****
****

NH3MI.DOC

TO VIEW OR PRINT THIS TEXT DOCUMENT:
IMPORT IT INTO A WORD PROCESSOR, SET YOUR MARGINS
TO ZERO, AND USE A FIXED-WIDTH FONT (e.g. COURIER)

YOU MAY ALSO VIEW OR PRINT THIS DOCUMENT USING THE
ADOBE ACROBAT READER, VERSION 4 OR LATER. THE ACROBAT-
READABLE VERSION OF THIS FILE IS CALLED

NH3MI.PDF

****
****

Third National Health and Nutrition Examination Survey
(NHANES III, 1988-1994):
Multiply Imputed Data Set on CD-ROM (Series 11, No. 7A)

NHANES III MULTIPLY IMPUTED DATA SET USER'S GUIDE

June 2001

Table of Contents

OVERVIEW OF NHANES III

The Third National Health and Nutrition Examination Survey (NHANES III) provides basic and detailed information on health and nutritional status of the civilian, noninstitutionalized U.S. population aged 2 months and older. It is the seventh in a series of similar surveys conducted periodically by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC). Data from NHANES III for public use were released in 1996 in the form of five data sets of files:

NHANES III Household Adult Data File (Catalog Number 77560)

NHANES III Household Youth Data File (Catalog Number 77550)

NHANES III Examination Data File (Catalog Number 76200)

NHANES III Laboratory Data File (Catalog Number 76300)

NHANES III Dietary Recall Data Files (Catalog Number 76700)

Public-use data files for the third National Health and Nutrition Examination Survey are also available from the National Technical Information Service (NTIS). A list of NCHS public-use data tapes available for purchase from NTIS may be obtained from the Data Dissemination Branch at NCHS. Information regarding a bibliography (on disk) of journal articles citing data from all the NHANES and the availability of NHANES III data in CD-ROM/SETS software format can be obtained from the Data Dissemination Branch (301-458-4636) or by writing to:

Data Dissemination Branch
National Center for Health Statistics
Room 1018
6525 Belcrest Road
Hyattsville, Maryland 20782-2003

NTIS can be contacted at:

NTIS - Computer Products Office
5285 Port Royal Road
Springfield, Virginia 22161
(703) 487-4807

Copies of all NHANES III questionnaires and data collection forms are included in the Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988-1994 (NCHS, 1994; U.S. DHHS, 1996). This publication, along with detailed information on NHANES procedures, interviewing, data collection, quality control techniques, survey design, nonresponse, and sample weighting can be found on the NHANES III Reference Manuals and Reports CD-ROM (U.S. DHHS, 1996). Information on how to order this CD-ROM is available from the Data Dissemination Branch at NCHS at the address and telephone number given above.

Referencing or Citing NHANES III Data

- In publications, please acknowledge NCHS as the original data source. For instance, the reference for the NHANES III Multiply Imputed Data Set is:

  U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set. CD-ROM, Series 11, No. 7A. Hyattsville, MD: Centers for Disease Control and Prevention, 2001. Includes access software: Adobe Systems, Inc. Acrobat Reader version 4.

- Please place the acronym "NHANES III" in the titles or abstracts of journal articles and other publications in order to facilitate the retrieval of such materials in bibliographic searches.

OVERVIEW OF THE NHANES III MULTIPLY IMPUTED DATA SET

Multiple imputation is a statistical technique in which missing data are replaced by several sets of plausible, alternative simulated values. The multiple imputations distributed on this CD-ROM provide an improved method for handling missing values in many but not all analyses of NHANES III data. These files are intended as a companion to--not a replacement for--other NHANES III public-use data sets. National estimates from the NHANES III Multiply Imputed Data Set may differ from those obtained from other public-use data sets (DHHS, 1996, CD-ROM, Series 1A or 2A). Users of the NHANES III Multiply Imputed Data Set are advised to consult the detailed analytic guidelines in this document and other documents provided on this CD-ROM.

MISSING DATA IN NHANES III

NHANES III is a complex, multistage area sample 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. The total NHANES III sample size is 39,695 individuals.

As with most large scale data collection efforts, NHANES III experienced moderate amounts of missing data due to various types of nonresponse. The data collection in NHANES III occurred in three major stages: (a) a household screener interview to obtain the age, sex, race, and ethnicity of each member of each sampled household, which determined eligibility and selection probabilities for the final stage of sampling; (b) a personal home interview to gather data on the family and each sampled person (SP), with questions pertaining to health status, medical history, diet, and other health-related behaviors; and (c) physical examination of each SP by medical staff in a mobile examination center (MEC). In the screener stage, every effort was made to obtain full and correct information, either from household members or proxy respondents (e.g. from neighbors). As a result, basic demographic information including age, sex, and race/ethnicity is available for each of the 39,695 SPs. At the personal home interview stage, however, data were obtained for only eighty-six percent of the entire sample (33,994 persons). All interviewed persons were invited to the MEC for a medical examination. Only seventy-eight percent (30,818) of the SPs were successfully examined in the MEC, however, and an additional 493 persons were given a special, limited examination in their homes. This basic pattern of nonresponse is depicted in Figure 1 below.

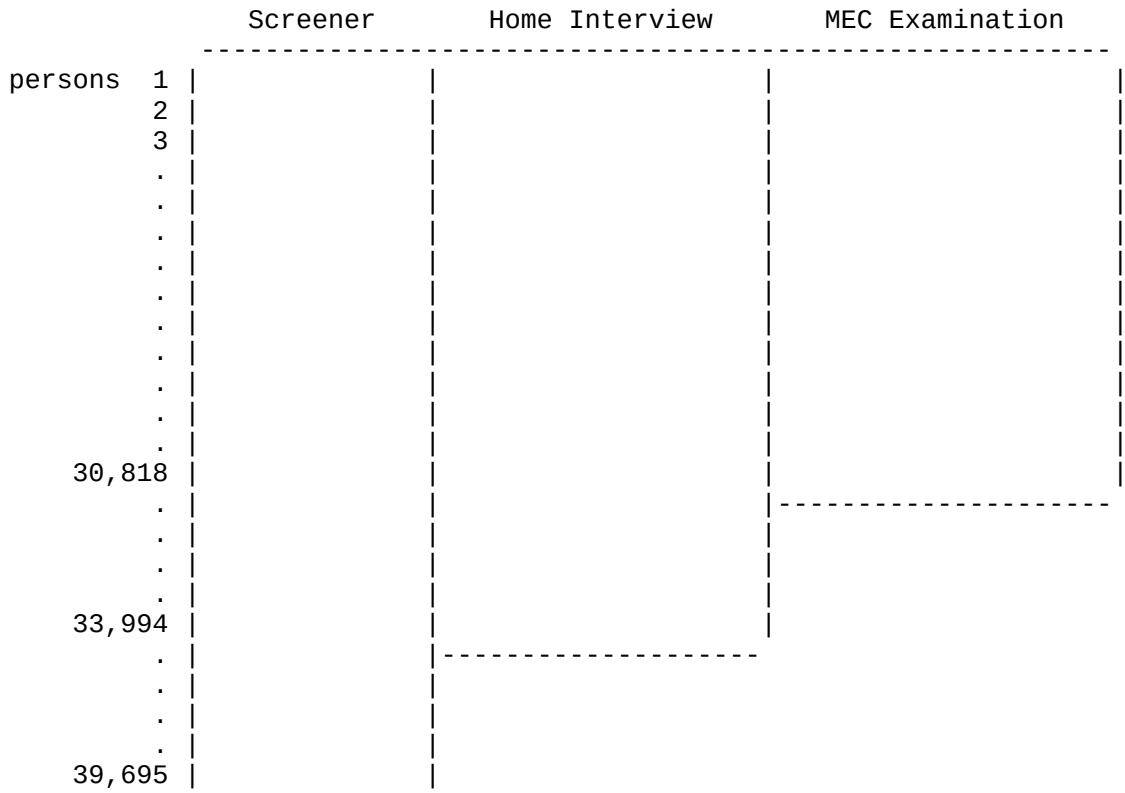

FIGURE 1: Patterns of nonresponse in NHANES III created by failure to interview and failure to examine sampled persons.

The rates of nonresponse varied appreciably across subgroups defined by the SPs' demographic characteristics. For illustration, the variation in response rates by age, race/ethnicity, and household size is shown in Table 1 below. Rates of nonresponse increase dramatically with age. Nonresponse rates were somewhat lower among African-Americans and Mexican-Americans than persons of other race/ethnicity. Nonresponse rates were highest among SPs in small households, and decrease dramatically as household size goes up. Unless special corrective measures are taken in the analysis of data from NHANES III, results may be biased toward the characteristics of those groups with highest rates of response.

TABLE 1: NHANES III response patterns by age, race/ethnicity and household size (NIN = not interviewed, INM = interviewed but not MEC examined, MEC = MEC examined)

	NIN %	INM %	MEC %
Overall (N=39,695)	14.4	8.0	77.6
Age			
Under 5	5.5	5.9	88.6
5-16	8.8	5.2	86.0
17-39	15.8	6.2	78.0
40-59	20.3	6.8	72.8
60+	21.2	15.5	63.3
Race/ethnicity			
Non-Hispanic Black	13.0	5.6	81.4
Mexican-American	12.2	5.9	81.8
Other	16.6	10.8	72.6
Household size			
1-2	20.8	12.9	66.3
3-4	13.9	7.2	78.9
5+	9.0	4.6	86.4

In NHANES III, as in previous NCHS examination surveys, nonresponse due to failure to interview and failure to examine has been handled by classical methods of reweighting. Because of the complex survey design, methods of statistical analysis that are based upon an assumption of simple random sampling are not appropriate. Sample weights are needed to correctly estimate prevalence, means, medians, and other population statistics, because individuals were not selected into the NHANES III sample with equal probability. To further compensate for differential rates of nonresponse, these sample weights were adjusted so that the weighted sample of respondents mirrored the demographic characteristics of the U.S. population. This adjustment was performed in two stages. In the first stage, the non-interview cases were removed from the sample, and their sample weights were distributed among respondents with similar demographic characteristics. The resulting adjusted weight, known as the "final interview weight" (variable WTPFQX6 in NHANES III public-use data files), is used for statistical analyses involving items collected in the home interview. In the second stage of adjustment, persons who were interviewed but not MEC-examined were assigned weights of zero, and their former weights were distributed among examined persons with similar characteristics. The resulting adjusted weight, known as the "final examination weight" (variable WTPFEX6), is appropriate for analyses involving items from the examination, and for analyses involving items from both the home interview and the examination. Detailed descriptions of the sample design and weighting procedures, and further guidelines for analysis, are available from The Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988-1994, (NCHS, 1994; U.S. DHHS, 1996), and from the NHANES III Reference Manuals and Reports CD-ROM (U.S. DHHS, 1996).

In addition to the nonresponse created by failure to interview and failure to examine, NHANES III also has moderate amounts of missing data on specific interview and examination items due to "Don't know" responses, refusals to answer specific questions or to submit to certain examination procedures, examinations that had to be terminated because SPs had to leave early, etc. For the most part, no statistical procedures or adjustments were applied to this type of item nonresponse. As a result, users of other NHANES III public-use data files will find that many interview and examination variables include codes such as "Blank but applicable," "Don't know," and other instances denoting failure to obtain usable data values.

HISTORY OF THE NHANES III MULTIPLE IMPUTATION RESEARCH PROJECT

In 1992, NCHS assembled a team of statistical researchers to investigate a variety of alternatives to current missing-data practice in NHANES III. Topics for investigation included improved methods of reweighting, procedures for single imputation (e.g. hot-deck methods), and model-based multiple imputation. This research effort evolved into the NHANES III Multiple Imputation Research Project. The history and efforts of the project team are summarized below.

"Imputation" is a generic term that encompasses many different methods for filling in the missing values in a data set, i.e. replacing the missing values with plausible data. For decades, imputation (often in conjunction with reweighting) has been applied in many large-scale data collection efforts at the U.S. Bureau of the Census, Bureau of Labor Statistics, and other public and private organizations. For an overview and critique of many commonly used survey imputation procedures, see Little and Rubin (1987, ch. 4) and their references.

From both practical and statistical viewpoints, imputation procedures are attractive. Imputation produces a "clean" data file which is easy for users to analyze. Users may not have the experience, time, or resources needed to perform state-of-the-art analyses of an incomplete data set. By releasing data files with imputed values for missing items, agencies like NCHS may provide a valuable service to data users, helping them to solve difficult missing-data problems that arise frequently in practice. Releasing imputed files helps to ensure that a variety of users performing similar statistical analyses will be led to the same results; variation due to different treatments of missing values by the user is removed. Finally, imputation can be more effective than reweighting in using the statistical relationships among survey variables to produce accurate predictions of the missing data values, leading to more efficient estimates (Little, 1986).

On the other hand, imputation can be difficult to implement well, particularly in a multivariate data set like NHANES III. A good imputation method should make use of whatever relevant information is known for an SP to impute his or her missing data values. The imputation method should attempt to preserve, rather than distort, the distribution of each variable and the plausible statistical relationships among them. Finally, unless special corrective measures are taken, an imputed data set may provide misleading results by understating actual levels of statistical uncertainty. If a data set has been "filled in" by imputation, and the imputed values are treated no differently from the observed values, then measures of uncertainty (e.g. standard errors) will have a downward bias because they ignore the fact the imputed values are only predictions and contain substantial error. This shortcoming of traditional single-imputation methods is well known and has been documented by Little and Rubin (1987), Rubin (1987), and many others.

In response to this shortcoming of single imputation, a new statistical paradigm called multiple imputation has recently emerged. Multiple imputation (MI) is a simulation-based approach to missing data in which each missing datum 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. Each version may be analyzed by standard complete-data methods, and the variation in results among the M versions provides a measure of missing-data uncertainty in addition to the usual variation due to sampling. The M sets of results may be formally combined, using simple rules of arithmetic, to provide standard errors and confidence intervals that incorporate missing-data uncertainty. The MI

method has been demonstrated to produce accurate confidence intervals in a variety of settings. Detailed presentations of the MI paradigm are given by Rubin (1987; 1996) and Schafer (1997).

In 1992 and 1993, the NHANES III MI Research Project team successfully produced multiple imputations for preliminary data from Phase 1. Using the 12,392 Phase 1 sampled adults (age 20+) and state-of-the-art computational methods, they formulated a statistical model and generated M = 10 multiple imputations for a modest selection of variables from the MEC examination and home interview. The modeling and imputation procedures are described in detail by Schafer, Khare, and Ezzati-Rice (1993), and by Khare, Little, Rubin and Schafer (1993). Results from this preliminary study were very encouraging. This study demonstrated the computational feasibility of using model-based MI in a large survey such as NHANES III. The quality of the imputations appeared to be comparable to or better than those produced by more traditional hot-deck procedures, which were also investigated by Ezzati-Rice, Fahimi, Judkins and Khare (1993). In particular, the joint modeling of variables used in the MI method was quite effective at preserving important inter-variable relationships. A detailed comparison of single and multiple imputations is given by Ezzati-Rice, Khare, Rubin, Little and Schafer (1993).

Through this preliminary study, the Project team concluded that MI offered significant advantages over reweighting in adjusting for nonresponse at the MEC examination stage; that is, substantial gains in precision could result by imputing examination variables for SPs who were interviewed but not examined, rather than removing the non-examined persons from the sample and reweighting the remaining examined SPs. MI also appeared to be an attractive method for handling the scattered missing values on interview and examination items resulting from refusals, "Don't know" responses, etc. However, MI appeared to offer little advantage over reweighting for the SPs who were not interviewed. The Project team recommended a combined strategy of MI for missing data among interviewed subjects and reweighting to adjust for non-interviews. Finally, the team concluded that for most analytic purposes, M = 5 imputations would be sufficient to produce high-quality statistical estimates. For discussion and recommendations regarding this prelimary study, see the report of Little and Rubin (1992).

In 1994 and 1995, the Project team designed and implemented an extensive simulation study to evaluate the performance of the MI method over repeated samples in an NHANES-type examination survey. The study was designed to answer questions such as: Would confidence intervals of a particular stated coverage level--say 95%--obtained from the MI method actually cover their respective population quantities 95% of the time? To address such questions, an artificial but realistic pseudo-population was created by pooling data from 31,847 individuals drawn from NHANES III and three previous health examination surveys: NHANES I (1971-74), NHANES II (1976-80), and HHANES (1982-84, Hispanic Americans only). This pseudo-population was weighted to resemble the projected year 2000 U.S. population with respect to important demographic and geographic characteristics. Probability samples of approximately 6,000 were drawn from this population according to a sampling plan with strata similar to those of NHANES III (the clustered aspects of the population were not simulated). Missing values were imposed upon each sample by a random mechanism that mimics the NHANES III nonresponse rates and patterns. Five imputations were generated by a model-based method, and estimates and confidence intervals were calculated for a large number of population quantities including means, prevalences, medians, quantiles, and regression coefficients. The entire process of sampling, imputation, and estimation was repeated 1,000 times. The MI procedure performed quite well. For example, among 448 population means, the average simulated coverage of 95% MI intervals over 1,000 repetitions was 949.3 (not significantly different from 950). The coverage showed no overall tendency to increase or decrease as the rates of missing information varied from 0% to nearly 60%. Detailed description of this simulation study and further discussion of its results and limitations is provided by Little et al. (1995).

Based on the encouraging results of the preliminary imputation work and simulation study, the Project team proceeded to develop and implement a workable MI procedure for NHANES III data as they became available in 1996 and 1997. In consultation with NCHS staff, a selection of approximately 60 key variables from NHANES III was assigned high priority for imputation. This set included selected body measurements (weight, height, waist circumference, skin folds, etc.), key variables from bone densitometry, fundus photography, blood pressure, and laboratory results from the analysis of blood and urine specimens (lead, iron, hemoglobin, hematocrit, cholesterol, triglycerides, etc.). The MI procedure was designed to preserve the essential relationships among these examination variables, and their relationships to some key variables from the home interview: health status, physical activity status, tobacco and alcohol use, self-reported height and weight, home blood pressure readings, and presence of select medical conditions (e.g. "Have you ever been told by a doctor or other health professional that your blood cholesterol level was high?"). The MI procedure also took into account basic demographic and economic characteristics of SPs and important features of the complex sample design. Five imputations were created for all 33,994 interviewed persons. This effort culminated with the production of the NHANES III Multiply Imputed Data Set, as described in this document and the accompanying reports on this CD-ROM.

IMPUTATION MODELS AND PROCEDURES

One striking feature of NHANES III is that the variables collected in the survey vary greatly by age. For example, bone densitometry was performed on adults of age 20 and over, whereas fundus photography applied to those 40 and over. Hemoglobin, hematocrit, and lead were measured for those 1 and over, cholesterol and triglycerides for those 4 and over, and selenium for those 12 and over. In the home interview, entirely different questionnaires were used for adults (17 and over) and youth (16 and under). To adequately describe the entire NHANES III sample by a single statistical model would have been difficult, because the data contain no information on the distribution of variables within the age groups for which they were not collected. To simplify the modeling task, the data were initially split into nine different age classes. A separate imputation model was applied to each class, and after imputation the classes were merged back into a single data set. The nine age classes and the number of interviewed SPs in each class are listed below.

TABLE 2. Age classes for imputation models in the NHANES III Multiply Imputed Data Set

Age class	Interviewed persons
1. Newborn (under 1 year)	2,107
2. 1 year old	1,339
3. 2-3 years old	2,536
4. 4-7 years old	3,426
5. 8-16 years old	4,536
6. 17-19 years old	1,225
7. 20-39 years old	7,377
8. 40-59 years old	4,852
9. 60+ years old	6,596
Total	33,994

Within each age class, a multivariate statistical model was constructed to simultaneously preserve the natural relationships among examination variables and their relationships to key variables from the home interview. Each model also preserves the intercorrelations among SPs within each NHANES III primary sampling unit. This latter feature is important because statistical methods recommended for the analysis of NHANES III--methods appropriate for data from complex surveys--rely heavily upon variation across primary sampling units to calculate standard errors. The models made use of some detailed demographic and geographic identifiers that cannot be released to the public in order to protect the confidentiality of SPs. For this reason, the model-fitting procedures cannot be duplicated by researchers outside of NCHS. After each model was determined, values for the missing variables of each SP were jointly simulated from a predictive probability distribution derived from the model, making use of whatever information was available from the SP's observed data. The simulation process was repeated five times, producing five multiple imputations for the missing data.

Detailed descriptions of the statistical models for the nine age classes are provided in the accompanying technical report by Schafer (2001a) found on this CD-ROM. Survey data like those from NHANES III can be quite complicated, and any statistical model used to describe them is at best only approximately true. The imputation models used here were designed to be compatible with many common analytic techniques including the estimation of prevalences, means, quantiles, linear and logistic regression modeling, etc. 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, their primary strengths and limitations.

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 to be imputed are (individually and jointly) normally distributed within demographic subgroups defined by age, sex, and race/ethnicity. 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 imputed 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. In some instances, power transformations that approximately removed skewness did not produce satisfactory results. For example, certain variables pertaining to skin folds, after being transformed to near-symmetry, still had lighter-than-normal tails; imputing them under a normal assumption would have generated unusually low and unusually high values outside the realm of physical plausibility. These problematic variables were transformed by a method based on the empirical cumulative distribution function (cdf), which forced 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. Regardless of which method was used--a power transformation, the empirical cdf method, or no transformation at all--the models may not 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 one's 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 imputed data will tend to be smaller than they otherwise would have been if no data had been missing. One notable exception, however, is that the models were designed to preserve potential two and three-way interactions among crucial demographic variables (gender and race/ethnicity). Moreover, because a separate imputation model was fit to each age class, interactions between age and the interview and examination variables are largely preserved as well.

Another limitation of the imputation procedure is that only a modest number of NHANES III variables could be included in the imputation models. The largest of the models, the one used for persons 60 and older, contained about 100 variables from the screener, 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 other variables drawn from NHANES III public-use data files; doing so could result in underestimation of the strength of these relationships.

Finally, as with any missing-data procedure, certain statistical assumptions were made about the manner in which the missing observations became missing. The imputation model and procedure assumed that nonresponse in NHANES III was "ignorable" in the sense defined by Little and Rubin (1987) and Rubin (1987). This assumption, which can neither be verified nor refuted from the NHANES III data, implies that the chance that any data value is missing may be related to quantities that are observed but not to quantities that are missing. Nearly all missing-data procedures that are commonly used in surveys, and in other areas of statistical practice, rely on an assumption of ignorability; often the assumptions made are even stronger. A detailed discussion of ignorability and its practical implications is given by Schafer (1997, ch. 2).

GUIDELINES FOR ANALYSIS

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 the previous section. 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 Analytic and Reporting Guidelines (U.S. DHHS, 1996) and in documentation accompanying previously released NHANES III public-use files (DHHS, 1996, CD-ROM, Series 1A or 2A). 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 and the choice of survey weights, as we now describe.

With other public-use files, users are instructed to use the final interview weight (WTPFQX6) for analyses involving items collected in the home interview, and the final examination weight (WTPFEX6) for analyses involving items from the MEC examination, or for joint analyses involving items from both the home interview and the MEC examination. The latter weight differs from the former in that it includes a nonresponse adjustment for SPs who were interviewed but not examined. In the NHANES III Multiply Imputed Data Set, however, examination variables for these SPs have been imputed, so the additional nonresponse adjustment becomes unnecessary. Therefore, users of this data set should use the final interview weight (WTPFQX6) for all estimation procedures. Variance estimation procedures based upon Taylor linearization should use the design information contained in the pseudo-PSU (SPPPSU6) and pseudo-stratum (SDPSTRA6) variables. Replication-type variance estimation procedures should use the fifty-two replicate versions of the final interview weight (WTPQRP1, WTPQRP2, ..., WTPQRP52). To minimize the possibilities for confusion, these are the only design variables and weights appearing in the NHANES III Multiply Imputed Data Set.

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; for more information, see the "Limitations" section of "IMPUTATION MODELS AND PROCEDURES" above.

Combining multiple estimates and standard errors

When producing statistical estimates from the NHANES III Multiply Imputed Data Set, the same estimation procedure should be applied to each of the five completed versions. The five sets of results may then be combined to produce a single statistical summary that formally incorporates uncertainty due to missing data into standard errors, significance levels, etc. Methods for combining the multiple sets of results are discussed by Rubin and Schenker (1986) and Rubin (1987). The easiest and most commonly used method is the one presented by Rubin (1987) for scalar (one-dimensional) estimands, which is summarized as follows.

Suppose that one is interested in producing an estimate and confidence interval for a population quantity Q, which may be a prevalence rate, mean, median, regression coefficient, etc. One must first calculate an estimate and standard error for Q from each of the M completed data sets, using methods that would be statistically appropriate if the data had no missing values. Because NHANES III has a complex sample design, one should use methods appropriate for complex samples, e.g. as implemented in SUDAAN or WesVar PC. Let Q1, Q2, ..., QM denote the estimates of Q, and let U1, U2, ..., UM denote the associated variance estimates (squared standard errors), from the repeated analysis of the imputed data. The combined estimate of Q is simply the mean or average of the individual estimates Q1, Q2, ..., QM. The combined standard error for this estimate comes from the following two sources.

- The within-imputation variance: This is the mean or average of the individual variance estimates U1, U2, ..., UM.

- The between-imputation variance: This is the sample variance (the sum of squared deviations from the mean, divided by M minus one) of the individual estimates Q1, Q2, ..., QM.

The total variance is the sum of the within-imputation variance and (M+1)/M times the between-imputation variance. The square root of this total variance is the overall standard error associated with the estimate of Q.

In many cases, an acceptable 95% confidence interval for Q can be formed based on a normal distribution: the overall point estimate, plus or minus two standard errors. A more accurate approximation derived by Rubin (1987) is based upon a Student's t-distribution, with degrees of freedom given by

$$D = (M - 1) * (1 + (M/(M+1))*(W/B))^{**}2 ,$$

where W and B are the within- and between-imputation variances, respectively. Under this t-approximation, a 95% confidence interval would be given by the estimate of Q, plus or minus the overall standard error times the 97.5th percentile of the t-distribution with D degrees of freedom.

Example analyses

The combination rule described above is easily carried out manually or programmed in standard statistical software packages (e.g. SAS). Several detailed numerical examples of this procedure are provided in a technical report found on this CD-ROM (Schafer, 2001b). These analyses were carried out in SUDAAN, WesVarPC, and SAS, and all computer programs are shown. Users are strongly encouraged to review that document before performing their own

analyses.

GENERAL REFERENCES

Department of Health and Human Services (DHHS) (1994) Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988-1994. 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.

Ezzati-Rice, T.M., Fahimi, M., Judkins, D. and Khare, M. (1993) Serial Imputation of NHANES III with Mixed Regression and Hot-Deck Techniques. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1:292-296. Alexandria, VA: American Statistical Association. Included with the Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set (DHHS, 2001, CD-ROM, Series 11, No. 7A).

Ezzati-Rice, T.M., Khare, M., Rubin, D.B., Little, R.J.A. and Schafer, J.L. (1993) A Comparison of Imputation Techniques in the Third National Health and Nutrition Examination Survey. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1:303-308. Alexandria, VA: American Statistical Association. Included with the Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set (DHHS, 2001, CD-ROM, Series 11, No. 7A).

Khare, M., Little, R.J.A., Rubin, B. and Schafer, J.L. (1993) Multiple Imputation of NHANES III. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1:297-302. Alexandria, VA: American Statistical Association. Included with the Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set (DHHS, 2001, CD-ROM, Series 11, No. 7A).

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 Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set (DHHS, 2001, CD-ROM, Series 11, No. 7A).

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

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

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 Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set (DHHS, 2001, CD-ROM, Series 11, No. 7A).

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. (1987) Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons.

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

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.

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schafer, J.L. (2001a) Multiple imputation models and procedures for NHANES III. Technical report accompanying the Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set (DHHS, 2001, CD-ROM).

Schafer, J.L. (2001b) Analyzing the NHANES III Multiply Imputed Data Set: Methods and Examples. Technical report accompanying the Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set (DHHS, 2001, CD-ROM, Series 11, No. 7A).

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. and Rubin, D.B. (1996) The NHANES III Multiple Imputation Project. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1:28-37. Alexandria, VA: American Statistical Association. Included with the Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set (DHHS, 2001, CD-ROM, Series 11, No. 7A).

Schafer J.L., Khare, M. and Ezzati-Rice, T.M. (1993) Multiple imputation of missing data in NHANES III. Proceedings of the Annual Research Conference, 459-487. Washington, DC: Department of Commerce, Bureau of the Census. Included with the Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set (DHHS, 2001, CD-ROM, Series 11, No. 7A).

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

GENERAL DATA FILE INFORMATION

The NHANES III Multiply Imputed Data Set was designed to facilitate repeated statistical analysis in current computing environments. It was originally proposed that the five imputations would be released in a single data file; each imputed variable would have appeared in the data set five times under slightly different names (e.g. BMPHT_1, BMPHT_2, ..., BMPHT_5). This idea was rejected because repeating an analysis five times might have become somewhat tedious for the user. For example, consider the task of writing a program to fit a multiple regression model. The user would have had to modify the name of each imputed variable in the program five times. Instead of placing the five imputations in a single data file, it was decided that they should be arranged in separate files with identical variable names. Under the latter arrangement, the user could run the same program five times, changing only the name of the input data file and possibly the name of the output data file that would store the results.

Data files

The NHANES III Multiply Imputed Data Set consists of six data files. Each file has 33,994 records, one for each interviewed person. The files are:

- CORE.DAT - This file contains 199 variables that do not change across the five imputation sets. The variables include general information from the screener (age, sex, race/ethnicity), design information and survey weights, imputation flags for the imputed variables, and additional NHANES III variables that may have appeared as predictors in the imputation models for one or more age classes but were not consistently imputed for all age classes.

- IMP1.DAT - This file contains the first version of 67 imputed variables, plus an identifying sequence number (SEQN) for merging with CORE.DAT.

- IMP2.DAT - Identical to IMP1.DAT, except that it contains the second version of the imputed variables.

- IMP3.DAT - Identical to IMP1.DAT, except that it contains the third version of the imputed variables.

- IMP4.DAT - Identical to IMP1.DAT, except that it contains the fourth version of the imputed variables.

- IMP5.DAT - Identical to IMP1.DAT, except that it contains the fifth version of the imputed variables.

The six files above are provided in ASCII format, along with SAS program files which read them in and convert them to SAS data sets. Another SAS program file, NH3MI.SAS, merges CORE.DAT with IMP1.DAT, ..., IMP5.DAT to create five versions of the merged SAS data set for analysis. Each of these merged data sets should be analyzed in the same fashion, producing five sets of estimates. These estimates should be saved and then combined using Rubin's (1987) procedure, as described above. Example analyses are provided in the accompanying technical report by Schafer (2001b), "Analyzing the NHANES III Multiply Imputed Data Set: Methods and Examples," included on this CD-ROM.

File layout

The appearance of IMP1.DAT, ..., IMP5.DAT is slightly different from that of other public-use files from NHANES III. In other public-use files, a blank field denotes a variable that does not apply to an individual because it was not collected for persons of that age group, a variable that was intentionally left blank because of a skip pattern in the questionnaire, etc. In the Multiply Imputed Data Set, these blank fields have been replaced by the numeric code -9 to facilitate use by software unaccustomed to blank fields. To users of SAS this new convention becomes irrelevant, because the program file NH3MI.SAS automatically changes every occurrence of -9 to the SAS missing-value code ".".

Documentation files

For each of the six .DAT data files listed above, a corresponding documentation file of the same name but with the .DOC extension provides complete information on file layout, tabulated distributions, and variable notes.

Variable names

Variable names in the NHANES III Multiply Imputed Data Set follow the same basic conventions as those in other public-use files. Names have no more than eight characters. The first two characters refer to the variable's origin; for example, variables beginning with "BD" come from the bone densitometry procedure in the MEC examination, and variables beginning with "HA" come from the Household Adult Questionnaire.

To distinguish the multiply imputed variables in the Multiply Imputed Data Set from their unimputed counterparts in other public-use files, these variables have been assigned new names ending in "MI". To stay within the eight-character limit, other characters within the original variable's name may have been deleted. For example, the variable recording the bone mineral density of the femur neck, which is called "BDPFNBMD" in the public use files, has become "BDPFNDMI" in the Multiply Imputed Data Set. A list of the variables that have been multiply imputed, and the variables to which they correspond in the NHANES III public use files, is shown in Table 3 below. (Two of the variables, HAN6SRMI and HAR3RMI, are recodes of variables from the Adult Questionnaire which have no counterparts in other public-use files.)

TABLE 3: Imputed variables in the NHANES III Multiply Imputed Data Set, and the variables from other NHANES III public-use files to which they correspond

In MI Data Set	In Public Use Files	Description	Age range
HOUSEHOLD FAMILY QUESTIONNAIRE ITEMS			
DMPPIRMI	DMPPIR	Poverty income ratio	2 mo +
HFF1MI	HFF1	Anyone living here smoke cigs in home	2 mo +
HOUSEHOLD ADULT QUESTIONNAIRE ITEMS			
HAB1MI	HAB1	Self-rating of health status	17 yr +
HAM5MI	HAM5	How tall are you without shoes-inchs	17 yr +
HAM6MI	HAM6	How much do you weigh in pounds	17 yr +
HAN6SRMI	********	Beer/wine/liquor (recode)	17 yr +
HAQ1MI	HAQ1	Condition of SPS natural teeth	17 yr +
HAR3RMI	********	Smoke cigarettes now (recode)	17 yr +
HAT28MI	HAT28	Compare own activity level to others	17 yr +
HAZAK1MI	HAZA8AK1	K1 for first BP measurement (home)	17 yr +
HAZAK5MI	HAZA8AK5	K5 for first BP measurement (home)	17 yr +
HAZBK1MI	HAZA8BK1	K1 for second BP measurement (home)	17 yr +
HAZBK5MI	HAZA8BK5	K5 for second BP measurement (home)	17 yr +
HAZCK1MI	HAZA8CK1	K1 for third BP measurement (home)	17 yr +
HAZCK5MI	HAZA8CK5	K5 for third BP measurement (home)	17 yr +
HOUSEHOLD YOUTH QUESTIONNAIRE ITEMS			
HYD1MI	HYD1	How is health of SP in general	2 mo-16 yr
HYF2MI	HYF2	Condition of natural teeth	2 yr-16 yr
BONE DENSITOMETRY			
BDPFNDMI	BDPFNBMD	Bone mineral density of femur neck-gm/cm sq	20 yr +
BDPINDMI	BDPINBMD	BMD of intertrochanter region-gm/cm sq	20 yr +
BDPKMI	BDPK	K value for scan	20 yr +
BDPTOAMI	BDPTOARE	Bone area of total region - cm sq	20 yr +
BDPTODMI	BDPTOBMD	Bone minrl density total region-gm/cm sq	20 yr +
BDPTRDMI	BDPTRBMD	BMD of trochanter region - gm/cm sq	20 yr +
BDPWTDMI	BDPWTBMD	BMD of Ward's triangle region-gm/cm sq	20 yr +
BODY MEASUREMENTS			
BMPBUTMI	BMPBUTTO	Buttocks circumference (cm)	2 yr +
BMPHEAMI	BMPHEAD	Head circumference (cm)	2 mo-7 yr
BMPHTMI	BMPHT	Standing height (cm)	2 yr +
BMPKNEMI	BMPKNEE	Knee height (cm)	60 yr +
BMPRECMI	BMPRECUM	Recumbent length (cm)	2 mo-3 yr
BMPSTHMI	BMPSITHT	Sitting height (cm)	2 yr +
BMPSB1MI	BMPSUB1	First subscapular skinfold (mm)	2 mo +
BMPSB2MI	BMPSUB2	Second subscapular skinfold (mm)	2 mo +
BMPSP1MI	BMPSUP1	First suprailiac skinfold (mm)	2 yr +
BMPSP2MI	BMPSUP2	Second suprailiac skinfold (mm)	2 yr +
BMPTR1MI	BMPTRI1	First triceps skinfold (mm)	2 mo +

BMPTR2MI	BMPTRI2	Second triceps skinfold (mm)	2 mo +
BMPWSTMI	BMPWAIST	Waist circumference (cm)	2 yr +
BMPWTMI	BMPWT	Weight (kg)	2 mo +

TABLE 3 (continued): Imputed variables in the NHANES III Multiply Imputed Data Set, and the variables from other NHANES III public-use files to which they correspond

In MI Data Set	In Public Use Files	Description	Age range
FUNDUS PHOTOGRAPHY			
FPPSUDMI	FPPSUDRU	Summary drusen score	40 yr +
FPPSUMMI	FPPSUMAC	Summary age-related maculopathy score	40 yr +
FPPSURMI	FPPSURET	Summary diabetic retinopathy score	40 yr +
BLOOD AND URINE ASSAY ITEMS			
FEPMI	FEP	Serum iron (ug/dl)	1 yr +
FRPMI	FRP	Ferritin (ng/ml)	1 yr +
HDPMI	HDP	Serum HDL cholesterol (mg/dL)	4 yr +
HGPMI	HGP	Hemoglobin (g/dl)	1 yr +
HTPMI	HTP	Hematocrit (%)	1 yr +
MCPSIMI	MCPSI	Mean cell hemoglobin: SI	1 yr +
MHPMI	MHP	Mean cell hemoglobin concentration (g/dl)	1 yr +
MVPSIMI	MVPSI	Mean cell volume: SI (fl)	1 yr +
PBPMI	PBP	Lead (ug/dl)	1 yr +
PHPFSTMI	PHPFAST	Length of calculated fast (in hours)	1 yr +
PXPMI	PXP	Serum transferrin saturation (%)	1 yr +
RCPMI	RCP	Red blood cell count (x 10**6)	1 yr +
RWPMI	RWP	Red cell distribution width (%)	1 yr +
SEPMI	SEP	Selenium (ng/ml)	12 yr +
TCPMI	TCP	Serum cholesterol (mg/dL)	4 yr +
TGPMI	TGP	Serum triglycerides (mg/dL)	4 yr +
TIPMI	TIP	Serum TIBC (ug/dl)	1 yr +
REPLICATE BLOOD PRESSURE (3x) FROM MEC EXAMINATION			
PEP6G1MI	PEP6G1	K1, systolic, for 1st BP (mmHg)	5 yr +
PEP6G2MI	PEP6G2	K4, diastolic, for 1st BP(mmHg)	5 yr-19 yr
PEP6G3MI	PEP6G3	K5, diastolic, for 1st BP (mmHg)	5 yr +
PEP6H1MI	PEP6H1	K1, systolic, for 2nd BP (mmHg)	5 yr +
PEP6H2MI	PEP6H2	K4, diastolic, for 2nd BP(mmHg)	5 yr-19 yr
PEP6H3MI	PEP6H3	K5, diastolic, for 2nd BP (mmHg)	5 yr +
PEP6I1MI	PEP6I1	K1, systolic, for 3rd BP (mmHg)	5 yr +
PEP6I2MI	PEP6I2	K4, diastolic, for 3rd BP(mmHg)	5 yr-19 yr
PEP6I3MI	PEP6I3	K5, diastolic, for 3rd BP (mmHg)	5 yr +

For each imputed variable, an imputation flag has been created and placed in the file CORE.DAT. These imputation flags have the same names as the imputed variables, except that the "MI" suffix is replaced with "IF". An imputation flag takes three possible values: "0" denotes that the variable in question is not applicable, "1" denotes that the variable is observed, and "2" denotes that the variables has been imputed.

The CORE.DAT file also contains a selection of variables from the Household Adult, Youth, and Family questionnaires that were not given high priority for imputation, but were used or considered for use as predictors in the imputation models for some or all of the age classes. These variables are given the same names as their counterparts in other public-use files. For these variables, a blank value indicates that the variable does not apply to the individual in question. Other types of missing values in these variables are indicated by 8-fills ("Blank but applicable") and 9-fills ("Don't know"). A list of these variables appears in Table 4 below.

TABLE 4: Variables in the NHANES III Multiply Imputed Data Set that served as potential predictors in the imputation models but were not imputed

Name	Description	Age range
HOUSEHOLD FAMILY QUESTIONNAIRE ITEMS		
HFA7	Highest grade or yr of school attended	2 mo +
HFA8	Finished highest grade/yr attended	2 mo +
HFA12	Marital status	14 +
HOUSEHOLD YOUTH QUESTIONNAIRE ITEMS		
HYE1G	Doc ever say had asthma	2 mo-16 yr
HYE1H	Doc ever say had chronic bronchitis	2 mo-16 yr
HYE6A	Doc ever say had high blood pressure	4 yr-16 yr
HYE6B	Doc ever say had high blood cholesterol	4 yr-16 yr
HYE15	Has ever had anemia	2 mo-16 yr
HYH2	Have trouble seeing w/one or both eyes	3 yr-16 yr
HYH10	Ever had troub hearing w/1 or both ears	2 mo-16 yr

TABLE 4 (continued): Variables in the NHANES III Multiply Imputed Data Set that served as potential predictors in the imputation models but were not imputed

Name	Description	Age range
HOUSEHOLD ADULT QUESTIONNAIRE ITEMS		
HAC1A	Ever told had arthritis	17 yr +
HAC1B	Which type of arthritis	17 yr +
HAC1C	Ever told had congestive heart failure	17 yr +
HAC1D	Ever told had stroke	17 yr +
HAC1E	Ever told had asthma	17 yr +
HAC1F	Ever told had chronic bronchitis	17 yr +
HAC1G	Ever told had emphysema	17 yr +
HAC1H	Ever told had hay fever	17 yr +
HAC1I	Ever told had cataracts	17 yr +
HAC1J	Ever told had goiter	17 yr +
HAC1K	Ever told had thyroid disease	17 yr +
HAC1L	Ever told had lupus	17 yr +
HAC1M	Ever told had gout	17 yr +
HAC1N	Ever told had skin cancer	17 yr +
HAC1O	Ever told had other type of cancer	17 yr +
HAD1	Ever told had diabetes	17 yr +
HAE2	Ever told had high blood pressure	17 yr +
HAE4A	Ever told to take prescr med for HBP	17 yr +
HAE4B	Ever told to ctrl/lose wt for HBP	17 yr +
HAE5A	Now taking prescr med for HBP	17 yr +
HAE5B	Is now ctrl/lose wt for HBP	17 yr +
HAE6	Ever had blood cholesterol checked	17 yr +
HAE7	Ever told had high cholesterol	17 yr +
HAF1	Ever had chest pain/discomfort	17 yr +
HAF10	Ever told had heart attack	17 yr +
HAG2	Ever had back pain most days for 1 mo	20 yr +
HAG3	Have back pain in past 12 months	20 yr +
HAG5A	Ever told had fractured hip	20 yr +
HAG5B	Ever told had fractured wrist	20 yr +
HAG5C	Ever told had fractured spine	20 yr +
HAG11	Ever told had osteoporosis	20 yr +
HAG12	Were treated for osteoporosis	20 yr +
HAN6HS	Beer and lite beer - times/month	17 yr +
HAN6IS	Wine, champagne - times/month	17 yr +
HAN6JS	Hard liquor - times/month	17 yr +
HAP1	Have total blindness	17 yr +
HAP1A	If yes, one or both eyes	17 yr +
HAP2	Use glasses, contacts, or both	17 yr +
HAP3	Trouble seeing with one or both eyes	17 yr +
HAP10	Have total deafness	17 yr +
HAP10A	If yes, one or both ears	17 yr +
HAR1	Smoked 100 cigarettes in life	17 yr +
HAR3	Smoke cigarettes now	17 yr +
HAR14	Used chewing tobacco, snuff	17 yr +
HAR16	Chew tobacco, snuff now	17 yr +
HAR23	Smoked 20 cigars in life	17 yr +

HAR24	Smoke cigars now	17 yr +
HAR26	Smoked 20 pipes of tobacco in life	17 yr +
HAR27	Smoke pipe now	17 yr +

Finally, CORE.DAT also contains a number of general information variables including demographic information from the screener, geographic identifiers, design information and survey weights. These variables (those not ending in "MI", "IF", or "MP") have names that are identical to those in the NHANES III public-use data files, and the data values within these variables should be entirely consistent with the public-use file data. These variables are listed in Table 5 below.

TABLE 5: General information variables in the NHANES III Multiply Imputed Data Set

In MI Data Set	In Public Use Files	Description
SEQN	SEQN	Sequence number
DMPFSEQ	DMPFSEQ	Family sequence number
DMPSTAT	DMPSTAT	Examination/interview status
DMARETHN	DMARETHN	Race-ethnicity
DMARACER	DMARACER	Race
DMAETHNR	DMAETHNR	Ethnicity
HSSEX	HSSEX	Sex
HSDOIMO	HSDOIMO	Date of screener (month)
HSAGEIR	HSAGEIR	Age at interview (screener) - qty
HSAGEU	HSAGEU	Age at interview (screener) - unit
HSAITMOR	HSAITMOR	Age in months at interview (screener)
HSFSIZER	HSFSIZER	Family size (persons in family)
HSHSIZER	HSHSIZER	Household size (persons in dwelling)
DMPCNTYR	DMPCNTYR	County code
DMPFIPSR	DMPFIPSR	FIPS code for State
DMPMETRO	DMPMETRO	Rural/urban code based on USDA code
DMPCREGN	DMPCREGN	Census region, weighting (Texas in south)
SDPPHASE	SPPHASE	Phase of NHANES III Survey
SPPPSU6	SPPPSU6	Total NHANES III pseudo-PSU
SDPSTRA6	SDPSTRA6	Total NHANES III pseudo-stratum
WTPFQX6	WTPFQX6	Total interviewed sample final weight
WTPQRP1	WTPQRP1	Replicate 1 final interview weight
WTPQRP2	WTPQRP2	Replicate 2 final interview weight
WTPQRP3	WTPQRP3	Replicate 3 final interview weight
WTPQRP4	WTPQRP4	Replicate 4 final interview weight
WTPQRP5	WTPQRP5	Replicate 5 final interview weight
WTPQRP6	WTPQRP6	Replicate 6 final interview weight
WTPQRP7	WTPQRP7	Replicate 7 final interview weight
WTPQRP8	WTPQRP8	Replicate 8 final interview weight
WTPQRP9	WTPQRP9	Replicate 9 final interview weight
WTPQRP10	WTPQRP10	Replicate 10 final interview weight
WTPQRP11	WTPQRP11	Replicate 11 final interview weight
WTPQRP12	WTPQRP12	Replicate 12 final interview weight
WTPQRP13	WTPQRP13	Replicate 13 final interview weight
WTPQRP14	WTPQRP14	Replicate 14 final interview weight
WTPQRP15	WTPQRP15	Replicate 15 final interview weight
WTPQRP16	WTPQRP16	Replicate 16 final interview weight
WTPQRP17	WTPQRP17	Replicate 17 final interview weight
WTPQRP18	WTPQRP18	Replicate 18 final interview weight
WTPQRP19	WTPQRP19	Replicate 19 final interview weight
WTPQRP20	WTPQRP20	Replicate 20 final interview weight
WTPQRP21	WTPQRP21	Replicate 21 final interview weight
WTPQRP22	WTPQRP22	Replicate 22 final interview weight
WTPQRP23	WTPQRP23	Replicate 23 final interview weight
WTPQRP24	WTPQRP24	Replicate 24 final interview weight
WTPQRP25	WTPQRP25	Replicate 25 final interview weight
WTPQRP26	WTPQRP26	Replicate 26 final interview weight
WTPQRP27	WTPQRP27	Replicate 27 final interview weight
WTPQRP28	WTPQRP28	Replicate 28 final interview weight
WTPQRP29	WTPQRP29	Replicate 29 final interview weight

WTPQRP30	WTPQRP30	Replicate 30 final interview weight
WTPQRP31	WTPQRP31	Replicate 31 final interview weight
WTPQRP32	WTPQRP32	Replicate 32 final interview weight
WTPQRP33	WTPQRP33	Replicate 33 final interview weight
WTPQRP34	WTPQRP34	Replicate 34 final interview weight

TABLE 5 (continued): General information variables in the NHANES III Multiply Imputed Data Set

In MI Data Set	In Public Use Files	Description
WTPQRP35	WTPQRP35	Replicate 35 final interview weight
WTPQRP36	WTPQRP36	Replicate 36 final interview weight
WTPQRP37	WTPQRP37	Replicate 37 final interview weight
WTPQRP38	WTPQRP38	Replicate 38 final interview weight
WTPQRP39	WTPQRP39	Replicate 39 final interview weight
WTPQRP40	WTPQRP40	Replicate 40 final interview weight
WTPQRP41	WTPQRP41	Replicate 41 final interview weight
WTPQRP42	WTPQRP42	Replicate 42 final interview weight
WTPQRP43	WTPQRP43	Replicate 43 final interview weight
WTPQRP44	WTPQRP44	Replicate 44 final interview weight
WTPQRP45	WTPQRP45	Replicate 45 final interview weight
WTPQRP46	WTPQRP46	Replicate 46 final interview weight
WTPQRP47	WTPQRP47	Replicate 47 final interview weight
WTPQRP48	WTPQRP48	Replicate 48 final interview weight
WTPQRP49	WTPQRP49	Replicate 49 final interview weight
WTPQRP50	WTPQRP50	Replicate 50 final interview weight
WTPQRP51	WTPQRP51	Replicate 51 final interview weight
WTPQRP52	WTPQRP52	Replicate 52 final interview weight