

December 21, 1992

Assessment of Trial Imputations for NHANES III

Final Report

Roderick J.A. Little and Donald B. Rubin
Datametrics Research, Inc.

Executive Summary

The Third National Health and Nutrition Examination Survey (NHANES III) is the third in a series of periodic surveys conducted by the National Center for Health Statistics to assess the health and nutritional status of the U.S. population. The NHANES III survey began in 1988 and will continue through 1994. Phase 1 of the survey, conducted in 1988-1991, involved data collection on a national probability sample of the U.S. population. This survey is subject to non-negligible levels of unit and item nonresponse, both in its interview and its examination components. In previous surveys, nonresponse was handled primarily using weighting adjustments. The goal of this project was to develop and compare alternative missing-data adjustments based on single and multiple imputation of missing data, and to assess mixed weighting and imputation strategies.

The data set used to develop the imputation models consisted of a subset of the Phase 1 data of NHANES III. The subset was restricted to adults over 17 years old, and included 6 completely-observed sample frame/household screening variables, namely:

stand, region, SMSA, age, sex, race;

22 interview variables, missing in about one out of every five cases, namely:

education, marital status, family income, self-reported health status, diabetes, hypertension/high blood pressure (diagnosis, medication ever, medication now), high cholesterol (diagnosis, medication ever, medication now), chest pain, heart attack diagnosis, cigarettes now, activity status, food/alcohol/cigarette in last 30 minutes, alcohol, drugs, height, weight, blood pressure (systolic, diastolic);

finally, 12 MEC variables from three components of the examination were included:

Body Measurements: height, weight, waist circumference, buttocks circumference

Blood Pressure: systolic 1-3, diastolic 1-3

Lipids: Serum Total Cholesterol, HDL Cholesterol.

These variables were missing in about one out of every three cases

Three imputation methods were developed as part of this project. Westat, Inc. imputed about 690 missing values for six MEC variables, namely log height, log weight, systolic and diastolic blood pressure, total serum cholesterol and HDL cholesterol, using two closely-related regression imputation methods. Datametrics Research and Joe Schafer multiply imputed 23 variables in the adult questionnaire and MEC, for over 4000 cases involving all types of nonresponse, using a Bayesian simulation method (DMS) for mixed normal and categorical data.

The major differences in the number of values imputed reflect in part differences in strategies regarding how much of the missing-data problem should be handled by

weighting adjustments and how much by imputation. Specifically, Westat propose a strategy (ILO -- imputation-LO) that confines imputation to item nonresponse for individuals who received the MEC examination, whereas Datametrics and Schafer apply a strategy (IHI -- imputation-HI) that also imputes all the other sources of unit and item nonresponse. In this report the relative strengths of weighting and imputation are debated, and a middle course (IMID) is proposed for current implementation that deals with nonresponse for individuals missing the questionnaire and MEC examination by a weighting adjustment, but multiply imputes important variables in the data set subject to other forms of nonresponse.

Comparisons of the imputes from the Westat and Datametrics / Schafer methods on the subset of values imputed by Westat indicate that both methods by and large produced reasonable imputations of the missing values, although Schafer's analysis did uncover the need to edit out some implausible respondent values. The Datametrics / Schafer DMS method is recommended, in view of its ability to handle general patterns of missing data, and to provide multiple imputations of the missing values that allow the uncertainty due to nonresponse to be incorporated into the analysis using standard complete-data methods.

Our main recommendations based on this study are as follows:

1. We propose the following IMID option for current implementation: cases missing the entire questionnaire and MEC examination are weighted, using the first-stage weight adjustments developed in Ezzati and Khare (1992). Missing values for key variables in other cases are replaced by $M = 5$ multiple-imputations, created by the DMS methodology described in this report. Missing-values of less important variables in the analysis file would be replaced by missing-value codes.
2. Multiply-imputed data bases should include in the documentation some summary of methods of analysis for multiply-imputed data, including examples of how to create and analyze repeated complete data sets.
3. In future, multiple imputation methodology can be expanded to more variables and more incomplete cases, as the method becomes established with users, documentation is refined, and computing limits are reduced by advances in statistical computing.
4. Ways should be considered of user-friendly packaging of multiple imputations in data bases, e.g. for SAS, SPSS, BMDP etc. For example, consider whether it is better to a) use pointers to the multiple imputes, or b) repeat each incomplete variable M times, with consequent duplication of complete information.
5. Future research on the first-stage weighting adjustment developed in Ezzati and Khare (1992) should consider the development of weights based on the estimated response propensity (Rosenbaum and Rubin 1983; Little 1986), and assess the importance of the component of variability missed by the current procedure.

6. More generally research is needed on improving the estimation of complete-data standard errors. Modeling to reduce the excessive variability of current SUDAAN estimates is needed.

7. Studies should be conducted concerning the reasons of nonresponse in the NHANES, and a comparison of respondents and nonrespondents with respect to observed characteristics. Is the missing-data mechanism missing completely at random, or do respondents and nonrespondents differ? The tables created for the current project and the work on weighting adjustments in Ezzati and Khare (1992) give some indication that differences between respondents and nonrespondents are minor, but more systematic assessments are needed.

8. Finally, projects such as the one we are assessing here demonstrate feasibility, and provide useful descriptive and comparative information, but they do not provide objective information about the operating characteristics of procedures - for example, do $(1-\alpha)$ confidence intervals created from the multiple-imputed data sets developed under the EMID procedure really cover the target quantity in a proportion $(1-\alpha)$ of repeated samples for NHANES? To convince skeptics, what is needed is an honest frequentist evaluation via simulation to assess the methods discussed here. That is, create hypothetical populations, draw samples from each, and apply existing and alternative procedures to the samples, and thereby assess the validity of inference about population quantities. The assessment should distinguish between deficiencies due to complete-data methodology, such as problems with the SUDAAN standard errors, and deficiencies due to alternative nonresponse adjustment methods such as weighting or multiple imputation.

1. Introduction

This report reviews work on imputation for NHANES III by Datametrics, Joe Schafer and Westat, provides an assessment of how we think imputation might be used in NHANES public use files currently under preparation, and suggests directions for future research in this area. Extensive documentation of the impressive body of work upon which these comments are based is provided in the memoranda listed in Table 1, which are attached as Appendix 1.

As the next section indicates, NHANES III contains a substantial number of missing values. There are three major problems created by missing data in data bases such as NHANES. First, if the units with missing values are systematically different from the units with complete data, naive analyses that ignore these differences will be biased. Second, the existence of missing data implies a loss of information, so that estimates will be less efficient than planned. Finally, most standard statistical methods are designed for complete (that is, rectangular) data sets; thus missing data often render the analysis of data from a study more complicated.

One common approach to missing values in public-use data bases is simply to flag missing values using missing-value codes. The use of missing-value codes is the easiest solution to missing data for the data producer, and it may be sensible for variables with very modest levels of nonresponse (for example, 1%-2% of cases), where the effort of developing imputation models has little pay-off. However, the approach simply passes the problem of analysis with missing-data on to the user. Most statistical packages discard cases from a statistical analysis that contain missing values on included variables (the method labeled complete-case analysis in Little and Rubin 1987), an option that is inefficient and potentially leading to nonresponse bias. Another problem with flagging missing values is that investigators may obtain different results for the same analysis applied to the same data set, because of differences in how the missing data are handled. A final problem with simply flagging missing values in variables with modest levels of nonresponse is that modest levels can become substantial in subgroups of interest (for example, in particular age x race x sex x urban/rural subgroups)

Imputation by the data producer, that is filling in the missing values with suitable estimates, yields a rectangular file that is amenable to standard analyses, retains incomplete cases, and promotes a uniform treatment of the missing data by users (to the extent that the supplied imputes are used). However, simple imputation methods, such as substituting the overall mean, can lead to worse estimates than those from complete-case analysis. The imputations need to condition on available information for nonrespondents, and to be based on a well-chosen implicit or explicit statistical model. For discussions of imputation methods in surveys see for example Kalton and Kasprzyk (1982, 1986), Madow et al. (1983, Volume 2), Little (1986, 1988a) and Rubin (1987).

A disadvantage of replacing missing values by a single estimate is that uncertainty in the imputation is not reflected, and thus inferences based on the filled-in data tend to be

understate uncertainty -- in particular, confidence intervals are too narrow, and P-values for hypothesis tests are too low. Multiple imputation (Rubin 1987) is an extension of imputation that replaces each missing value by $M > 1$ values, drawn from the predictive distribution of the missing values under a model. This extension retains the ability to analyze the data by complete-data methods, but also allows uncertainty in the imputation to be assessed and formally incorporated into the analysis in a straightforward way, as discussed in Section 8.

We believe that multiple imputation under a judiciously-chosen model is one of the best ways of handling missing values in a public-use data base; reasons are given in Rubin (1987) and Little (1988a). Recent work by Fay (1991, 1992) has attempted to undermine the method with examples where the imputer and user adopt different models, but we believe these criticisms are based on a misinterpretation of his results and hence not well founded. A more serious practical objection is that with large multivariate data sets, the modeling task of creating multiple imputes for all the missing data may be formidable. Less valid but simpler methods may be adequate, particularly for variables of secondary importance.

One alternative to imputation is to drop incomplete cases, but weight the complete cases to compensate for nonresponse bias. The simplest form of this method is to weight respondents by the inverse of response rates computed within adjustment cells based on values of variables measured for respondents and nonrespondents. This method was used to handle nonresponse in previous NHANES surveys, and weighting methods for NHANES III are considered in Ezzati and Khare (1992). A useful extension of this approach is to base adjustment cells on predictions from a logistic regression of a response indicator on observed covariates (Rosenbaum and Rubin 1983; Little 1986). Weighting provides a useful mechanism for bias adjustment, but it is not a good tool for controlling variance, especially with many survey variables. One reason is that weighting provides the same nonresponse adjustment for all variables, regardless of their degree of association with nonresponse, whereas imputation allows the nonresponse adjustment to be tuned to each missing variable. Imputation also has optimal properties if the imputation model is correct. For more discussion see Little (1986), Rubin (1987).

Two types of missing data are commonly distinguished in surveys -- *unit* nonresponse, where basic sample and household information is available for an individual but the survey information is missing through noncontact or refusal to participate, and *item* nonresponse, where an interview was conducted but particular variables in the survey are missing. A common strategy is to deal with unit nonresponse by weighting adjustments and item nonresponse by imputation. In NHANES III, this simple recipe is complicated by the fact that the survey has two data collection instruments, namely the questionnaire and the MEC examination, and more than two if the components of the MEC examination are treated as distinct. Hence, some thought is needed to assess the most judicious use of weighting and imputation for this survey.

The structure of this report is as follows. Section 2 discusses the pattern of missing data in NHANES III and introduces some terminology for types of missing data. Section 3 discusses general properties of weighting and imputation adjustments, and presents the specific alternatives considered in this evaluation. Section 4 discusses the current weighting methods for NHANES III developed in Ezzati and Khare (1992), proposes future enhancements, and relates this work to the imputation methods developed in this project. Section 5 presents and argues in favor of the **MIID** strategy for addressing the problem of missing data in NHANES III. Section 6 discusses the estimation of standard errors of NHANES III estimates. Section 7 compares the particular imputation models used by Datametrics / Schafer and Westat in this project, and presents arguments in favor of the multiple imputation methods of Datametrics / Schafer. Section 8 discusses some issues in the implementation of multiple imputation, and sketches the analysis of multiply-imputed data sets by survey users. Finally, Section 9 restates the recommendations given above in the Executive Summary.

2. Types of Missing data in NHANES III

Variables in NHANES III can be usefully classified into three groups:

1. Sample frame / household screening variables
2. Interview variables (family and health history variables)
3. Mobile Examination Center (MEC) variables

Missing data in the sample frame / household screening variables are referred to here as screening nonresponse. The level of screening nonresponse is minor, and for the purposes of this report we shall treat household/screening variables as fully observed.

Missing data in the interview variables are referred to here as interview nonresponse. The interview data consist of family questionnaire variables, and health variables obtained for sampled individuals. Schafer (Document 3, Table 1) reports 15%-16% of values missing on the family questionnaire variables, and 18%-23% of values missing on the selected adult questionnaire variables, the latter reflecting somewhat higher amounts of item nonresponse. (The high nonresponse rate of 62% for cholesterol diagnosis is misleading, since much of it attributable to a filter in the questionnaire.)

Missing data in the MEC variables are referred to here as examination nonresponse. Schafer (Document 3, Table 1) reports levels of nonresponse of 31%-34% of examination nonresponse.

When missing or present as a set, these three blocks of variables (screening, interview, examination) have an approximately monotone structure, with screening variables fully observed, questionnaire variables missing when the interview is not conducted, and examination variables missing when either (a) the interview is not conducted or (b) the interview is conducted but the MEC examination does not take place. However, nonresponse for individual items spoils this monotone structure.

Table 1. Description of Memoranda for NHANES III Imputation Project

Doc	Date	Author	Title	Short Title
1.	June 12	Little/Rubin	NHANES Survey: Evaluation of Imputation Methods	
2.	June 22	Schafer	Recommendations on Model for NHANES III Imputation Project	
3.	July 9	Schafer	Rates and Patterns of Missingness in the NHANES III Imputation File	
4.	July 31	Fahimi	Imputation of MEC Variables	WES 1
5.	Aug 5	Fahimi	Alternative Imputations for the Cholesterol Measurement	WES 2
6.	July 29	Rowland	Evaluation of WESTAT Single Variable Measurements	WES 3
7.	Aug 3		Tables 1-3 for WES 3	WES 4
8.	Aug 4	Schafer	Multiply Imputed Data Files for NHANES III	MI 1
9.	Aug 7	Ezzati/Khare	NHANES III Imputation Project	
10.	Aug 7	Ezzati/Khare	Tables 1 and 2 for memo 10	MI 2
11.	Aug 7	Rowland	Evaluation of Single Variable and Multi-Variable Imputations	MI 3
12.	Aug 10	Schafer	NHANES III Imputation Group Meeting (tables and figures)	
13.	Sept 3	Ezzati/Khare	NHANES III Imputation Project - additional tables 1A-D, 2A-D, 3A-B	
14.	Sept 7	Schafer	Model and Procedures Used to Create Multiply Imputed Datasets for NHANES III	
15.	Sept 11	Schafer	Exploratory Analysis of Imputed Values in the NHANES III Imputation Project	
16.	Sept 17	Judkins, Winglee	Variance Estimation with Imputed Data for NHANES III	
17.	Sept. 22	Schafer	Westat's Recommendations on Missing Data and Variance Estimation Procedures for NHANES III	

The MEC examination involves *components* corresponding to related sets of measurements. Item nonresponse for the MEC often arises when all the variables in a particular component are missed. Thus, item nonresponse for the MEC variables is classified as either *component* nonresponse, where an individual is examined but all the variables in one component of the exam are missing, or *item-within-component* nonresponse for particular items, typically a minor problem.

3. Weighting vs. Imputation

As noted in the introduction, two general strategic approaches are commonly considered for dealing with missing values: either drop incomplete cases from the analysis and apply a weighting adjustment to the remaining cases, or retain cases in the file and impute one or more values for each missing datum. A key decision concerns the extent to which weighting and imputation are used to handle the various types of nonresponse.

Westat's (Document 18) proposed procedures for handling missing data, henceforth labeled WES, handle nearly all the missing data problems by weighting adjustments, confining imputation to MEC component and item-within-component nonresponse - that is, to cases that received the MEC exam but were missing one or more components of the exam (e.g. all body measurements), or missing items within a component. For example, as shown in Tables 4-6 of Document 3, only about 17% of the missing values for the MEC variables studied here was attributable to component and item-within-component nonresponse, accounting for about 6% of all missing values. Under this proposal, all cases not receiving the MEC exam would be dropped from the analysis file, with the remaining cases reweighted to compensate for this component of missing data. We call this strategy ILO for imputation-LO, since its use of imputation is limited.

The multiple imputation approach, labeled henceforth as DMS, was developed by Datametrics and Joe Schafer and implemented on the test NHANES III data sets by Dr. Schafer. It is based on Bayesian simulation for the general location model for mixed normal and categorical data (Documents 2 and 14). Maximum likelihood methods for incomplete data from this model are described in Little and Schluchter (1985), and the Bayesian methods based on the Gibbs' sampler are described in Schafer's Ph.D. dissertation (Schafer 1991). DMS was applied to *all* missing values of variables in the trial data set. This approach relies much more on imputation and less on weighting than the Westat approach, since all cases are retained in the file and the role of weighting confined to sampling adjustments. We call this strategy IHI for imputation-HI, since it uses imputation to handle all nonresponse.

An intermediate strategy between ILO and IHI is to drop and weight for cases missing both the interview and examination variables, but retain and impute cases with the interview present but the examination missing. We call this strategy IMID.

Precise formulations of the IMID and ILO strategies require the specification of what is meant by an interview or exam being present or missing -- for example, an interview where some basic family questionnaire variables were recorded, but all the adult health variables were missing, might be treated as missing even though a small number of interview variables are present. Such details are not addressed here, the focus being on a broader assessment of the three alternative strategies. The choice between ILO, IMID and IHI should be based on statistical properties of the methods, as well as more practical considerations such as ease of implementation and use. Some general comments may assist in the choice:

In favor of imputation:

1) Imputation (dumb or smart) has the clear advantage of retaining values of recorded variables in incomplete cases, which are dropped by weighting (or, if retained in the file, are given zero weight). From this perspective, imputation becomes increasingly attractive as the relative number of observed variables in an incomplete case increases. In particular IMIN has the undesirable feature of sacrificing information on individual interviews from people who were interviewed but missed the MEC exam.

2) Weighting is an inferior tool for missing-data adjustment, particularly when the set of observed covariates is extensive. It can provide a useful mechanism for bias adjustment if weighting classes are appropriately formulated, but is not a good tool for controlling variance, especially with many survey variables. Imputation also has optimal properties if the imputation model is correct. For more discussion see Little (1986), Rubin (1987).

In favor of weighting:

3) Weighting has the operational advantage of providing a single adjustment for all variables simultaneously: it is often much less work. As Westat correctly point out (Document 16), imputing the entire set of examination variables for cases who missed the MEC exam could be a mammoth task in multivariate modeling; we are not sure how many variables are involved, but presumably it is much greater than the set of about 30 variables included in the test analysis.

4) Weighting avoids potential problems of inconsistencies between imputed values of missing variables, such as can arise when using an imputation model that fails to satisfy logical editing constraints. We tend to view this problem as more an inconvenience than a serious issue, arising from the tendency to want to treat an imputed record as the truth rather than an estimate with uncertainty. Also, recorded data usually need to be edited for the same reasons.

5) A more serious issue can occur when imputation models impute subsets of variables (for example, components of the MEC) independently rather than jointly, and thus fail to incorporate conditional associations between variables in different sets. Such associations are preserved by weighting. This potential limitation in imputation methods need to be documented in user information. However we note that Schafer's analysis managed to accomplish joint imputation of certain important examination variables that occurred in three different components of the exam.

More generally, the issue of what constitutes a "correct imputation model" deserves extended comment, for it in fact is the only theoretical issue that limits the applicability of multiple imputation: general theory in Rubin (1987, chapter 4) shows that "proper" imputations under a correct model provides valid subsequent inferences from the model-based or randomization / frequentist perspectives. Also, certain assumptions of the model can be relaxed (such as normality of error distributions) without affecting the validity of

multiple imputation inferences in large samples, and approximate models often work extremely well -- as George Box is quoted as saying, "All models are wrong, but some are very useful." In general, the propriety of the complete-data model is far more important than that of the multiple imputation model, because the latter is only used for the fraction of information that is missing.

The following guidelines are useful when selecting an imputation model. In principle, any useful predictor of a missing value should be part of the model, including interactions that appear to have any predictive power. Predictive ability as measured, for example by R^2 in a linear regression, is typically more important than significance tests for coefficients, especially in small samples. Use of transformations to improve the fit of the model is highly desirable, just as with complete data analyses. Variables that are known to be used in future planned analyses should be included, whether or not they appear to be significant or powerful predictors. This seems to be a prescription for including many, many variables, and it is. As discussed in Section 7, the technology of iterative simulation (e.g., the Gibbs sampler), allows multiple imputation of increasingly large models, as computational algorithms and available computing resources improve. For the DMS models fitted here, the software developed by Schafer (1991) also provides the tools for diagnostic checking, model modification, and editing.

The inclusion of "too many (e.g. zero true coefficient)" predictors does no harm to the validity of the resulting inferences, nor does their exclusion even if the data analyst decides to include them--this is true despite Fay's recent (1991, 1992) proclamations to the contrary, which are based on a misinterpretation of his analytic and simulation results. Consequently, the objective is to include all predictive variables in the model, although worthless variables can always be excluded even if a future analyst uses them; the model only fails to the extent that the left-out variables orthogonal to the included variables have predictive power. Note the use here of "orthogonal" components of the left-out variables, not the left-out variables themselves--with large data sets and many included variables, often the left-out orthogonal components are very minor, even if the left-out variables are important before adjusting for the included variables. Several real world examples support this claim, e.g., the occupation and industry example briefly described in Rubin and Schenker (1990), which has been studied by a variety of statisticians and social scientists without any evidence of failure.

4. The Current Proposal for Weighting NHANES III.

The nonresponse weighting adjustment for NHANES III developed in Ezzati and Khare (1992) computes weights for the ILO procedure; that is, non-zero weights are assigned only to individuals who received the MEC examination. The weights are computed as the product $w = w_1 w_2$, where w_1 is the inverse of estimated response rate of individuals for the survey questionnaire, and w_2 is the inverse of the estimated response rate for the individual exam, given completion of the questionnaire. The weights w_1 are computed as inverse

response rates within a 72-cell cross-classification of all cases by the screening variables age (2 categories), race (3), region (3), SMSA (2) and household size (2). The weight w_2 is computed within a 72-cell cross-classification of cases responding to the questionnaire, by screening variables age (2), race (3), household size (2) and questionnaire variables income (3) and self-reported health status (2). The weights w are then multiplied by the sample design weight and a post-stratification factor.

We make the following comments about this procedure.

4.1 Consistent with the ILO strategy, cases who responded to the questionnaire but did not take the examination are assigned a weight $w = 0$, that is, are effectively dropped from the file. If two weights were retained in the file, namely w and w_1 , then analyses involving the questionnaire variables could include these cases and use a weight based on w_1 .

4.2 The need to limit the variance from small adjustment cells results in variables that are very coarsely classified (for example, age in two categories). A refinement that might be useful in this context is to form adjustment cells based on the response propensity, estimated from a logistic regression of response/nonresponse on the classification variables. (E.g. Rosenbaum and Rubin 1983; Little 1986; Czajka et al., 1992; Judkins et al. 1992).

4.3 If the IHI method were adopted, nonresponse weights would be set to one, and weighting confined to adjustments for the survey design and post-stratification. If the MID method were adopted, cases who missed the interview and MEC examination would be dropped and remaining cases assigned the nonresponse weight w_1 .

4.4 The results in Ezzati and Khare (1992) suggest that bias is not a major concern here. The ranges of the stage 1 weights w_1 (1.02 to 1.37), and the stage 2 weights w_2 (1.00 to 1.42) are quite small compared with the range of the final weights (0.4 to 2.6). Furthermore, most of the associations of questionnaire variables with MEC response are modest in size and statistically insignificant after adjusting for the screening variables. Given these findings, the main utility of the questionnaire data may lie in imputation for variance reduction rather than weighting for bias reduction, particularly since Schafer notes in Document 17 that the questionnaire data contain good predictors of some MEC variables.

5. A Proposal for Discussion: IMID.

Taking the previous considerations into account, we propose the following form of IMID as a baseline for discussion, in the context of current plans for addressing nonresponse in NHANES III:

A) Delete cases missing both the interview and MEC exam, and apply their nonresponse weight w_1 to the remaining cases (with adjustments for the survey design and post-stratification).

B) For cases with some interview data present (that is, not dropped in A), multiply-impute for item nonresponse in the interview variables, and also for nonresponse in important examination variables. This strategy allows the observed data in the personal interview to be retained for these cases, and used to predict the missing values of key MEC variables. Other MEC variables are assigned missing value codes.

This scheme is preferred to IHI on pragmatic grounds, since it avoids the imputation of entire records for "unit nonrespondents" that contain only screening variables. The decision seems justified on grounds of simplicity given that covariate information for prediction is limited for these cases. A practical reason for avoiding imputation for these cases is that it tends to reduce the number of multiple imputes per missing value needed to adequately reflect the imputation uncertainty. The analysis of ten multiple imputes (Document 13) suggested that when imputation is applied to unit nonrespondents, 5, and for some estimands more than 5, multiple imputes are needed to stabilize the between-imputation variance. If nonrespondents missing the interview and the MEC examination are weighted, 5 appears to be a reasonable number - this was the number used successfully in the Census Industry and Occupational Recoding project (Rubin and Schenker 1987; Treiman, Bielby and Cheng 1988; Clogg, Rubin, Schenker, Schultz and Weidman 1991).

The IMID scheme is preferred to ILO since a) interview data on cases with interview present and MEC missing are retained in the file, and used to predict key MEC variables. The analysis in Ezzati and Khare (1992) suggests that residual nonresponse bias after adjustment for the screening variables is small. This implies that the information carried in the second-stage weight w_2 is small, and the ILO scheme is approximately equivalent to simply discarding the interview data for interview present/MEC missing cases. With regard to the MEC variables, this approach is equivalent to the strategy of retaining these cases and assigning them missing-value codes to *all* the MEC variables. The IMID scheme improves on this scheme by selectively imputing the more important MEC variables. We expect the set of variables and cases to be multiply imputed could gradually increase as more experience is gained about imputation models for the NHANES data set, and as computational resources improve.

6. Standard Errors: Propagating Uncertainty from the Missing Data

6.1 SUDAAN Linearization vs Replication

A major problem with assessing and properly reflecting the added uncertainty from missing data in NHANES III is that it is widely agreed that current SUDAAN linearization methods for assessing standard errors from the complete data are problematic, given the small number of clusters. The wild fluctuations in design effects across multiple imputations (e.g. Document 10, Table 1) are a reflection of this problem.

Westat's view (Document 16) is that elaborate methods for propagating the additional uncertainty from imputation may not be worth the effort, given their ILO proposal to limit imputation to component and item-within-component nonresponse, and the poor quality of the SUDAAN standard errors. For weighting, they note that SUDAAN methods are not easily applied to assess the added uncertainty from estimating the nonresponse weights; the size of this uncertainty is unclear. They propose applying replication methods, where nonresponse weights are computed for a set of replicate samples, as an alternative approach to variance estimation that takes into account the added uncertainty from nonresponse adjustment. Since this method formally incorporates the component of uncertainty from nonresponse weighting missing in the SUDAAN analyses, Westat suggests that it could form a basis for calibrating the SUDAAN estimates.

However, as Schafer notes in his response (Document 17), the replication method does not overcome the major defect in the SUDAAN approach, that is, the instability of estimates caused by the small number of clusters. Given the instability as reflected in the design effects from multiple imputation, it may be very difficult to measure the added component of sampling error from the estimated nonresponse weights, except perhaps at very gross levels of aggregation. Our view is that better methods of variance measurement require some form of modeling. The development of more reliable estimates of sampling error for their survey designs seems an important research priority for NCHS.

6.2 Imputation Error: Multiple Imputation vs Adjustment Factors

The other issue with variance estimation concerns the assessment of imputation error. Westat (Document 16) argues that multiple imputation requires "sophisticated computing equipment and large storage capacity for the multiple sets of imputed data", and proposes formulae for adjusting standard errors of singly-imputed data.

Schafer (Document 17) points out theoretical limitations of the latter approach compared with the general validity of multiple imputation approach. He also suggests that the practical difficulties have been overestimated. Since the analysis of multiply-imputed data sets simply involves repeated complete-data analyses with different sets of imputes substituted, "sophisticated computing software" is not required of the user -- indeed the avoidance of specialized missing-data software is one of the main rationales for multiple imputation, rather than direct analysis of the incomplete data. With regard to storage

requirements, additional storage is only needed for variables that have missing values and are multiply-imputed. A simple but wasteful approach to storage with five multiple imputes is simply to replace the incomplete variable by five copies, with each set of imputations substituted. Since singly-imputed data requires an additional storage location for a missing-value flag, the number of locations for each multiply-imputed variable is increased from two to five. Since we suggest that only a subset of important variables are multiply imputed, the net increase in the size of the file would be less than a factor of two. Other arrangements of the multiply-imputed data could reduce the additional storage capacity to much lower levels, at the expense of some added data-base manipulation. We do not see this as a limitation.

7. Comparisons of Imputation Methods

In this section we provide our assessment of the test evaluations of the alternative imputation procedures conducted by Westat (WES) and Schafer / Datametrics (DMS).

7.1 Three imputation methods were developed as part of this project. WES imputed about 690 component and item-within component missing values for six MEC variables: log height (LGHT), log weight (LGWT), systolic blood pressure (D-SYS), diastolic blood pressure (D-DIAS), total serum cholesterol (TCRESULT) and HDL cholesterol (HDRESULT), using two closely related regression imputation methods (Document 5). DMS was used to multiply impute 23 variables in the adult questionnaire and MEC, for over 4000 cases involving all types of nonresponse (Documents 2 and 14), a much more extensive imputation exercise.

7.2 Since the methods applied seem to be very different, it is worth noting their similarities. Both essentially apply linear regression models to predict means of missing variables, and then add noise to create a draw from the predictive distribution of the missing values. WES adds noise in the form of residuals or values from matched cases, whereas DMS in essence adds normal deviates.

7.3 WES fills in variables sequentially, which is appropriate for data with a monotone missing-data pattern. However, DMS works for any missing-data pattern. Although the missing-data pattern from missing entire questionnaires and entire examinations is close to monotone, questionnaire nonresponse for particular items, and component and item-within-component nonresponse in the MEC (which is the type of nonresponse to which WES is applied) does not have a monotone pattern, and thus ad-hoc fixes are needed to deal with multivariate imputation under the WES approaches.

7.4 In his implementation of DMS, Schafer included the survey design variable, STAND, as a predictor in his models, which was omitted in the WES models. This inclusion is in principle necessary for valid inferences from the design-based perspective.

7.5 DMS adds normal noise, and hence may be somewhat more sensitive to deviations from normality for some estimands. However, scatter plots of observed and imputed

values (Document 15) suggest that this problem is effectively resolved by transformations of the non-normal variables. Schafer's comparisons of the MEC component and item-within-component imputations (Document 15) suggest that in broad terms the imputes from all the methods reflect the distribution of the variables quite well. However, Schafer's careful data analysis revealed a number of gross outliers in the data set (see for example Figures 1-3 in Document 14), which were not adequately handled by Westat. Removal of these outliers improved the DMS model and apparently prevented some isolated problematic imputations that surfaced in the WES results (Document 15).

7.6 DMS also provides "proper" multiple imputations that include the uncertainty from estimating the parameters of the model, and allow the added component of uncertainty from imputation to be included in inferences (see Section 6.1 below); the WES imputations are single, and (unlike DMS) do not provide simple and rigorous procedures for assessing imputation uncertainty. The effects of ignoring the added uncertainty from imputation can be serious, particularly for hypothesis tests involving vector parameters (Li, Raghunathan and Rubin 1991). Adjustment factors are proposed by Westat (Document 16), but as noted by Schafer (Document 17), these appear limited to estimates of means and have questionable theoretical properties.

7.7 We regard the DMS analysis with ten multiple imputes (MI10) as the gold standard, since it deals adequately with the multivariate missing data pattern, it reflects all sources of uncertainty, and the model, editing and estimation procedures are superior.

7.8 The MI10 results include an analysis of the fraction of missing information. Estimates tend to be highly variable because of instability of the SUDAAN estimates of within-imputation variance. However, it is clear that the percent missing information is generally low for item-within-component nonresponse (e.g. 3%), and higher for component and unit nonresponse (e.g. 25%).

7.9 Design effects vary remarkably across the multiply-imputed data sets, reflecting the instability of the SUDAAN standard errors. This problem hinders the comparisons of MI3, MI5 and MI10. However, our assessment is that $M = 5$ multiple imputes should be sufficient, particularly under the proposed IMID strategy that weights for cases missing the interview and MEC examination.

7.10 As noted in Section 3, imputation has theoretical advantages over weighting in that it uses available information about covariates in an optimal way. Gains in efficiency are difficult to assess given the noisy variance estimates, although the high R^2 values for some of the predictive models (Document 14) suggest that imputation yields good predictive power for some MEC variables.

7.11 In contrast, Westat (Document 16, Table 2) provide a table suggesting that the increase in variance from discarding cases missing the MEC is surprisingly modest, averaging about 5% over all the variables considered. The smallness of the increase seems to arise because the estimated design effects for estimates based on all persons ($n=11662$)

average about 10% more than the estimated design effects for estimates based on examined persons only ($n=8212$). This finding suggests that the information loss from discarding incomplete cases is muted by the effects of clustering. Note that such an effect, if real, would tend to be reduced for estimates for subclasses of the sample, which tend to have smaller design effects due to clustering of the sample. However, given the unreliability of the design effect estimates we think this issue needs further study before any firm conclusions can be reached.

7.12 From our perspective, DMS methods are more principled in that they flow directly from the choice of model and rigorous principles of probabilistic predictive inference (e.g. Rubin 1987, chapter 5), whereas WES methods tend to be hampered by the need for repeated ad-hoc fixes to deal with the problems of non-monotone missingness. (We ourselves have considered such fixes: see Little, 1988) Since such fixes are not needed in the DMS approach, effort can be concentrated on the details of the model (e.g. which variables to include as predictors), editing and other data analysis to monitor model fit.

7.13 Westat argues against DMS on the grounds of excessive complexity. However the computational tools underlying DMS methods are simply the standard tools of regression combined with chi-squared and normal random number generation, and (on the WES side), matching algorithms that deal appropriately with the multivariate pattern of missing data are not simple to devise and program.

The DMS method follows the theoretically most rigorous approach to multiple imputation, drawing from the posterior predictive distribution of the missing values under an appropriate model. Until very recently, this task has been extremely demanding even in relative small multivariate data sets. Much has changed recently, however, with the advent of iterative simulation methods (see for example Tanner, 1991). The task still is not necessarily easy, but the major limitation in many contexts involves practical issues of computational storage and speed, rather than theoretical breakthroughs.

The DMS method is based on iterative simulation using the Gibbs sampler and data augmentation (Schafer 1991; Tanner 1991; Tanner and Wong 1987). The underlying general location model assumes a multivariate normal distribution for missing continuous variables with a mean that is linear in observed continuous and discrete variables, and has broad applicability in many applied settings. The basic computational ideas are quite straightforward, although Schafer's software is full of intelligent options, shortcuts, and diagnostics. This software is currently available and will be fully documented in a forthcoming Chapman and Hall monograph. Recent work by C.H Liu, a current Ph.D. student of D.B. Rubin, has further increased the efficiency of the computations. Liu's speeded algorithms are working and available, although lacking the extensive documentation and diagnostic displays in Schafer's software. Both programs have been used extensively on personal computers, both in the U.S in academic and government environments and in Europe. When planning for the future, it seems unwise to be very concerned about current computational limitations; as we have seen, even in our hurried

exploratory project, we have been able to handle substantial aspects of a rather large real world problem.

We feel that ultimately the proof of the imputing is in the results. The fact is that Schafer accomplished a much larger imputation task than Westat in a similar time-frame, and at the same time provide a wealth of useful diagnostic information concerning the choice of covariates and outliers. Moreover, DMS software is more readily applicable to missing data in other data bases, since it handles a more general pattern of missing data and has the ability to handle categorical data.

8. Specific Issues Associated with Multiple Imputation

8.1 File Storage and Analysis of Multiply-Imputed Data

We propose that multiple imputation be confined (at least initially) to a set of important variables for analysis, and that missing values of unimportant variables are handled by assigning missing-value codes. We suggest that $M = 5$ multiple imputes be created for each missing value, by creating five versions of each variable, with one set of multiple imputations substituted for the missing cases, and the observed value substituted for the observed cases. This scheme is somewhat wasteful in terms of storage space, but makes analysis as simple as possible for the user. An analysis of the multiply-imputed file simply requires a complete-data analysis to be repeated five times, once with each version of the multiply-imputed variables. For example, suppose that the analysis is a regression of Y on X_1 , X_2 and X_3 , where Y and X_2 are multiply-imputed and X_1 and X_3 are fully observed (e.g. design variables). Then the five analyses consist in regressing $Y_{(m)}$ on X_1 , $X_{2(m)}$, and X_3 , for $m = 1, \dots, 5$, where $Y_{(m)}$ is the version of Y with the m^{th} set of imputes substituted, etc. The key point is that each regression (or any other analysis procedure adopted) is a complete-data analysis of the filled-in data, that is, no special software is required to allow for the fact that some data are imputed.

Large-sample inferences involve simple manipulations of the results from the complete-data analyses. In particular, for inferences about a scalar quantity Q (say the regression coefficient of X_1 in the above example), let \hat{Q}_m be the estimate of Q from the m^{th} complete-data analysis, and let U_m be the associated estimate of variance. The quantities

$$\bar{Q} = \sum_{m=1}^M \hat{Q}_m, \quad \bar{U} = \frac{1}{M} \sum_{m=1}^M U_m, \quad (1)$$

$$B = \frac{1}{M-1} \sum_{m=1}^M (\hat{Q}_m - \bar{Q})^2, \quad (2)$$

$$r = \left(1 + \frac{1}{M}\right) \frac{B}{\bar{U}}. \quad (3)$$

are computed, where M is the number of imputed data sets. A $100(1-\alpha)\%$ confidence interval for Q is then

$$\bar{Q} \pm t_{\nu(1-\alpha/2)} \sqrt{T}, \quad (4)$$

where

$$T = \bar{U} + (1 + M^{-1})B,$$

is the total variance, and $t_{\nu(1-\alpha/2)}$ is the $1-\alpha/2$ quantile of the t distribution with $\nu = (M-1)(1+1/r)^2$ degrees of freedom (Rubin 1987; Rubin and Schenker 1986).

Methods for multivariate statistics and associated variance-covariance matrices are also described in Rubin (1987). These methods usually work well and have much better coverage properties than single-imputation methods. More refined methods are described in Li, Raghunathan and Rubin (1991), and are very accurate. Methods for directly combining P-values from complete-data analyses are described in Li, Meng, Raghunathan and Rubin (1991); these methods can be inaccurate with large fractions of missing data, but should work well for the levels of nonresponse in NHANES III. Methods for working with likelihood-ratio functions are described in Meng and Rubin (1992), and are very accurate. Specific formulae for these procedures are given in the review article (Rubin and Schenker 1991) attached as Appendix 2.

It is important to realize that these multivariate analyses on a multiply-imputed data set are nearly as easy to carry out as scalar analyses. For example, consider the multivariate estimand consisting of prevalence estimates by age, sex, and race, where each complete-data analysis produces the three-way table of estimates with associated standard errors, as well as p-values for several log-linear models. Each complete-data analysis on each multiply-imputed data set produces this collection of statistics. The M tables are averaged to produce the estimate of the three-way prevalence rates, and the standard errors are combined exactly as with scalar estimates. There are several options for combining the p-values. The easiest is to directly combine them using the method of Li, Meng, Raghunathan, and Rubin (1991). The most accurate is to use the method of Meng and Rubin (1992), which uses the estimates under the log-linear models and a second pass through the M data sets to obtain precise p-values. A third option is to estimate the complete-data correlations among the prevalence rates and use the equivalently accurate method of Raghunathan and Rubin. A fourth option, being developed by Raghunathan and Rubin will be available shortly. All of these methods are computationally straightforward, the first being extremely simple. More experience in the context of particular data sets is needed to formulate reliable advice on when it is worthwhile going beyond the first method, which is no more difficult than the computation of the total variance from within and between imputation components.

Documentation of public-use data sets that include multiple imputations of some variables should include material on how to carry out these inferential procedures, together with

some worked examples. It is noteworthy that other organizations have released multiply-imputed data sets (Clogg et al. 1991; Kennickell 1991) for use by non-specialists, so this task is practically feasible.

8.2 Efficiency of Multiple Imputation

All of the imputation methods in this project impute a draw from the predictive distribution rather than a mean. This approach is strongly recommended since imputing means yields biased estimates of quantities that are estimated by nonlinear functions of the data (for example variances or percentiles). However, imputing draws does lead to some loss of efficiency relative to asymptotically efficient procedures such as maximum likelihood. Multiple imputation reduces this loss of efficiency, since the estimate of Q in the inference (4) is an average \bar{Q} over the M multiply-imputed data sets -- the loss of efficiency tends to zero as M tends to infinity.

Specifically, assuming proper imputations from a correct model, the variance of estimates from multiple imputation is increased asymptotically by the factor

$$\left(1 + \frac{\gamma}{M}\right)$$

where γ is the fraction of missing information, and is estimated in the multiple-imputation output by the quantity $r/(1+r)$ where r is given by (3). The fraction of incomplete cases to be imputed under our proposed IMID scheme is less than 20% of all cases, since individuals missing the entire questionnaire and MEC examination are weighted. The fraction of missing information is typically less than this, since the fraction of incomplete cases does not take into account information in the incomplete cases used to predict the missing values. Setting $\gamma = 0.15$, the increase in variance is 15% with single imputation ($M=1$) and 3% with multiple imputation with $M = 5$, corresponding to increases in standard error of 8% and 1.5%, respectively. In practice values of γ for the whole sample are generally smaller than 0.15, although larger values may be found in subsets of the data with high levels of nonresponse and *poor predictors* of the missing values. Thus multiple imputation provides some useful improvements in efficiency over single imputation, in addition to its advantages with respect to reflecting imputation uncertainty.

It should be noted that these losses in efficiency are relative to fully-efficient estimation methods, and not to sub-optimal fixes such as weighting class adjustments, which could be considerably less efficient even than single imputation. The loss of information seems very minor compared with the practical benefits of multiple imputation for users. Since the data can be analyzed using complete-data methods, multiple imputation takes care of the missing-data problem once and for all, eliminating differences between results that are artifacts of different ways of treating the missing values (Little and Rubin 1987, Part 1).

9. Recommendations for Current Practice and Additional Research

In summary, we think this project has demonstrated the feasibility of multiple imputation methods based on Gibbs' sampling for missing data in NHANES 3. Public-use data bases that are multiply imputed have been released by other agencies. The DMS methods are state-of-the-art, and if implemented would demonstrate the agency's commitment to high-quality and rigorous statistical methodology in this area. Our recommendations are as follows:

9.1 We propose initially the IMID option, where cases missing the entire questionnaire and MEC examination are weighted, using the first-stage weight developed in Ezzati and Khare (1992). Furthermore, multiple imputation would be confined to a subset of important variables, with missing values of other variables left as missing-value codes.

9.2 Multiply-imputed data bases should include in the documentation some summary of methods of analysis for multiply-imputed data, including examples of how to create and analyze repeated complete data sets.

9.3 In future, multiple imputation methodology can be expanded to more variables and more incomplete cases, as the method becomes established with users, documentation is refined, and computing limits are reduced by advances in statistical computing.

9.4 Ways should be considered of user-friendly packaging of multiple imputations in data bases, e.g. for SAS, SPSS, BMDP etc. For example, consider whether it is better to a) use pointers to the multiple imputes, or b) repeat each incomplete variable M times, with consequent duplication of complete information.

9.5 Future research on the first-stage weighting adjustment developed in Ezzati and Khare (1992) should consider the development of weights based on the estimated response propensity (Rosenbaum and Rubin 1983; Little 1986), and assess the importance of the component of variability missed by the current procedure.

9.6 More generally research is needed on improving the estimation of complete-data standard errors. Modeling to reduce the excessive variability of current SUDAAN estimates is needed.

9.7 Studies should be conducted concerning the reasons of nonresponse in the NHANES, and to compare respondents and nonrespondents with respect to observed characteristics. Is the missing-data mechanism missing completely at random, or do respondents and nonrespondents differ? The tables created for the current project and the work on weighting adjustments in Ezzati and Khare (1992) give some indication that observed differences between respondents and nonrespondents are minor, but more systematic assessments are needed.

9.8 Finally, projects such as the one we are assessing here demonstrate feasibility, and provide useful descriptive and comparative information, but they do not provide objective information about the operating characteristics of procedures - for example, does the multiple-imputation interval (4), as implemented in DMS, really cover the quantity Q in a proportion $(1-\alpha)$ of repeated samples for NHANES, and similarly for the WES proposals? To convince skeptics, what is needed is an honest frequentist evaluation via simulation to assess the methods discussed here. That is, create hypothetical populations, draw samples from each, apply existing and competing alternative procedures to the samples, and thereby assess the validity of inference about population quantities. The assessment should distinguish between limitations due to complete-data methodology, such as SUDAAN standard errors, and those due to alternative nonresponse adjustment methods such as weighting or multiple imputation.

Although such a project requires a major effort, methods for drawing samples as in the NHANES and deriving inferences from the samples are already available. The primary new task involves design and implementation of the software needed to create hypothetical populations that provide useful information for the NHANES setting. Also, software is needed to create missing values in a realistic manner. Software to assess the performance of sample inferences appears relatively easy to develop.

References

- Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B. and Weidman, L., "Multiple Imputation of Industry and Occupational Codes in Census Public-Use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, 86, 68-78.
- Czajka, J.L., Hirabayashi, S.M., Little, R.J.A. and Rubin, D.B. (1992). "Projecting from advance data using propensity modeling; an application to income and tax statistics", *Journal of Business and Economic Statistics*, 10, 117-132.
- Ezzati, T. and Khare, M. (1992), "Nonresponse adjustments in a National Health Survey", *Proceedings of the Survey Research Methods Section, American Statistical Association 1992*.
- Fay, R.E. (1991), "A design-based perspective on missing data variance," *1991 Annual Research Conference Proceedings*, Bureau of the Census, 429-440, Washington DC: U.S. Department of Commerce.
- Fay, R.E. (1992), "When are inferences from multiple imputation valid?", *Proceedings of the Survey Research Methods Section, American Statistical Association 1992*.
- Goksel, H., Judkins, D.R. and Mosher, W.D. (1991), "Nonresponse adjustments for a telephone follow-up to a national in-person survey," *Proceedings of the Survey Research Methods Section, American Statistical Association 1991*, 581-586.
- Kalton, G. and Kasprzyk, D. (1982), "Imputing for Missing Survey Responses," *Proceedings of the Survey Research Methods Section, American Statistical Association 1982*, 22-31.
- Kalton, G. and Kasprzyk, D. (1986), "The Treatment of Missing Survey Data," *Survey Methodology*, 12, 1-16.
- Kennickell, A. B. (1991), "Imputation of the 1989 Survey of Consumer Finances: stochastic relaxation and multiple imputation," *Proceedings of the Survey Research Methods Section, 1991 American Statistical Association*, 1-10.
- Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991), "Large-sample significance tests from multiply-imputed data using moment-based statistics and a reference F distribution," *Journal of the American Statistical Association*, 86, 1065-1073.
- Li, K.H., Meng, X.L., Raghunathan, T.E. and Rubin, D.B. (1991), "Significance levels from repeated P-values with multiply-imputed data," *Statistica Sinica*, 1, 65-92.
- Little, R.J.A. (1986), "Survey nonresponse adjustments," *International Statistical Review*, 54, 139-157.
- Little, R.J.A. (1988a), "Missing data in large surveys," *Journal of Business and Economic Statistics*, 6, 287-301 (with discussion).
- Little, R.J.A. (1988b), "A test of missing completely at random for multivariate data with missing values," *Journal of the American Statistical Association*, 83, 1198-1202.