAGEU   18
      SDPPSU6  41
      SDPSTRA6 42-43
      WTPFEX6  59-67
      PEP6G1   1393-1395
      PEP6H1   1403-1405
      PEP6I1   1413-1415
      BMPBMI   1524-1527
      BMPWAIST 1590-1594
      BDPFNBMD 5276-5280;
   LABEL
      SEQN     = "Respondent identification number"
      HSSEX    = "Sex"
      HSAGEIR  = "Age at interview (Screener)"
      HSAGEU   = "Age at interview-unit (Screener)"
      SDPPSU6  = "Total NHANES III pseudo-PSU"
      SDPSTRA6 = "Total NHANES III pseudo-stratum"
      WTPFEX6  = "Total MEC-examined sample final weight"
      PEP6G1   = "K1, systolic, for 1st BP (mmHg)(age 5+)"
      PEP6H1   = "K1, systolic, for 2nd BP (mmHg)(age5+)"
      PEP6I1   = "K1, systolic, for 3rd BP (mmHg)(age 5+)"
```

Figure 5: Program in SAS and SAS-callable SUDAAN for conventional analysis of population means and prevalence rates

```
      BMPBMI   = "Body mass index"
      BMPWAIST = "Waist circumference (cm) (2+ years)"
      BDPFNBMD = "Bone mineral density femur neck-gm/cm sq";
   RUN;

*****************************************************************;
*   Read in select variables from the NHANES III public-use     ;
*   file LAB.DAT                                                ;
*****************************************************************;
DATA LABVARS;
   INFILE LAB MISSOVER;
   INPUT
      SEQN     1-5
      HSSEX    15
      HSAGEIR  16-17
      HSAGEU   18
      SDPPSU6  41
      SDPSTRA6 42-43
      WTPFEX6  59-67
      FEP      1441-1443
      TCP      1598-1600;
   LABEL
      SEQN     = "Respondent identification number"
      HSSEX    = "Sex"
      HSAGEIR  = "Age at interview (Screener)"
      HSAGEU   = "Age at interview-unit (Screener)"
      SDPPSU6  = "Total NHANES III pseudo-PSU"
      SDPSTRA6 = "Total NHANES III pseudo-stratum"
      WTPFEX6  = "Total MEC-examined sample final weight"
      FEP      = "Serum iron (ug/dL)"
      TCP      = "Serum cholesterol (mg/dL)";
   RUN;

*****************************************************************;
*   Merge the two files by SEQN                                 ;
*****************************************************************;
PROC SORT DATA=EXAMVARS;
   BY SEQN;
   RUN;

PROC SORT DATA=LABVARS;
   BY SEQN;
   RUN;

DATA BOTH;
   MERGE EXAMVARS LABVARS;
   BY SEQN;
   RUN;

*****************************************************************;
*   Prepare variables for use by SUDAAN's PROC DESCRIPT         ;
*****************************************************************;
DATA FORSUDN;
   SET BOTH;
   ************************************;
   * categorize age                   ;
   ************************************;
   AGE = HSAGEIR;
   IF HSAGEU = 1 THEN AGE = AGE / 12;
   IF AGE GE 20 AND AGE LE 39 THEN AGEGRP = 1;
```

Figure 5 (continued)

```
   ELSE IF AGE GE 40 AND AGE LE 59 THEN AGEGRP = 2;
   ELSE IF AGE GE 60 THEN AGEGRP = 3;
   ELSE AGEGRP = 0;
   ************************************;
   * recode 8-fills as missing values  ;
   ************************************;
   IF PEP6G1   =   888 THEN PEP6G1   = .;
   IF PEP6H1   =   888 THEN PEP6H1   = .;
   IF PEP6I1   =   888 THEN PEP6I1   = .;
   IF BMPBMI   =  8888 THEN BMPBMI   = .;
   IF BMPWAIST = 88888 THEN BMPWAIST = .;
   IF BDPFNBMD = 88888 THEN BDPFNBMD = .;
   IF FEP      =   888 THEN FEP = .;
   IF TCP      =   888 THEN TCP = .;
   ************************************;
   * calculate overweight indicator    ;
   ************************************;
   OVERWT = 0;
   IF HSSEX = 1 AND BMPBMI GE 27.8 THEN OVERWT = 100;
   IF HSSEX = 2 AND BMPBMI GE 27.3 THEN OVERWT = 100;
   ************************************;
   * average systolic blood pressure   ;
   ************************************;
   SYSTOLIC = (PEP6G1 + PEP6H1 + PEP6I1) / 3;
   ************************************;
   * output select variables for       ;
   * examined adults                   ;
   ************************************;
   KEEP SDPSTRA6 SDPPSU6 WTPFEX6 AGE AGEGRP HSSEX
      BDPFNBMD BMPWAIST BMPBMI OVERWT SYSTOLIC FEP TCP;
   IF AGE GE 20 AND WTPFEX6 GT 0 THEN OUTPUT;
   RUN;

************************************;
* sort data by pseudo-stratum and PSU ;
* in preparation for SUDAAN          ;
* linearization (WR) method          ;
************************************;
PROC SORT DATA=FORSUDN;
   BY SDPSTRA6 SDPPSU6;
   RUN;

************************************;
* call SUDAAN Proc Descript using the ;
* linearization (WR) method, storing  ;
* the results in a temporary data set ;
* called SUDNOUT                      ;
************************************;
PROC DESCRIPT DATA=FORSUDN FILETYPE=SAS DESIGN=WR MEANS;
   NEST SDPSTRA6 SDPPSU6 / MISSUNIT;
   WEIGHT WTPFEX6;
   VAR BDPFNBMD BMPWAIST BMPBMI OVERWT SYSTOLIC FEP TCP;
   SUBGROUP HSSEX AGEGRP;
   LEVELS 2 3;
   TABLES HSSEX*AGEGRP;
   OUTPUT MEAN SEMEAN / FILENAME=SUDNOUT FILETYPE=SAS REPLACE;
   RUN;

************************************;
* read in SUDNOUT, arrange results  ;
```

Figure 5 (continued)

```
************************************;
DATA RESULTS;
   SET SUDNOUT;
   EST = MEAN;
   SE = SEMEAN;
   ************************************;
   * labels for the estimands         ;
   ************************************;
   LENGTH QTYLABEL $25.;
   IF      VARIABLE = 1 THEN QTYLABEL = 'Mean BMD femur neck';
   ELSE IF VARIABLE = 2 THEN QTYLABEL = 'Mean waist circumference';
   ELSE IF VARIABLE = 3 THEN QTYLABEL = 'Mean body mass index';
   ELSE IF VARIABLE = 4 THEN QTYLABEL = 'Pct overweight';
   ELSE IF VARIABLE = 5 THEN QTYLABEL = 'Mean systolic BP';
   ELSE IF VARIABLE = 6 THEN QTYLABEL = 'Mean serum iron';
   ELSE IF VARIABLE = 7 THEN QTYLABEL = 'Mean serum cholesterol';
   ************************************;
   * labels for the demographic groups  ;
   ************************************;
   LENGTH GRPLABEL $25.;
   IF HSSEX = 0 THEN DO;
     IF      AGEGRP=0 THEN GRPLABEL = 'All Adults (20+ years)';
     ELSE IF AGEGRP=1 THEN GRPLABEL = 'M&F 20-39 years';
     ELSE IF AGEGRP=2 THEN GRPLABEL = 'M&F 40-59 years';
     ELSE IF AGEGRP=3 THEN GRPLABEL = 'M&F 60+ years';
     END;
   ELSE IF HSSEX = 1 THEN DO;
     IF      AGEGRP=0 THEN GRPLABEL = 'All Males (20+ years)';
     ELSE IF AGEGRP=1 THEN GRPLABEL = 'Males 20-39 years';
     ELSE IF AGEGRP=2 THEN GRPLABEL = 'Males 40-59 years';
     ELSE IF AGEGRP=3 THEN GRPLABEL = 'Males 60+ years';
     END;
   ELSE IF HSSEX = 2 THEN DO;
     IF      AGEGRP=0 THEN GRPLABEL = 'All Females (20+ years)';
     ELSE IF AGEGRP=1 THEN GRPLABEL = 'Females 20-39 years';
     ELSE IF AGEGRP=2 THEN GRPLABEL = 'Females 40-59 years';
     ELSE IF AGEGRP=3 THEN GRPLABEL = 'Females 60+ years';
     END;
   ************************************;
   * calculate endpoints of 95% interval ;
   * using t-distribution with df = 49   ;
   ************************************;
   DF = 49;
   LOWER95 = EST - TINV(.975,DF)*SE;
   UPPER95 = EST + TINV(.975,DF)*SE;
   FORMAT EST SE LOWER95 UPPER95 8.4 DF 10.1;
   RUN;

*****************************************************************;
*   print results                                                ;
*****************************************************************;
OPTIONS LINESIZE=132;
PROC PRINT DATA=RESULTS;
   VAR QTYLABEL GRPLABEL EST SE DF LOWER95 UPPER95;
   RUN;
```

Figure 5 (continued)

Table 5: Results from Conventional Analysis Example A—estimates, standard errors, degrees of freedom, lower and upper endpoints of 95% interval estimates

	Est.	SE	df	lower	upper
Mean BMD femur neck					
All Adults (20+ years)	0.8231	0.0032	49.0	0.8167	0.8295
M/F 20–39 years	0.8944	0.0026	49.0	0.8891	0.8997
M/F 40–59 years	0.8103	0.0031	49.0	0.8041	0.8166
M/F 60+ years	0.7013	0.0037	49.0	0.6939	0.7087
All Males (20+ years)	0.8688	0.0030	49.0	0.8627	0.8749
Males 20–39 years	0.9310	0.0033	49.0	0.9243	0.9376
Males 40–59 years	0.8378	0.0040	49.0	0.8299	0.8458
Males 60+ years	0.7700	0.0045	49.0	0.7610	0.7790
All Females (20+ years)	0.7786	0.0039	49.0	0.7707	0.7865
Females 20–39 years	0.8539	0.0037	49.0	0.8465	0.8612
Females 40–59 years	0.7835	0.0040	49.0	0.7756	0.7915
Females 60+ years	0.6495	0.0041	49.0	0.6412	0.6577
Mean waist circumference					
All Adults (20+ years)	91.8908	0.2418	49.0	91.4048	92.3768
M/F 20–39 years	87.4279	0.3207	49.0	86.7833	88.0724
M/F 40–59 years	95.1082	0.3759	49.0	94.3527	95.8637
M/F 60+ years	96.7181	0.2734	49.0	96.1687	97.2674
All Males (20+ years)	95.3062	0.2604	49.0	94.7828	95.8295
Males 20–39 years	90.7536	0.3888	49.0	89.9723	91.5348
Males 40–59 years	98.8187	0.3567	49.0	98.1019	99.5355
Males 60+ years	100.5990	0.3481	49.0	99.8994	101.2985
All Females (20+ years)	88.7752	0.3848	49.0	88.0019	89.5486
Females 20–39 years	84.2114	0.5005	49.0	83.2055	85.2173
Females 40–59 years	91.5637	0.5664	49.0	90.4254	92.7020
Females 60+ years	93.7901	0.3754	49.0	93.0358	94.5444
Mean body mass index					
All Adults (20+ years)	26.5207	0.1102	49.0	26.2993	26.7422
M/F 20–39 years	25.6412	0.1428	49.0	25.3542	25.9282
M/F 40–59 years	27.5412	0.1571	49.0	27.2256	27.8569
M/F 60+ years	26.9144	0.1170	49.0	26.6793	27.1495
All Males (20+ years)	26.6008	0.1104	49.0	26.3790	26.8226
Males 20–39 years	25.8760	0.1514	49.0	25.5717	26.1803
Males 40–59 years	27.5125	0.1598	49.0	27.1913	27.8337
Males 60+ years	26.8779	0.1373	49.0	26.6019	27.1538
All Females (20+ years)	26.4480	0.1575	49.0	26.1314	26.7646
Females 20–39 years	25.4133	0.2097	49.0	24.9918	25.8347
Females 40–59 years	27.5685	0.2353	49.0	27.0956	28.0414
Females 60+ years	26.9418	0.1467	49.0	26.6470	27.2366
Percent overweight					
All Adults (20+ years)	34.8851	0.6950	49.0	33.4884	36.2818
M/F 20–39 years	27.8049	0.9616	49.0	25.8725	29.7372
M/F 40–59 years	41.5506	1.1439	49.0	39.2518	43.8494
M/F 60+ years	40.1713	1.1196	49.0	37.9214	42.4212
All Males (20+ years)	33.3356	0.8845	49.0	31.5581	35.1130
Males 20–39 years	26.3661	1.1777	49.0	23.9995	28.7327
Males 40–59 years	40.0124	1.3705	49.0	37.2584	42.7665
Males 60+ years	39.2653	1.4440	49.0	36.3634	42.1672
All Females (20+ years)	36.2935	0.9781	49.0	34.3279	38.2592
Females 20–39 years	29.2015	1.5359	49.0	26.1151	32.2879
Females 40–59 years	43.0102	1.5465	49.0	39.9023	46.1181
Females 60+ years	40.8487	1.5229	49.0	37.7883	43.9090

Table 5 (continued)

	Est.	SE	df	lower	upper
Mean systolic blood pressure					
All Adults (20+ years)	120.8019	0.4020	49.0	119.9940	121.6097
M/F 20–39 years	112.6410	0.2823	49.0	112.0738	113.2083
M/F 40–59 years	121.4090	0.3832	49.0	120.6388	122.1791
M/F 60+ years	136.9561	0.5085	49.0	135.9342	137.9780
All Males (20+ years)	123.1275	0.4393	49.0	122.2447	124.0102
Males 20–39 years	117.3881	0.3903	49.0	116.6038	118.1723
Males 40–59 years	124.0380	0.5418	49.0	122.9491	125.1268
Males 60+ years	135.5036	0.6816	49.0	134.1338	136.8733
All Females (20+ years)	118.6862	0.5227	49.0	117.6357	119.7367
Females 20–39 years	108.0326	0.3209	49.0	107.3877	108.6775
Females 40–59 years	118.8992	0.5092	49.0	117.8759	119.9224
Females 60+ years	138.0333	0.6432	49.0	136.7407	139.3259
Mean serum iron					
All Adults (20+ years)	91.2931	0.7213	49.0	89.8436	92.7427
M/F 20–39 years	96.7025	1.0207	49.0	94.6514	98.7537
M/F 40–59 years	88.0736	0.9260	49.0	86.2126	89.9345
M/F 60+ years	84.8436	0.8021	49.0	83.2317	86.4555
All Males (20+ years)	98.4621	0.7274	49.0	97.0004	99.9238
Males 20–39 years	103.8286	1.2379	49.0	101.3409	106.3163
Males 40–59 years	95.3632	1.0762	49.0	93.2006	97.5258
Males 60+ years	90.7811	1.2450	49.0	88.2792	93.2829
All Females (20+ years)	84.6862	0.9056	49.0	82.8662	86.5062
Females 20–39 years	89.6269	1.3886	49.0	86.8364	92.4175
Females 40–59 years	81.0888	1.2959	49.0	78.4846	83.6929
Females 60+ years	80.3906	0.8741	49.0	78.6340	82.1472
Mean serum cholesterol					
All Adults (20+ years)	204.3741	0.7638	49.0	202.8392	205.9090
M/F 20–39 years	188.3648	0.9843	49.0	186.3868	190.3429
M/F 40–59 years	213.3335	1.0322	49.0	211.2593	215.4077
M/F 60+ years	224.2486	1.0902	49.0	222.0579	226.4394
All Males (20+ years)	202.1969	0.9453	49.0	200.2972	204.0965
Males 20–39 years	190.7335	1.2552	49.0	188.2110	193.2560
Males 40–59 years	212.9113	1.2919	49.0	210.3152	215.5074
Males 60+ years	212.0264	1.3248	49.0	209.3640	214.6887
All Females (20+ years)	206.3780	0.9161	49.0	204.5371	208.2189
Females 20–39 years	186.0224	1.0760	49.0	183.8601	188.1848
Females 40–59 years	213.7382	1.2061	49.0	211.3144	216.1621
Females 60+ years	233.4301	1.5039	49.0	230.4078	236.4524

4.2 Logistic regression example

Our final example, shown in Figure 6, replicates the logistic regression analysis described earlier (Figure 4) using the previously released NHANES III data files. In this example, the relevant variables are extracted from the NHANES III examination and adult questionnaire data files and merged into a single data set. Missing items denoted by 8- and 9-fills are converted to the SAS missing value code, and the logistic model is fit by SUDAAN's PROC LOGISTIC using the replicate method of variance estimation. Because this conventional analysis involves variables from both the interview and the examination, the 'final examination weight' (variable `WTPFEX6`) is used for estimation, and the replicate examination weights (`WTPXRP1--WTPXRP52`) are used for variance estimation.

The results obtained from this program are displayed in Table 6. After omitting subjects with missing values on any of the required variables, the model-fitting procedure used data from 16,327 subjects. Comparing these results to those of Table 4, we see that the discrepancies among the estimates are not necessarily small; three coefficients have changed by more than 70% of their standard errors. The standard errors from the two procedures on average are approximately the same size. In general, one should expect that for more complicated analyses involving many variables at once, the discrepancies in results between the two methods will become larger, because as more variables are included the proportion of cases that will be discarded in conventional analyses tends to grow.

```
*****************************************************************;
*                                                                ;
*   Conventional Analysis - Example D                            ;
*                                                                ;
*   Analysis of NHANES III public-use data files using           ;
*   SAS-callable SUDAAN, PROC LOGISTIC                           ;
*                                                                ;
*   Models the log-odds of being classified as overweight        ;
*   (adults only) given the following covariates:                ;
*                                                                ;
*       sex (1=Male, 2=Female)                                   ;
*       age group (1=20-39, 2=40-59, 3=60+)                      ;
*       race-ethnicity (1=non-Hispanic white/other, non-         ;
*          Hispanic black, Mexican-American)                     ;
*       self-rating of health status (1=excellent, 2=very good,  ;
*          3=good, 4=fair, 5=poor)                               ;
*       compare own activity level to others (1=more active,     ;
*          2=less active, 3=about the same)                      ;
*       poverty status (1=at or below poverty line, 2=above)     ;
*                                                                ;
*   Variance estimation by replicate (BRR) method with Fay       ;
*   adjustment.                                                  ;
*****************************************************************;

*****************************************************************;
*   Specify the paths for NHANES III public-use ASCII data       ;
*   on CD-ROM, EXAM and ADULT files. You will need to modify     ;
*   these paths if your CD-ROM is not drive E:                   ;
*****************************************************************;
FILENAME EXAM "E:\EXAM\EXAM.DAT" LRECL=6235;
FILENAME ADULT "E:\ADULT\ADULT.DAT" LRECL=3348;

*****************************************************************;
*   Read in select variables from the NHANES III public-use      ;
*   file EXAM.DAT                                                ;
*****************************************************************;
DATA EXAMVARS;
   INFILE EXAM MISSOVER;
   INPUT
      SEQN     1-5
      DMARETHN 12
      HSSEX    15
      HSAGEIR  16-17
      HSAGEU   18
      BMPBMI   1524-1527;
   LABEL
      SEQN     = "Respondent identification number"
      DMARETHN = "Race-ethnicity"
      HSSEX    = "Sex"
      HSAGEIR  = "Age at interview (Screener)"
      HSAGEU   = "Age at interview-unit (Screener)"
      BMPBMI   = "Body mass index";
   RUN;

*****************************************************************;
*   Read in select variables from the NHANES III public-use      ;
*   file ADULT.DAT                                               ;
*****************************************************************;
DATA ADLTVARS;
   INFILE ADULT MISSOVER;
```

Figure 6: Program in SAS and SAS-callable SUDAAN for conventional logistic regression analysis, with standard errors calculated by replicate method

```
INPUT
   SEQN     1-5
   DMARETHN 12
   HSSEX    15
   HSAGEIR  18-19
   HSAGEU   20
   DMPPIR   36-41
   WTPFEX6  61-69
   WTPXRP1  763-771
   WTPXRP2  772-780
   WTPXRP3  781-789
   WTPXRP4  790-798
   WTPXRP5  799-807
   WTPXRP6  808-816
   WTPXRP7  817-825
   WTPXRP8  826-834
   WTPXRP9  835-843
   WTPXRP10 844-852
   WTPXRP11 853-861
   WTPXRP12 862-870
   WTPXRP13 871-879
   WTPXRP14 880-888
   WTPXRP15 889-897
   WTPXRP16 898-906
   WTPXRP17 907-915
   WTPXRP18 916-924
   WTPXRP19 925-933
   WTPXRP20 934-942
   WTPXRP21 943-951
   WTPXRP22 952-960
   WTPXRP23 961-969
   WTPXRP24 970-978
   WTPXRP25 979-987
   WTPXRP26 988-996
   WTPXRP27 997-1005
   WTPXRP28 1006-1014
   WTPXRP29 1015-1023
   WTPXRP30 1024-1032
   WTPXRP31 1033-1041
   WTPXRP32 1042-1050
   WTPXRP33 1051-1059
   WTPXRP34 1060-1068
   WTPXRP35 1069-1077
   WTPXRP36 1078-1086
   WTPXRP37 1087-1095
   WTPXRP38 1096-1104
   WTPXRP39 1105-1113
   WTPXRP40 1114-1122
   WTPXRP41 1123-1131
   WTPXRP42 1132-1140
   WTPXRP43 1141-1149
   WTPXRP44 1150-1158
   WTPXRP45 1159-1167
   WTPXRP46 1168-1176
   WTPXRP47 1177-1185
   WTPXRP48 1186-1194
   WTPXRP49 1195-1203
   WTPXRP50 1204-1212
   WTPXRP51 1213-1221
   WTPXRP52 1222-1230
```

Figure 6 (continued)

```
   HAB1     1451
   HAT28    2499;
LABEL
   SEQN     = "Respondent identification number"
   DMARETHN = "Race-ethnicity"
   HSSEX    = "Sex"
   HSAGEIR  = "Age at interview (Screener)"
   HSAGEU   = "Age at interview-unit (Screener)"
   DMPPIR   = "Poverty Income Ratio (unimputed income)"
   WTPFEX6  = "Total MEC-examined sample final weight"
   WTPXRP1  = "Replicate 1 final exam weight"
   WTPXRP2  = "Replicate 2 final exam weight"
   WTPXRP3  = "Replicate 3 final exam weight"
   WTPXRP4  = "Replicate 4 final exam weight"
   WTPXRP5  = "Replicate 5 final exam weight"
   WTPXRP6  = "Replicate 6 final exam weight"
   WTPXRP7  = "Replicate 7 final exam weight"
   WTPXRP8  = "Replicate 8 final exam weight"
   WTPXRP9  = "Replicate 9 final exam weight"
   WTPXRP10 = "Replicate 10 final exam weight"
   WTPXRP11 = "Replicate 11 final exam weight"
   WTPXRP12 = "Replicate 12 final exam weight"
   WTPXRP13 = "Replicate 13 final exam weight"
   WTPXRP14 = "Replicate 14 final exam weight"
   WTPXRP15 = "Replicate 15 final exam weight"
   WTPXRP16 = "Replicate 16 final exam weight"
   WTPXRP17 = "Replicate 17 final exam weight"
   WTPXRP18 = "Replicate 18 final exam weight"
   WTPXRP19 = "Replicate 19 final exam weight"
   WTPXRP20 = "Replicate 20 final exam weight"
   WTPXRP21 = "Replicate 21 final exam weight"
   WTPXRP22 = "Replicate 22 final exam weight"
   WTPXRP23 = "Replicate 23 final exam weight"
   WTPXRP24 = "Replicate 24 final exam weight"
   WTPXRP25 = "Replicate 25 final exam weight"
   WTPXRP26 = "Replicate 26 final exam weight"
   WTPXRP27 = "Replicate 27 final exam weight"
   WTPXRP28 = "Replicate 28 final exam weight"
   WTPXRP29 = "Replicate 29 final exam weight"
   WTPXRP30 = "Replicate 30 final exam weight"
   WTPXRP31 = "Replicate 31 final exam weight"
   WTPXRP32 = "Replicate 32 final exam weight"
   WTPXRP33 = "Replicate 33 final exam weight"
   WTPXRP34 = "Replicate 34 final exam weight"
   WTPXRP35 = "Replicate 35 final exam weight"
   WTPXRP36 = "Replicate 36 final exam weight"
   WTPXRP37 = "Replicate 37 final exam weight"
   WTPXRP38 = "Replicate 38 final exam weight"
   WTPXRP39 = "Replicate 39 final exam weight"
   WTPXRP40 = "Replicate 40 final exam weight"
   WTPXRP41 = "Replicate 41 final exam weight"
   WTPXRP42 = "Replicate 42 final exam weight"
   WTPXRP43 = "Replicate 43 final exam weight"
   WTPXRP44 = "Replicate 44 final exam weight"
   WTPXRP45 = "Replicate 45 final exam weight"
   WTPXRP46 = "Replicate 46 final exam weight"
   WTPXRP47 = "Replicate 47 final exam weight"
   WTPXRP48 = "Replicate 48 final exam weight"
   WTPXRP49 = "Replicate 49 final exam weight"
   WTPXRP50 = "Replicate 50 final exam weight"
```

Figure 6 (continued)

```
      WTPXRP51 = "Replicate 51 final exam weight"
      WTPXRP52 = "Replicate 52 final exam weight"
      HAB1     = "Is health in general excellent,...,poor"
      HAT28    = "Active compared with men/women your age";
   RUN;

*****************************************************************;
*   Merge the two files by SEQN                                 ;
*****************************************************************;
PROC SORT DATA=EXAMVARS;
   BY SEQN;
   RUN;

PROC SORT DATA=ADLTVARS;
   BY SEQN;
   RUN;

DATA BOTH;
   MERGE EXAMVARS ADLTVARS;
   BY SEQN;
   RUN;

*****************************************************************;
*   Prepare variables for use by SUDAAN's PROC DESCRIPT         ;
*****************************************************************;
DATA FORSUDN;
   SET BOTH;
   ************************************;
   * categorize age                   ;
   ************************************;
   AGE = HSAGEIR;
   IF HSAGEU = 1 THEN AGE = AGE / 12;
   IF AGE GE 20 AND AGE LE 39 THEN AGEGRP = 1;
   ELSE IF AGE GE 40 AND AGE LE 59 THEN AGEGRP = 2;
   ELSE IF AGE GE 60 THEN AGEGRP = 3;
   ELSE AGEGRP = 0;
   ************************************;
   * race-ethnicity classification    ;
   ************************************;
   RACEETHN = DMARETHN;
   IF DMARETHN = 4 THEN RACEETHN = 1;
   ************************************;
   * recode 8 or 9-fills as missing   ;
   ************************************;
   IF BMPBMI   =   8888 THEN BMPBMI   = .;
   IF DMPPIR   = 888888 THEN DMPPIR   = .;
   IF HAB1     =      8 THEN HAB1     = .;
   IF HAB1     =      9 THEN HAB1     = .;
   IF HAT28    =      8 THEN HAT28    = .;
   IF HAT28    =      9 THEN HAT28    = .;
   ************************************;
   * calculate overweight indicator   ;
   ************************************;
   OVERWT = 0;
   IF HSSEX = 1 AND BMPBMI GE 27.8 THEN OVERWT = 100;
   IF HSSEX = 2 AND BMPBMI GE 27.3 THEN OVERWT = 100;
   ************************************;
   * poverty status classification    ;
   ************************************;
   IF DMPPIR LE 1.0 THEN POVERTY = 1;
```

Figure 6 (continued)

```
   ELSE IF DMPPIR GT 1.0 THEN POVERTY = 2;
   ************************************;
   * output select variables for      ;
   * examined adults                  ;
   ************************************;
   KEEP WTPFEX6 WTPXRP1--WTPXRP52 AGE AGEGRP HSSEX RACEETHN
      HAB1 HAT28 POVERTY OVERWT;
   IF AGE GE 20 AND WTPFEX6 GT 0 THEN OUTPUT;
   RUN;

*******************************************************;
* Because this is the SAS-callable version, SUDAAN     ;
* PROC LOGISTIC is invoked as PROC RLOGIST             ;
*******************************************************;
PROC RLOGIST SUDDATA=FORSUDN FILETYPE=SAS DESIGN=BRR;
   WEIGHT WTPFEX6;
   REPWGT WTPXRP1--WTPXRP52 / ADJFAY=2.0408;
   SUBGROUP HSSEX AGEGRP RACEETHN HAB1 HAT28 POVERTY;
   LEVELS 2 3 3 5 3 2;
   REFLEVEL HSSEX=1 AGEGRP=1 RACEETHN=1 HAB1=1 HAT28=3 POVERTY=2;
   MODEL OVERWT = HSSEX AGEGRP RACEETHN HAB1 HAT28 POVERTY;
   OUTPUT / BETAS=DEFAULT FILENAME=SUDNOUT FILETYPE=SAS REPLACE;
   RUN;

************************************;
* read in SUDNOUT, arrange results    ;
************************************;
DATA RESULTS;
   SET SUDNOUT;
   EST = BETA;
   SE = SEBETA;
   ************************************;
   * labels for the coefficients      ;
   ************************************;
   LENGTH QTYLABEL $25.;
   IF      MODELRHS =  1 THEN QTYLABEL = 'Intercept';
   ELSE IF MODELRHS =  2 THEN QTYLABEL = 'Male';
   ELSE IF MODELRHS =  3 THEN QTYLABEL = 'Female';
   ELSE IF MODELRHS =  4 THEN QTYLABEL = 'Age 20-39';
   ELSE IF MODELRHS =  5 THEN QTYLABEL = 'Age 40-59';
   ELSE IF MODELRHS =  6 THEN QTYLABEL = 'Age 60+';
   ELSE IF MODELRHS =  7 THEN QTYLABEL = 'Non-Hispanic white/other';
   ELSE IF MODELRHS =  8 THEN QTYLABEL = 'Non-Hispanic black';
   ELSE IF MODELRHS =  9 THEN QTYLABEL = 'Mexican-American';
   ELSE IF MODELRHS = 10 THEN QTYLABEL = 'Health excellent';
   ELSE IF MODELRHS = 11 THEN QTYLABEL = 'Health very good';
   ELSE IF MODELRHS = 12 THEN QTYLABEL = 'Health good';
   ELSE IF MODELRHS = 13 THEN QTYLABEL = 'Health fair';
   ELSE IF MODELRHS = 14 THEN QTYLABEL = 'Health poor';
   ELSE IF MODELRHS = 15 THEN QTYLABEL = 'More active than others';
   ELSE IF MODELRHS = 16 THEN QTYLABEL = 'Less active than others';
   ELSE IF MODELRHS = 17 THEN QTYLABEL = 'About the same';
   ELSE IF MODELRHS = 18 THEN QTYLABEL = 'At or below poverty line';
   ELSE IF MODELRHS = 19 THEN QTYLABEL = 'Above poverty line';
   ************************************;
   * calculate t-ratios and p-values  ;
   * using t-distribution with df = 52 ;
   ************************************;
   DF = 52;
   TRATIO = EST/SE;
```

Figure 6 (continued)

```
   PVALUE = 2 * (1 -PROBT( ABS(TRATIO),DF ) );
   FORMAT EST SE 8.4 TRATIO 6.2 PVALUE 6.4 DF 4.1;
   RUN;

*****************************************************************;
*   print results                                                ;
*****************************************************************;
OPTIONS LINESIZE=132;
PROC PRINT DATA=RESULTS;
   VAR QTYLABEL EST SE DF TRATIO PVALUE;
   RUN;
```

Figure 6 (continued)

Table 6: Results from Conventional Analysis Example D—estimated coefficients, standard errors, degrees of freedom, T-ratios, and p-values

	Est.	SE	df	T-ratio	p-value
Intercept	−1.5090	0.0834	52.0	−18.10	0.0000
Male	0.0000	0.0000	52.0	—	—
Female	0.0477	0.0503	52.0	0.95	0.3475
Age 20–39	0.0000	0.0000	52.0	—	—
Age 40–59	0.6567	0.0538	52.0	12.20	0.0000
Age 60+	0.5944	0.0759	52.0	7.83	0.0000
Non-Hispanic white/other	0.0000	0.0000	52.0	—	—
Non-Hispanic black	0.4368	0.0591	52.0	7.39	0.0000
Mexican-American	0.3914	0.0684	52.0	5.72	0.0000
Health excellent	0.0000	0.0000	52.0	—	—
Health very good	0.4407	0.0766	52.0	5.76	0.0000
Health good	0.6870	0.0872	52.0	7.88	0.0000
Health fair	0.8295	0.0899	52.0	9.23	0.0000
Health poor	0.4744	0.1171	52.0	4.05	0.0002
More active than others	−0.3805	0.0631	52.0	−6.03	0.0000
Less active than others	0.2764	0.0571	52.0	4.84	0.0000
About the same	0.0000	0.0000	52.0	—	—
At or below poverty line	−0.0574	0.0550	52.0	−1.04	0.3017
Above poverty line	0.0000	0.0000	52.0	—	—

5 Discussion

Analyzing data from a complex survey such as NHANES III can be a complicated task even apart from missing data. The NHAMES III Multiply Imputed Data Set was designed to produce population estimates and standard errors with better statistical properties than those coming from ad hoc case-deletion or single-imputation techniques traditionally used by analysts. Initially, analyzing a multiply imputed data set may require slightly more effort than traditional methods, because estimates and standard errors must be computed several times and then combined by Rubin's rules. In the long run, however, using multiply imputed data sets may prove to be simpler and more convenient, because many subjective decisions formerly made by analysts (e.g. which subset of cases to use for a particular analysis) have been eliminated.

For the most straightforward types of analyses of NHANES III, such as creating national estimates of means and prevalences, results obtained from the NHANES III Multiply Imputed Data Set may appear to be similar to those obtained by traditional methods. As discussed in the previous section, however, the two approaches should not be regarded as equivalent because they do perform differently over repeated application. When used repeatedly, the statistical benefits of the new method (greater precision, reduced probability of Type I error) will begin to accrue. For regression modeling and more complicated analyses involving many variables at once, the new method may produce results that are substantially different from those of traditional methods. These larger discrepancies arise because as the number of variables grows, the proportion of cases discarded by traditional methods tends to grow rapidly, whereas the multiple-imputation method is always based on the entire sample. In these situations, results from the NHANES III Multiply Imputed Data Set will tend to be less biased and more precise, because they are based on the full sample rather than a subset of complete cases.

References

Barnard, J. and Rubin, D.B. (1999) Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948–955.

Department of Health and Human Services (DHHS) (1994) *Plan and Operation of the Third National Health and Nutrition Examination Survey*, 1988–94. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Department of Health and Human Services (DHHS) (1996) *NHANES III Reference Manuals and Reports*. CD-ROM. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Department of Health and Human Services (DHHS) (1997) *National Health and Nutrition Examination Survey, III*, 1988–1994. CD-ROM, Series 11, No. 1A, ASCII Version. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Department of Health and Human Services (DHHS) (1998) *National Health and Nutrition Examination Survey, III*, 1988–1994. CD-ROM, Series 11, No. 2A, ASCII Version. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Judkins, D.R. (1990) Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223–239.

Little, R.J.A. (1986) Survey nonresponse adjustments for estimation of means. *International Statistical Review*, 54, 139–157.

Little, R.J.A., Ezzati-Rice, T.M., Johnson, W., Khare, M., Rubin, D.B. and Schafer, J.L. (1995) A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. *Proceedings of the Annual Research Conference*, 257–266. Washington, DC: Department of Commerce, Bureau of the Census. Included with the NHANES III Multiple Imputation Research Data Set (DHHS, 1999, CD-ROM).

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data,* New York: Wiley.

Little, R.J.A. and Rubin, D.B. (1992) Assessment of Trial Imputations for NHANES III. Waban, MA: Datametrics Research, Inc. Technical report accompanying the NHANES III Multiple Imputation Research Data Set (DHHS, 2000, CD-ROM).

Meng, X.L. (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 10, 538–573.

Research Triangle Institute (1998) *SUDAAN: Software for the Statistical Analysis of Correlated Data*, Version 7. Research Triangle Park, NC: Research Triangle Institute.

Rubin, D.B. and Schenker, N. (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366–374.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys,* New York: Wiley.

Rubin, D.B. (1996) Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473–489.

Stata Corporation (1997) *Stata Reference Manual.* College Station, TX: Stata Corporation.

Westat, Inc. (1997) *A User's Guide to WesVarPC.* Rockville, MD: Westat, Inc.

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