THE NHANES III MULTIPLE IMPUTATION PROJECT

J.L. Schafer, T.M. Ezzati-Rice, W. Johnson, M. Khare, R.J.A. Little, and D.B. Rubin *
Joseph L. Schafer, 325 Classroom Building, University Park, PA 16802

Key words: Missing data, sample surveys.

1 INTRODUCTION

The Third National Health and Nutritional Examination Survey (NHANES III) was designed to provide statistics on the 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). Data were collected over two three-year periods, 1988–91 (Phase 1) and 1991–94 (Phase 2), with a total sample size of approximately 40,000; national estimates are produced for each three-year period and for the entire six years. NHANES III is a complex, multistage area sample with oversampling of young children (under 5), the elderly (60+) Mexican Americans and African Americans; details of the design are given by Ezzati *et al.* (1992).

The unique feature of NHANES is that data were collected through actual physical examinations of the sampled persons. This examination strategy is both a strength and a weakness. The strength lies in the wealth of scientific data that were gleaned about a wide variety of health characteristics; the full exam included detailed body measurements, blood and urine samples, a pregnancy test, measurements of blood pressure and bone density, fundus photography, a dental exam, etc. The weakness of this strategy is its cost, both in terms of operating expense and in the burden placed on survey respondents. The examinations were conducted in mobile examination centers (MEC's) staffed by medical professionals. These centers had to be continually moved throughout the country to serve each geographic cluster of sampled persons. The high cost of MEC relocation meant that relatively few primary sampling units could be used. Moreover, because of the inconvenience associated with going to the MEC and completing the exam, nonresponse rates for the examinations were understandably high; despite the monetary incentives offered to participants, a substantial proportion of the sampled persons did not show up.

*J.L. Schafer, The Pennsylvania State University; T.M. Ezzati-Rice, W. Johnson, and M. Khare, National Center for Health Statistics; R.J.A. Little, University of Michigan; D.B. Rubin, Harvard University. With the exception of J.L. Schafer, authors' names are given in alphabetical order.

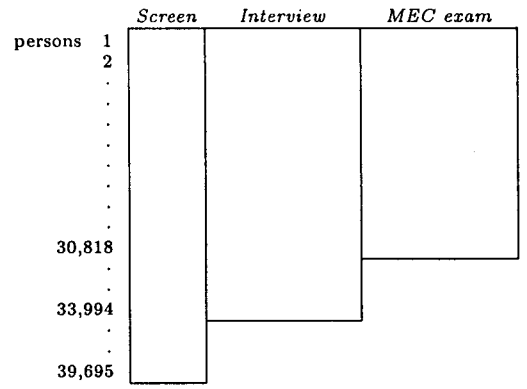

Figure 1: Unit nonresponse in NHANES III

Rates and patterns of nonresponse

The basic patterns of nonresponse in NHANES III are shown in Figure 1. The data collection occurred in three stages: (a) a household screening interview to obtain the age, sex, and race/ethnicity of each member of the sampled household, which determined the selection probabilities for each person in the final stage of sampling; (b) a personal home interview for each sampled person, with questions pertaining to health status, medical history, diet and other health-related behaviors; and (c) the physical examination in the MEC. If after repeated attempts the data collectors were unsuccessful in obtaining a screening interview, demographic information on household members was obtained from neighbors. As a result, age, sex, and race/ethnicity are known for all persons in the final dataset; there are no missing values for these variables. At the personal home interview stage, however, only 33,994 (85.6%) of the 39,695 sampled persons were successfully interviewed. None of the non-interviewed persons were scheduled for examination in the MEC. Among the interviewed persons, 30,818 (90.7%) later showed up for the MEC exam; the examination rate was thus 30,818/39,695 or 77.7% of the entire sample.

Rates of nonresponse varied appreciably by certain characteristics of the sampled persons. For illustration, the variation in rates by person's age, by race/ethnicity, and by household size is shown in Table 1.

In addition to the unit nonresponse described above, individual variables from the personal home

Table 1: NHANES III response rates by age, race/ethnicity and household size (*NIN* = not interviewed, *INM* = interviewed but no MEC exam, *MEC* = MEC exam)

	NIN	*INM*	*MEC*
Overall	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

Table 2: Overall missingness rates for select NHANES III variables

Variable	% missing
Self-rating of health status (17+)	18.8
Family income (all ages)	21.1
Body weight at exam (all ages)	21.6
Systolic blood pressure at exam (5+)	28.1
Serum cholesterol (4+)	29.4
Drusen score (40+)	41.3

interview and MEC exam exhibited varying rates of item nonresponse. Immediate reasons for item nonresponse included refusal to answer specific questions, inability to participate in certain examination procedures, examinations that were terminated early because the subject had to leave, and so on. Together, the unit and item nonresponse led to missingness rates for key survey variables of 30% or more. Overall missingness rates for a few select variables are shown in Table 2. Some variables are missing by design for certain age groups; for example, bone density measures were not taken for youths under 20. Therefore, the missingness rates in Table 2 are calculated as a percentage of the sampled persons eligible to receive the question or procedure.

Previous work

In previous NCHS health examination surveys, unit nonresponse at the personal home interview and MEC exam stages was handled by classical methods of reweighting. The nonrespondents were removed from the sample, and the respondents' sample weights—representing the reciprocals of the marginal probabilities of inclusion in the sample—were adjusted to make the total weights in reduced sample agree with those of the original sample within cells defined by age, race/ethnicity, household size, and other variables that seemed to be related to unit nonresponse. Relatively little was done to compensate for item nonresponse. NCHS staff imputed missing values for some variables on a limited basis, but these imputations were for internal agency use only; data files released to the public contained missing value codes for all item nonresponses. Decisions about how to analyze the incomplete data were essentially left to the data user.

The NHANES imputation project

In 1992, NCHS assembled a team of researchers to investigate a variety of alternatives to the current practice, including multiple imputation (Rubin, 1987). This project will culminate in Fall, 1996 with the release of multiply imputed NHANES III research data files to the public. In the remainder of this paper, we review the major elements of this effort: exploring the feasibility of multiply imputing a substantial number of examination variables (Section 2); a simulation study to assess the performance of the multiple-imputation procedure over repeated samples (Section 3); results from the simulations (Section 4); and discussion of some theoretical and practical issues surrounding the production and public release of these multiply imputed microdata files (Section 5).

2 EXPLORATORY WORK

Brief summary of 1993 paper here. Describe the imputation procedure and the analysis that led us to decide to impute missing values for all interviewed persons, but reweight for the noninterviews about whom relatively little is known.

3 SIMULATION PROCEDURES

Purpose

In 1994–1995 we conducted a simulation study to evaluate the performance of our multiple-imputation procedures. Our procedures were formulated within a Bayesian framework and were based on a probability model which was, at best,

only approximately true. The primary goal of the simulation was to evaluate the performance of the imputation procedures from purely frequentist perspective, without reference to any particular model. For example, we wanted to learn whether multiple-imputation $100(1-\alpha)\%$ interval estimates in typical applications would really cover the stated quantity $100(1-\alpha)\%$ of the time over repetitions of the sampling and imputation procedure. A secondary goal of the simulation was to compare the performance of multiple-imputation procedures to those of ad hoc methods that would be employed by the majority of data users if no multiple imputations were provided. Below we review the key elements of the simulation study and summarize the results we have obtained to date.

Constructing a population

Perhaps the most challenging aspect of the simulation was to create an artificial population from which we could draw actual NHANES-like samples. One possible approach was an adaptation of the bootstrap: Draw units with replacement from the current NHANES data to create samples of roughly the same size as the current data. We decided against this approach for a number of reasons. First, we feared that the samples would be quite unrealistic; they would exhibit excessive duplication of units, and thus would fail to adequately reflect the diversity and variablity found in realistic populations, particularly in the tails of the population distributions. Moreover, we feared that the summary statistics obtained from bootstrap resampling would not reflect realistic levels of variability over repeated samples. Another possibility was to use a computer to generate pseudorandom values from an actual multivariate probability model. This approach was also rejected, because of the obvious dangers that the population model would be too simplistic and would resemble the imputation model too closely.

In the end, we created an artificial population by pooling data from NHANES III, Phase 1 and three previous NCHS examination surveys: HANES I (1971–74), HANES II (1976–80), and HHANES (1982–84, Hispanic Americans only). We first identified a set of ten examination variables whose definitions were consistent across the four surveys: standing and sitting height, weight, systolic and diastolic blood pressure, total serum cholesterol, hemoglobin, hematocrit, iron, and total iron binding capacity. We then extracted the adults (20+) from each survey with complete data on all ten exam variables. In addition to these ten, we also retained twelve pre-exam variables: age, sex, race/ethnicity (3 levels), geographic location (13 levels), household size, marital status, years of education, poverty index, self-reports of ever having been diagnosed with diabetes and heart attack, and self-reported height and weight. These pre-exam variables—with the exception of self-reported height and weight—had a modest number of missing values. Self-reported height and weight, however, were missing for all cases from HANES I because these questions were not asked in that survey. Missing values in the pre-exam variables were imputed by hot-deck procedures that made minimal parametric assumptions (Schafer, 1994a). The end result was a population of cases more realistic than any probability model that we could have invented, because each case in the population represented a person from an actual health examination survey. Moreover, the population size (31,847 cases) was considerably larger than the samples that we drew (about 6,000 cases per sample), so the problems any problems associated with duplication of units were greatly reduced.

Weighting the population

Without further adjustments, the 31,847 cases would not have been representative of any population of real interest. To remedy this situation, we assigned the cases to 48 population strata, cross-classifying them by age (20–59, 60+) and 24 non-overlapping race/geography cells. We then reweighted the cases so that the total weight within each stratum represented a projected population count for the year 2000 obtained from the U.S. Census Bureau. Unit i in stratum h received *population weight*

$$w_i^* = N_h \frac{\pi_i^{-1}}{\sum_{j \in h} \pi_j^{-1}}$$

rounded off to the nearest integer, where N_h is the projected count and π_i^{-1} is the original weight of unit i in the survey from which it came.

Drawing samples

From this adjusted population, we drew stratified random samples to mimic essential features of NHANES III. The overall expected sample size was fixed at 6,000 to ensure that the expected number of sample cases in each stratum amounted to no more than one third of the actual number of population cases; this total of 6,000 was then divided among

the 48 strata in proportions occurring in NHANES III, Phase 1, to obtain the expected sample sizes $n_1, n_2, \ldots, n_{48}$. Each sample was drawn as follows: For $i = 1, 2, \ldots, 31,847$, case i was included in the sample m_i times, where m_i was drawn from a binomial distribution with index w_i^* and success probability n_h/N_h, $i \in h$. This method differed slightly from typical stratified sampling schemes in that the realized sample sizes $\tilde{n}_1, \tilde{n}_2, \ldots, \tilde{n}_{48}$ were random rather than fixed. After selecting a sample, we calculated *sample weights*

$$w_i = \frac{N_h}{\tilde{n}_h} \quad \text{for all } i \in h.$$

These sample weights allowed us to get unbiased estimates of population means. Consider a population mean for a survey variable Y, defined as

$$Q = \frac{\sum_i w_i^* y_i}{\sum_i w_i^*}, \tag{1}$$

where the sums are taken over the population. An unbiased complete-data estimate of Q is

$$\hat{Q} = \frac{\sum_i w_i y_i}{\sum_i w_i}, \tag{2}$$

where the sums are taken over all units in the sample. Notice that the denominator of $\hat{Q}$ does not vary, because the w_i always sum to the size of the artificial population,

$$\sum_i w_i = \sum_{h=1}^{48} \sum_{i \in h} \frac{N_h}{\tilde{n}_h} = \sum_{h=1}^{48} N_h.$$

Patterns of nonresponse

After drawing a sample, we imposed nonresponse on the sample units by a nonparametric hot-deck procedure based on data from NHANES III. The interviewed adults (20+) from NHANES III, Phase 1 were classified into a contingency table by race/geography (24 levels), sex, age (4 levels) and household size (3 levels). Any cell containing fewer than five persons was collapsed with adjacent cells. Each sampled person drawn from the artificial population was matched to an NHANES III donor selected at random from the relevant cell, and the donor's response pattern was assigned to the sampled person. The variables used to define the cells of this hot deck were completely observed in NHANES III; thus the resulting nonresponse mechanism was ignorable in the sense defined by Rubin (1987), because the probabilities of missingness depended only on quantities that were observed.

Multiple imputation

After imposing patterns of missingness, we multiply imputed missing values five times under a relatively simple version of the general location model (Schafer 1994c). The model included all ten exam variables, plus seven of the twelve pre-exam variables: age, sex, race/ethnicity, household size, geography, and self-reported height and weight. The other five pre-exam variables were deliberately left out. Some recent criticisms of multiple-imputation methodology have focused on distortions that may arise when a data analyst focuses on a relationship between an imputed variable and another covariate not included in the imputation model (e.g. Fay, 1992). By using these omitted variables in subsequent analyses, we could see whether the apparent inconsistencies between the imputation and analysis models would have any discernible effect.

Limitations

Our simulation procedure had several important limitations. First, the population and resulting samples did not reflect the geographic clustering that is present in NHANES III due to the relatively small number of primary sampling units. To capture this structure would have required additional, and probably unrealistic, modeling assumptions or data not available from previous NCHS surveys. Second, the procedure used to create missingness corresponds to a purely ignorable mechanism; the simulation provides no information on the impact of possible deviations from ignorable nonresponse. Realistic nonignorable alternatives could be specified in the future, preferably based on information from real followup interviews of nonrespondents.

4 SIMULATION RESULTS

We repeated the entire procedure—drawing a sample from the population, imposing missingness, and generating five multiple imputations—a total of 1,000 times, and proceeded to analyze the performance of multiple-imputation inferences for a variety of estimands. The basic method for multiple-imputation inference, due to Rubin (1987), proceeds as follows. Let Q be a population scalar quantity of interest. For a particular sample, let $\hat{Q}_j$ be the point estimate and $\sqrt{U_j}$ the standard error obtained from the jth imputed dataset, $j = 1, \ldots, m = 5$. Specific formulas for $\hat{Q}_j$ and U_j will depend on the nature of Q, and will be described below. The combined point estimate is sim-

ply the average of the individual estimates, $\bar{Q} = m^{-1}\sum_{j=1}^{m}\hat{Q}_j$. The uncertainty associated with $\bar{Q}$ has two components. The within-imputation component, is the average of the squared standard errors, $\bar{U} = m^{-1}\sum_{j=1}^{m}U_j$. The between-imputation component is the sample variance of the m estimated coefficients, $B = (m-1)^{-1}\sum_{j=1}^{m}(\hat{Q}_j - \bar{Q})^2$. The total variance is $T = \bar{U} + (1+m^{-1})B$. For a crude confidence interval, we can refer to the normal curve, 95% C.I. $\approx \bar{Q} \pm 1.96\sqrt{T}$. For a better approximation, we refer to a t-distribution with ν degrees of freedom, where

$$\nu = (m-1)\left(1 + \frac{m\bar{U}}{(m+1)B}\right)^2.$$

For diagnostic purposes, it is helpful to calculate two additional quantities: the relative increase in variance due to nonresponse $r = (1+m^{-1})B/\bar{U}$, and the estimated fraction of missing information

$$\gamma = \frac{r + 2/(\nu+3)}{r+1}. \qquad (3)$$

A crucial assumption underlying this procedure is that the a valid complete-data inference can be obtained with a normal approximation; that is, if $\hat{Q}$ and U represent the point and variance estimates that would be used if the data were complete, then

$$(\hat{Q} - Q)/\sqrt{U} \sim N(0,1) \qquad (4)$$

would be approximately true, so that the complete-data interval $\hat{Q} \pm 1.96\sqrt{U}$ would have approximately 95% coverage.

Results for means and proportions

If Q represents a population mean (1), then (2) is the natural complete-data estimate. A variance estimate for (2) appropriate in stratified samples is

$$U = \sum_{h=1}^{48}\left[\frac{N_h}{N}\right]^2 \frac{S_h^2}{\tilde{n}_h}, \qquad (5)$$

where $N = \sum_h N_h$ is the size of the artificial population, and

$$S_h^2 = \frac{1}{\tilde{n}_h - 1}\sum_{i\in h}(y_i - \bar{y}_h)^2$$

is the ordinary sample variance calculated from the sample units within stratum h (e.g. Cochran, 1977). This method applies to a population proportion as well, simply by defining a binary survey variable taking values 0 or 1.

This method can also be extended to a subdomain mean, the average of a survey variable for a subset of the population. Define

$$x_i' = \begin{cases} 1 & \text{if unit } i \text{ is in the subdomain,} \\ 0 & \text{otherwise} \end{cases}$$

and $y_i' = x_i' y_i$. The subdomain mean for the population is

$$Q = \frac{\sum_i w_i^* y_i'}{\sum_i w_i^* x_i'},$$

where the sums are taken over the population. An approximately unbiased estimate of Q is

$$\hat{Q} = \frac{\sum_i w_i y_i'}{\sum_i w_i x_i'} = \frac{\hat{Y}}{\hat{X}},$$

where $\hat{Y} = \sum_i w_i y_i'$, $\hat{X} = \sum_i w_i x_i'$, and the sums are taken over all units in the sample. This estimate is only approximately unbiased because the denominator is random. An appropriate Taylor-linearized variance estimate is

$$U = \frac{1}{\hat{X}^2}\hat{V}(\hat{Y}) + \frac{\hat{Y}^2}{\hat{X}^4}\hat{V}(\hat{X}) - 2\frac{\hat{Y}}{\hat{X}^3}\hat{\mathrm{Cov}}(\hat{Y},\hat{X}),$$

where

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{48} N_h^2\left[\frac{1}{\tilde{n}_h(\tilde{n}_h-1)}\sum_{i\in h}(y_i' - \bar{y}_h')^2\right],$$

and similarly for $\hat{V}(\hat{X})$ and $\hat{\mathrm{Cov}}(\hat{Y},\hat{X})$.

We evaluated nominal 95% interval estimates for the means of the ten exam variables, along with the proportions of persons falling into six categories according to standard NCHS definitions: hypertensive, high cholesterol, underweight, overweight, severely overweight, and anemic. This was done for the entire population, within 3 categories of race/ethnicity, and within 24 cells of age by race/ethnicity by sex, for a total of $16\times(1+3+24) = 448$ means. In 1,000 independent repetitions, the average coverage of an actual 95% interval is 950 with a standard deviation of $\sqrt{1000(.95)(.05)} = 6.9$. The average coverage of our multiple-imputation intervals over all 448 estimands was 949.3, not significantly different from 950.

The coverages tended to vary more in the domains with smaller sample sizes; a plot of coverage by the average sample size is shown in Figure 2. This suggests that departures from 95% coverage for some estimands could be due in large part to failure of the normal approximation for the complete data (4), not a shortcoming of multiple imputation itself. It is useful to compare the performance of the multiple-imputation intervals not only

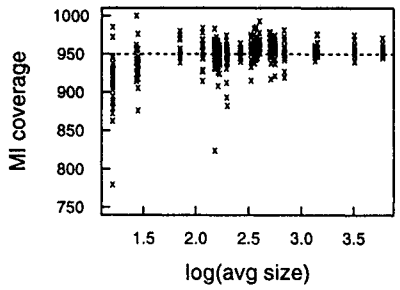

Figure 2: Coverage of multiple-imputation (MI) interval estimates for 448 means by logarithm (base 10) of the average domain sample size

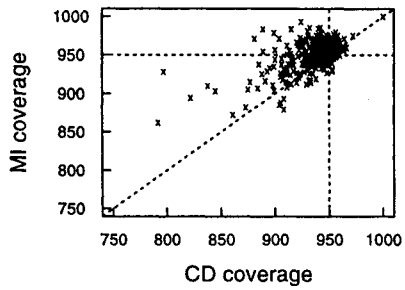

Figure 3: Coverage of complete-data (CD) versus multiple-imputation (MI) interval estimates for 448 means, with points (507,824), (608,799), and (479,876) not shown

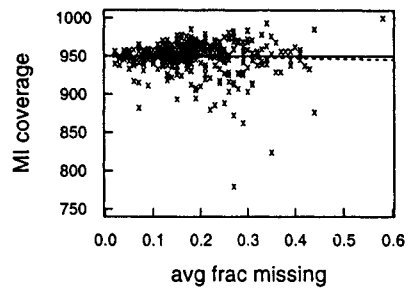

Figure 4: Coverage of multiple-imputation (MI) interval estimates for 448 means by average fraction of missing information

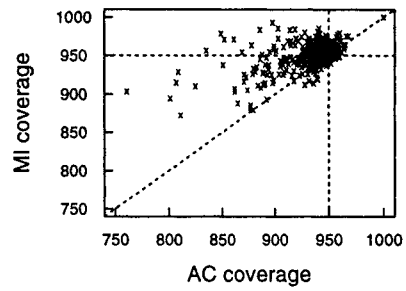

Figure 5: Coverage of available-case (AC) versus multiple-imputation (MI) interval estimates for 448 means, with points (443,824), (703,928), (530,779), (579,862), and (383,876) not shown

to the baseline of 95%, but to the actual coverage of the normal-based intervals $\hat{Q} \pm 1.96\sqrt{U}$ that one would have used if no data were missing. A plot of the coverages of the multiple-imputation (MI) intervals versus their complete-data counterparts (CD) is shown in Figure 3. The two coverages are positively correlated; the performance of MI is indeed tied to the performance of CD. Somewhat surprisingly, among the estimands for which the CD intervals exhibited gross undercoverage—and especially for the three pathological cases that fell outside the plotting region—the MI intervals performed substantially *better* than their CD counterparts. It seems that the process of deleting some observations, and then imputing them under the general location model, actually made the sampling distribution of the resulting point estimates more nearly normal. On the other hand, there were no estimands for which CD did well but MI did poorly.

With multiple imputation, we expect in general that as the fraction of missing information (3) for an estimand increases, the number of imputations needed for the MI procedure to perform well should also go up. In this simulation, however, the performance of the MI intervals based on five imputations did not appear to deteriorate with increasing missing information. A plot of MI coverage by average fraction of missing information is shown in Figure 4. The least squares fit (dashed line) is nearly indistinguishable from a horizontal line through 950 (solid).

Finally, we also compared MI to what secondary NHANES data users are likely do if no imputations are provided. This procedure, which we call the available-case (AC) method, is to simply omit the nonrespondents and calculate weighted averages and proportions from the cases that remain. The coverage of MI and AC is compared in Figure 5. The coverages are again positively correlated, with MI doing substantially better when AC exhibits gross undercoverage; there are no cases for which AC does well but MI does not. For the five cases that fell outside the plotting region, MI was very helpful in correcting the disastrous performance of AC. The coverage of AC averaged 928.4 with a standard deviation of 51.81, which is significantly below

950 ($z = -8.82$) if the 448 means are regarded as a sample from a population of estimands.

Results for quantiles

In addition to means and proportions, we investigated the properties of MI intervals for quantiles. From a medical and public-health standpoint, quantiles are important for establishing ranges considered to be normal and abnormal. From a statistical point of view, they are interesting because of their nonlinear functional form.

To obtain point estimates and standard errors for quantiles, we applied a variation of an approximate method described by Woodruff (1952). Let $F(y)$ denote the population cumulative distribution function for a continuous survey variable Y, so that $Q = F^{-1}(p)$ is the pth quantile. Woodruff's method is based on the fact that $Q_1 \leq Q \leq Q_2$ if and only if $F(Q_1) \leq p \leq F(Q_2)$, because F and F^{-1} are strictly increasing. Rather than finding a confidence interval for the Q directly, we found an interval (p_1, p_2) for the proportion of the population below the sample estimate $\hat{Q}$, and took the estimated p_1th and p_2th quantiles as the endpoints of the confidence interval for Q. Then, appealing to the large-sample normal approximation for quantiles (Francisco and Fuller, 1991), we used $\sqrt{U} = (Q_2 - Q_1)/4$ as a complete-data standard error, where Q_1 and Q_2 are the p_1th and p_2th sample quantiles.

To calculate a sample quantile, we ordered the sample values $y_i, \ldots, y_n$ from smallest to largest, carrying along their respective sample weights w_i. Denote the ordered values by $y_{(i)}$, and their respective sample weights by $w_{(i)}$. Then $\hat{Q}$, the estimated pth quantile of Y, was taken to be $y_{(j)}$, where j was the greatest integer such that

$$\frac{\sum_{i=1}^{j} w_{(i)}}{\sum_{i=1}^{n} w_{(i)}} \leq p.$$

The true quantile Q was calculated in the same manner, using the values of y_i from the population and the corresponding population weights w_i^*.

Finally, to improve the properties of interval estimate (p_1, p_2), particularly in the vicinity of 0 or 1, we applied a normal approximation on the logit scale. The estimated logit is

$$\hat{\phi} = \log\left(\frac{\hat{Y} + .5}{\hat{X} - \hat{Y} + .5}\right),$$

where $\hat{Y}$ and $\hat{X}$ are the estimated number of successes and estimated total population size, respectively, in the domain of interest. A linearized estimate of the variance of $\hat{\phi}$ is

$$\begin{aligned}\hat{V}(\hat{\phi}) = &\left(\frac{\hat{Y} + .5}{\hat{X} + 1}\right)^{-2} (\hat{X} - \hat{Y} + .5)^{-2} \hat{V}(\hat{Y}) \\ &+ (\hat{X} - \hat{Y} + .5)^{-2} \hat{V}(\hat{X}) \\ &-2\left(\frac{\hat{Y} + .5}{\hat{X} + 1}\right)^{-1} (\hat{X} - \hat{Y} + .5)^{-2} \hat{\mathrm{Cov}}(\hat{Y}, \hat{X}).\end{aligned}$$

The interval estimate (p_1, p_2) was found by applying the inverse logit transformation $e^{\phi}/(1 + e^{\phi})$ to the endpoints of $\hat{\phi} \pm 2\hat{V}^{1/2}(\hat{\phi})$.

This method for quantiles assumes that the underlying distribution function F is continuous. Because of rounding, however, the variables in our artificial population were actually quite discrete; most of them took values on a relatively coarse grid. This discreteness, when combined with the occasional duplication of population units in the samples for certain strata, caused the interval estimates for quantiles to perform rather poorly for most of the variables before missingness was imposed. Two variables for which the CD intervals were fairly well behaved were body mass index (weight/ht^2) and total serum cholesterol, so we report the results for the 50, 90, and 95th percentiles of these two variables in the total population, by sex (2 classes), by age (3 classes), and by race/ethnicity (3 classes)—a total of $2 \times 3 \times (1 + 2 + 3 + 3) = 54$ quantiles.

Plots of the coverage rates for MI showed patterns similar to those of Figures 2–5; the only notable difference was a slight tendency for the MI coverage to *increase* with the fraction of missing information; the MI intervals based on five imputations appear to become more conservative as rates of missing information rise. Summaries of the performance of CD, AC, and MI averaged over the quantiles are shown in Table 8.3 and 8.4 for body mass index and total cholesterol, respectively. These tables report the bias of the point estimate as a percentage of the standard error for CD; the average standard error as a percentage of the standard error for CD; and the coverage of the 95% intervals. The final row of each table gives the average percentage of cases deleted in the AC method, and the average fraction of missing information for the MI method. The results for MI are very encouraging. The MI point estimates are less biased than those for AC. The MI intervals are on average 8–9% narrower than those for AC, yet have higher coverage probabilities. The CD and AC intervals show some undercoverage, whereas the coverage of the MI intervals equals or exceeds the nominal rate.

Table 3: Average performance of CD, AC, and MI for 27 quantiles of body mass index

	CD	AC	MI
bias (% of CD se)	−5.76	−5.39	−1.27
standard error (% of CD)	100	109	99
coverage	935.9	934.6	950.2
missing (%)	0	16.7	12.6

Table 4: Average performance of CD, AC, and MI for 27 quantiles of total serum cholesterol

	CD	AC	MI
bias (% of CD se)	−10.1	−16.3	−8.75
standard error (% of CD)	100	109	101
coverage	919.9	917.8	960.4
missing (%)	0	16.0	19.6

Results for regression coefficients

Finally, we are now able to report some limited results for coefficients in a logistic regression model. Our goal was to assess the performance of MI for some complex estimands that may be of interest to secondary data users. We also wanted to perform a realistic analysis involving variables not used the imputatation procedure, to see whether inconsistencies between the imputer's and analyst's model would create difficulties for MI. For the complete-data procedure, we adopted the weighted point estimate and linearized variance estimate for stratified samples used by SUDAAN, a popular commercial software package for the analysis of survey data; details of the procedure are given by Shah *et al.* (1993).

We decided to focus on an effect of possible scientific interest: the odds ratio relating ever having been diagnosed with diabetes to an indicator for hypertension derived from the exam blood-pressure readings. Persons having been diagnosed with diabetes tend to be very different from those who have not in terms of basic demographic variables (e.g. age), so a simple comparison of blood pressure by diabetes status may not accurately reflect the underlying relationship; we need to control or adjust for these demographic differences. Therefore, we fit a logistic model for diabetes status that included indicator variables to distinguish among 24 cells of age by sex by race/ethnicity, plus a main effect for hypertension. Note that diabetes status had been used in the imputation procedure.

Results for the coefficient are shown in Table 5. The average rate of missing information was 17%.

Table 5: Average performance of CD, AC, and MI for logistic regression coefficient relating hypertension to diabetes (population coefficient .3691)

	CD	AC	MI
Avg. estimate	.3722	.3750	.3091
Avg. standard error	.1645	.1832	.1631
Coverage	948	938	941

In this example, neither AC nor MI does badly. The point estimate for MI appears to be biased toward zero, which was anticipated because the response variable did not appear in the imputation model; the bias is not large, however, in comparison to the average standard errors. Despite the apparent bias, the intervals from MI are still slightly superior to those from AC; the MI intervals are narrower, yet have higher coverage. These findings are consistent with the theoretical results of Meng (1995) and Rubin (1996), who discuss the possibility of superefficient MI estimates that perform better than any procedure based on the observed data alone, because the imputer has an opportunity to introduce extra information through an intelligent specification of the imputation model. Further results for regression coefficients will be available in the near future.

5 MULTIPLY IMPUTED DATA FILES FOR PUBLIC RELEASE

Production

The combined data from both phases of NHANES III have recently become available, and we are now in the process of creating a multiply-imputed data file for public release. This file, intended for reserach purposes, will contain five imputations of more than 60 basic variables for the 33,994 interviewed persons; the 5,071 noninterview cases will be handled by a reweighting procedure as described by Ezzati and Khare (1992). The imputation method will quite similar to that described in Section 2, but with two important exceptions.

The first major difference is that we are now imputing missing values for the entire NHANES III sample; all age groups are now present. It would be difficult to describe the data for all ages by a single general location model, because some of the assumptions—in particular, the homogeneity of covariance structures across demographic cells—will be seriously violated. It is unreasonable to believe,

for example, that body measurements for infants could be modeled with the same residual variances as for adults, because the magnitude of the measurements is so different. Moreover, the variables that are collected in the NHANES III home interview and exam vary markedly by age; many components are missing by design for certain age categories. To simplify matters, we are splitting the dataset by age and modeling each age group separately.

A second major difference between the new imputation methods and the procedure described in Section 2 is that a larger number of variables is now involved. For example, among the body measurements we are now imputing variables such as sitting height, head circumference, replicate measurements of skin folds, etc. As a result, in some age groups it is not possible to model all the variables simultaneously; our computing power is still too limited. Moreover, the data themselves appear too limited to support a single joint model for all variables at once.

To overcome the difficulties inherent with a single joint model for all variables, we are adopting a two-stage imputation procedure. In the first stage, we impute lower-dimensional summaries of each of the major examination components: body measurements, blood pressure, lipids, etc. These summaries include variables that are known to be strongly related to data from other components (e.g. height and weight), and/or composite measures that capture essential features of the component variables (e.g. the first few principal components). The summary measures for all components are modeled and imputed jointly under a general location model that includes demographic and geographic indicators. In the second stage, the remaining variables withinin each component are imputed conditionally given demographic and geographic indicators and the imputed values from the first stage. The second-stage models for each component are assumed to be conditionally independent of one another. In this way, we are attempting to preserve the most important aspects of the joint distribution of variables within each component, and the most important aspects of the relationships among the components.

Documentation

The public release will be accompanied by documentation. One main purpose of this documentation is to educate users about how to analyze a multiply imputed dataset. The booklet will include simple examples of analyses with results that can be easily reproduced, so that users can gain confidence that they are implementing the procedures correctly. A second, equally important, purpose of the documentation is to inform users of the assumptions underlying the imputation method. No single set of multiple imputations should be expected to work for all analyses; the best we can hope for is to impute under a model that is general enough for a variety of possible uses, and make users aware of the model's shortcomings so that they will be less likely to use the imputations for a purpose for which they were never intended.

Conclusions

In this application of multiple imputation, we are imputing under a model that is, at best, only approximately true. Yet a growing body of evidence suggests that these model-based imputations are more than adequate to provide valid inferences for a variety of statistical analyses. The fact that the modeling assumptions are being applied not to the entire dataset, but only to its missing part, is one reason why we should expect the method to be quite robust to departures from the imputation model. By releasing multiply-imputed data files, NCHS is providing a valuable service to secondary data users who might otherwise lack the resources to implement statistically sound missing-data procedures on their own.

6 REFERENCES

Cochran, W.J. (1977) *Sampling Techniques*, Second edition, New York: Wiley.

Ezzati, T. and Khare, M. (1992) Nonresponse adjustments in a national health survey. *Proceedings of the Social Statistics Section*, American Statistical Association, 203–208.

Ezzati, T., Massey, J., Waksberg, J., Chu, A. and Maurer, K. (1992) Sample design: Third National Health and Nutrition Examination Survey. *Vital Health Statistics*, Series 2, No. 113, National Center for Health Statistics.

Ezzati-Rice, T., Johnson, W., Khare, M., Little, R., Rubin, D. 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*, Bureau of the Census, 257–266.

Fay, R.E. (1992) When are inferences from multiple imputation valid? *Proceedings of the Survey Re-*

search Methods Section, American Statistical Association, 227–232.

Francisco, C.A. and Fuller, W. (1991) Quantile estimation with a complex survey design. *Annals of Statistics*, **19**, 454–469.

Meng, X.L. (1995) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, **10**, 538–573.

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

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

Schafer, J.L. (1994a) Imputation of missing interview variables prior to the MI simulation study. Unpublished memorandum to National Center for Health Statistics, 7/19/94.

Schafer, J.L. (1994b) Recommendations on sampling strategy for multiple-imputation simulation study. Unpublished memorandum to National Center for Health Statistics, 7/5/94.

Schafer, J.L. (1994c) Imputation model and procedures for the MI simulation study. Unpublished memorandum to National Center for Health Statistics, 7/21/94.

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*, Bureau of the Census, 459–487.

Shah, B.V., Folsom, R.E., LaVange, L.M., Wheeless, S.C., Boyle, K.E., and Williams, R.L. (1993) Statistical methods and mathematical algorithms used in SUDAAN. North Carolina: Research Triangle Institute.

Woodruff, R.S. (1952) Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, **47**, 653–646.