Multiple Imputation of Missing Income Data in the National Health Interview Survey

Nathaniel SCHENKER, Trivellore E. RAGHUNATHAN, Pei-Lu CHIU, Diane M. MAKUC, Guangyu ZHANG, and Alan J. COHEN

The National Health Interview Survey (NHIS) provides a rich source of data for studying relationships between income and health and for monitoring health and health care for persons at different income levels. However, the nonresponse rates are high for two key items, total family income in the previous calendar year and personal earnings from employment in the previous calendar year. To handle the missing data on family income and personal earnings in the NHIS, multiple imputation of these items, along with employment status and ratio of family income to the federal poverty threshold (derived from the imputed values of family income), has been performed for the survey years 1997–2004. (There are plans to continue this work for years beyond 2004 as well.) Files of the imputed values, as well as documentation, are available at the NHIS website (*http://www.cdc.gov/nchs/nhis.htm*). This article describes the approach used in the multiple-imputation project and evaluates the methods through analyses of the multiply imputed data. The analyses suggest that imputation corrects for biases that occur in estimates based on the data without imputation, and that multiple imputation results in gains in efficiency as well.

KEY WORDS: Health insurance; Health status; Missing data; Poverty; Public-use data; Sequential regression multivariate imputation.

1. INTRODUCTION

The National Health Interview Survey (NHIS) is a multipurpose health survey that is the principal source of information on the health of the civilian, noninstitutionalized household population of the United States (National Center for Health Statistics 2005). It is conducted by the Bureau of the Census for the National Center for Health Statistics of the Centers for Disease Control and Prevention, and it includes approximately 40,000 households containing approximately 100,000 people each year. The survey provides a rich source of data for studying relationships between income and health and for monitoring health and health care for persons at different income levels. There is particular interest in the health of vulnerable populations, such as those with low income, as well as these persons' access to and use of health care. However, the nonresponse rates are high for two key items—total family income and personal earnings from employment—both referring to the previous calendar year.

To handle the missing data on family income and personal earnings in the NHIS, multiple imputation of these items has been performed. This article describes the approach used to create the multiple imputations and evaluates the methods through analyses of the multiply imputed data. The remainder of Section 1 provides further information on the questions on income in the NHIS, the missing data, and the products of the imputation project. Section 2 discusses the imputation procedure used. Section 3 gives examples in which multiply imputed data from the 2001 NHIS are analyzed. The multiple-imputation analyses are compared with those based on no imputation and on single imputation, as well as with analyses using information from the Annual Social and Economic Supplement to the Current Population Survey (CPS), a monthly survey of households conducted by the Bureau of the Census for the Bureau of Labor Statistics. Section 4 contains a concluding discussion.

1.1 Income Items in the NHIS

The NHIS is conducted through personal household interviews. The questionnaire, which underwent a major revision in 1997, consists of a basic module as well as various supplements. The basic module, which remains largely unchanged from year to year, consists of three components: the family core, the sample adult core, and the sample child core.

The family core component, which contains the questions on family income and personal earnings, collects information on every member of a family and includes sections on family relationships, health status and activity limitations, injuries, health care access and utilization, health insurance, sociodemographic background, and income and assets. All members of the household age 17 years and older who are at home at the time of the interview are invited to participate and to respond for themselves. For those under age 17 and those not at home during the interview, information is provided by a knowledgeable adult (age 18 and older) family member residing in the household.

The sociodemographic background section of the family core component includes the following question on personal earnings for each adult who had at least one job or business: "What is your best estimate of {your/subject name's} earnings {including hourly wages, salaries, tips and commissions} before taxes and deductions from all jobs and businesses in {last calendar year}?" The response is not taken into account in the subsequent section (income and assets).

In the section on income and assets, the respondent is first asked whether any family members of any age (and if so, who) received income from each of several different sources. The respondent is then asked about total combined family income for

Nathaniel Schenker is Senior Scientist for Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782 (E-mail: *nschenker@cdc.gov*). Trivellore E. Raghunathan is Professor of Biostatistics, School of Public Health and Research Professor, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106 (E-mail: *teraghu@umich.edu*). Pei-Lu Chiu is Survey Statistician, Division of Health Interview Statistics, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782 (E-mail: *pchiu@cdc.gov*). Diane M. Makuc is Associate Director for Science, Office of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782 (E-mail: *dmakuc@cdc.gov*). Guangyu Zhang is a doctoral student, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48106 (E-mail: *guangyuz@umich.edu*). Alan J. Cohen is Statistician, Office of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782 (E-mail: *acohen@cdc.gov*). The work of Raghunathan and Zhang was supported by a cooperative agreement with the National Center for Health Statistics (DHHS/CDC/NCHS/UR6/CCU51748) and a grant from the National Science Foundation (SES-0106914).

Journal of the American Statistical Association
September 2006, Vol. 101, No. 475, Applications and Case Studies
DOI 10.1198/016214505000001375

all family members including children as follows: "Now I am going to ask about the total combined income of your family in {last calendar year}, including income from all sources we have just talked about, such as wages, salaries, Social Security or retirement benefits, help from relatives, and so forth. Can you tell me that amount before taxes?" If the respondent does not provide the amount, then the following question is asked: "You may not be able to give us an exact figure for your total combined family income, but can you tell me if your income was \$20,000 or more or less than \$20,000?" If one of these two income groups is specified by the respondent, a card is shown to the respondent with the goal of placing the income into 1 of 44 detailed income categories, and the respondent is asked which category best represents the total combined family income. Note that the total combined income of all family members is estimated by the respondent. An estimate is not obtained by summing responses to more detailed questions, as is done in some surveys that include more extensive questions on income, such as the Annual Social and Economic Supplement to the CPS.

1.2 Missing Data on Income in the NHIS

For the 8 years 1997–2004, the respective weighted percentages of families with unknown family incomes were: 24, 29, 31, 32, 32, 32, 33, and 33 for the "exact" value; 20, 25, 28, 29, 29, 29, 31, and 29 for the 44-category value; and 6, 8, 8, 8, 9, 9, 10, and 11 for the 2-category ($\geq$\$20,000 or $<$\$20,000) value. The respective weighted percentages of employed adults with unknown personal earnings were 24, 30, 32, 33, 31, 30, 33, and 31. (The weighted missing-data rates given in this paragraph are all close to their unweighted counterparts.) Missing-data rates for most other variables in the NHIS are very low.

In addition to the high rates of missing income data, there is evidence that the missingness is related to several person-level and family-level characteristics, including items pertaining to health. For example, Table 1 displays the results of fitting a logistic regression for persons age <65, with an indicator variable for nonresponse on both the exact and 44-category values of family income as the outcome, and selected variables as predictors. Statistically significant predictors of nonresponse (as indicated by 95% confidence intervals for odds ratios that exclude 1) include variables for not having health insurance, having activity limitations, age, race, being born outside of the United States, and region of residence. Variance estimates underlying the confidence intervals in Table 1, and those used in computing estimated standard errors presented elsewhere in this article, account for weighting, stratification, and clustering through the survey estimation procedures available in Stata, release 8.0 (Stata Corporation 2003). The variance estimates are based on variability at the primary sampling unit level, and they are generally either approximately unbiased or conservative (Stata Corporation 2003, sec. 30.2.1).

Because missingness of income data is related to several important characteristics, the respondents generally cannot be treated as a random subset of the original sample. It follows that the most common method for handling missing data in software packages, complete-case analysis (Little and Rubin 2002, sec. 3.2), which deletes cases that are missing any of the variables involved in the analysis, will often be biased. Moreover, complete-case analysis discards some of the observed data and thus is also generally inefficient relative to methods using all of the observed data.

Table 1. Results From a Logistic Regression, With an Indicator Variable for Nonresponse on Both the Exact and 44-Category Values of Family Income as the Outcome and Selected Variables as Predictors, for Persons Age <65, 2001 NHIS

Variable	*Odds ratio*	*95% confidence interval*
Has health insurance?		
No	1.54	(1.43, 1.66)
Yes (reference)		
Has limitations of activities?		
Yes	.77	(.72, .83)
No (reference)		
Age (years)		
<18	.64	(.60, .69)
18–24	.69	(.63, .76)
25–34	.58	(.53, .62)
35–44	.70	(.64, .76)
45–54	.82	(.76, .88)
55–64 (reference)		
Gender		
Male	.98	(.96, 1.01)
Female (reference)		
Race/ethnicity		
Hispanic	1.01	(.91, 1.12)
Non-Hispanic black	1.21	(1.09, 1.34)
Non-Hispanic other	.92	(.78, 1.08)
Non-Hispanic white (reference)		
Born in the U.S.?		
No	1.14	(1.05, 1.25)
Yes (reference)		
Region of residence		
Northeast	1.05	(.90, 1.23)
South	.79	(.71, .88)
West	.94	(.84, 1.06)
Midwest (reference)		
Resides in metropolitan area?		
No	.91	(.79, 1.04)
Yes (reference)		

1.3 The NHIS Multiple-Imputation Project

To handle the missing data on family income and personal earnings in the NHIS, multiple imputation (Rubin 1987) of these items was performed for the survey years 1997–2004, with five sets of imputed values created to allow the assessment of variability due to imputation. (There are plans to continue this work for years beyond 2004 as well.) Because personal earnings were collected only for employed adults, employment status was also imputed for the small percentage (<4%) of adults for whom it was unknown. Finally, the ratio of family income to the applicable federal poverty thresholds was derived for families with missing incomes, based on the imputed values. The imputation procedure incorporated many predictors, including demographic and health-related variables (see Sec. 2.4).

For each year in the period 1997–2004, there are five data files for the NHIS multiply imputed data, one file for each set of imputed values. For each person, each file contains the values of family income, personal earnings, employment status, and the poverty ratio; flags indicating whether the value of each variable was imputed; and information for linking the data to other data from the NHIS. In the public-use version of the multiply imputed data, family income and personal earnings are given in 11 categories, and the poverty ratio is given in 14 categories. Datasets containing the imputed values, along with documentation, can be obtained from the NHIS website (*http://www.cdc.gov/nchs/nhis.htm*).

2. PROCEDURE FOR CREATING IMPUTATIONS FOR THE NHIS

A detailed description of the procedure used in the imputation project, including lists of all of the predictors in the imputation models, can be found in the technical documentation that accompanies the multiply imputed data (the direct link to the documentation is *http://www.cdc.gov/nchs/data/nhis/tecdoc.pdf*). The following four sections discuss major features of the imputation procedure. Section 2.1 summarizes complicating issues that made the imputation problem especially interesting methodologically. Section 2.2 provides an overview of the steps in the procedure. Section 2.3 outlines the sequential regression multivariate imputation (SRMI) algorithm (Raghunathan, Lepkowski, Van Hoewyk, and Solenberger 2001) that was used in each step. Finally, Section 2.4 summarizes the predictors used in the imputation models.

2.1 Complicating Issues

The imputation of family income and personal earnings in the NHIS was complicated by several issues. First, these variables are hierarchical in nature, with one reported at the family level and the other at the person level. Second, there are structural dependencies among the variables in the survey. For example, individuals can have earnings (given by one variable) only if they are employed (as indicated by other variables). Third, in some cases, the income and earnings items needed to be imputed within bounds; for example, as discussed in Sections 1.1 and 1.2, some families did not report exact income values but did report income categories, which were used to form bounds for exact income. As another example, an intermediate step in the imputation procedure, as discussed in Section 2.2, was to impute "family earnings," that is, the total of personal earnings within a family. When personal earnings within a family were reported for some employed adults but missing for others, the sum of the reported personal earnings was used as a lower bound for family earnings. Finally, several variables of various types (categorical, continuous, count) were used as predictors in the imputation procedure, and they often required a small amount of imputation themselves.

2.2 Steps in the Imputation Procedure

To handle the hierarchical nature of family income and personal earnings, it was decided to first impute the missing values of family income and the missing values of family earnings, the latter of which occurred for families that included employed adults with unknown personal earnings. Once these family-level items were imputed, missing values of personal earnings within each family were imputed through imputation of the proportion of family earnings to be allocated to those family members with missing personal earnings.

Family income and family earnings were imputed first because other variables were expected to be especially useful in predicting these items. For example, as described in Section 1.2, although exact family income was not reported for 24–33% of the families, an income category was available for most of these families. In addition, some families with missing values of family income had information available on family earnings and vice versa, and these two variables were expected to be highly correlated. Finally, the (log) mean and (log) standard deviation of reported family incomes were calculated by secondary sampling unit (SSU), and these contextual variables were included as predictors in the regressions used in the imputation procedure (see Secs. 2.3.1 and 2.4). (The SSUs in the NHIS were small clusters of housing units.)

In the imputation of family income, family earnings, and personal earnings, several covariates were used. The family-level covariates were primarily summaries of the person-level covariates within each family. Most of the person-level covariates had very low rates of missingness. To facilitate their use, their missing values, along with missing values of employment status, were imputed for adults (because employment and earnings items, as well as many of the person-level covariates, apply only to adults in the NHIS) before the imputation of family income and family earnings. Any remaining missing values in the family-level covariates, due primarily to missingness in person-level covariates for children, were imputed together with family income and family earnings.

To summarize, the sequence of steps in the imputation procedure was as follows:

1. Impute missing values of person-level covariates and employment status for adults.
2. Create family-level covariates.
3. Impute missing values of family income and family earnings, as well as any missing values of family-level covariates (due primarily to missing person-level covariates for children).
4. Impute the proportion of family earnings to be allocated to each employed adult with missing personal earnings, and calculate the resulting personal earnings.

In the initial imputation of variables in step 1, income and earnings items were not used as predictors. To fully incorporate any relationships between income and earnings items and the person-level covariates imputed in step 1, the procedure cycled through steps 1–4 five more times, with the income and earnings items (including the imputed values) now included as predictors in step 1. (During the additional cycles, imputed values of employment status were kept constant at their initial imputed values, to avoid incompatibilities with imputed values of personal earnings.) In each of these five additional cycles, the SSU level (log) mean and (log) standard deviation of family incomes were also recalculated, with the imputed values included in the calculations.

To create multiple imputations, the entire imputation process was repeated independently five times.

Although missing values were imputed for several variables other than family income (and the poverty ratio derived from it), personal earnings, and employment status, and several new variables were created, these additional variables and imputed values were not retained in the final public-use data files.

2.3 Sequential Regression Multivariate Imputation

The imputations in each of steps 1, 3, and 4 described in Section 2.2 were created through the SRMI algorithm (Raghunathan et al. 2001) using the module IMPUTE in the software package IVEware (*http://www.isr.umich.edu/src/smp/ive*). This section provides a brief description of SRMI, discusses aspects of SRMI in the context of steps 1, 3, and 4, and compares SRMI with imputation based on a full joint model.

2.3.1 Brief Description. Let X denote the fully observed variables, and let $Y^{(1)}, Y^{(2)}, \dots, Y^{(k)}$ denote the k variables with missing values, ordered by the amount of missingness from least to most. The imputation process for $Y^{(1)}, Y^{(2)}, \dots, Y^{(k)}$ proceeds in c rounds. In the first round, the regression of $Y^{(1)}$ on X is fitted to the cases with $Y^{(1)}$ observed, and the missing values of $Y^{(1)}$ are imputed (randomly from an approximate predictive distribution based on the fitted regression). Then the regression of $Y^{(2)}$ on X and $Y^{(1)}$ (including the imputed values of $Y^{(1)}$) is fitted to the cases with $Y^{(2)}$ observed, and the missing values of $Y^{(2)}$ are imputed. Next the regression of $Y^{(3)}$ on X, $Y^{(1)}$, and $Y^{(2)}$ is fitted to the cases with $Y^{(3)}$ observed, and the missing values of $Y^{(3)}$ are imputed, and so on, until the regression of $Y^{(k)}$ on $X, Y^{(1)}, Y^{(2)}, \dots, Y^{(k-1)}$ is fitted to the cases with $Y^{(k)}$ observed, and the missing values of $Y^{(k)}$ are imputed.

In rounds 2 through c, the imputation process carried out in round 1 is repeated, except that in each regression, all variables except for the variable to be imputed are used as predictors (with their most recent imputed values included). Thus the regression of $Y^{(1)}$ on $X, Y^{(2)}, Y^{(3)}, \dots, Y^{(k)}$ is fitted to the cases with $Y^{(1)}$ observed, and the missing values of $Y^{(1)}$ are reimputed; then the regression of $Y^{(2)}$ on $X, Y^{(1)}, Y^{(3)}, \dots, Y^{(k)}$ is fitted to the cases with $Y^{(2)}$ observed, and the missing values of $Y^{(2)}$ are reimputed, and so on. After c rounds, the final imputations of the missing values of $Y^{(1)}, Y^{(2)}, \dots, Y^{(k)}$ are used.

For each regression in the SRMI procedure, IVEware allows the use of a normal linear regression model if the outcome variable is continuous, a logistic regression model if the outcome variable is binary, a multinomial logit model if the outcome variable is categorical with more than two categories, a Poisson regression model if the outcome variable is a count, and a two-stage model if the outcome variable is semicontinuous (see Raghunathan et al. 2001 for more details).

As discussed in Section 2.1, there were structural dependencies among some of the variables being imputed, and bounds needed to be placed on some variables as well. Besides allowing the use of several types of regression models, IVEware allows for restrictions to account for structural dependencies (by subsetting the data during estimation) and allows imputation within bounds (through the use of truncated distributions).

Because fitting of regressions in the SRMI procedure does not automatically account for features of the sample design, variables reflecting the design were included as predictors in the regression models, as discussed further in Section 2.4.

The idea of imputing variables sequentially using regression models dates back at least to Kennickell (1991), who used such an algorithm for multiply imputing missing values for the Survey of Consumer Finances of the Federal Reserve Board. Another software package that implements such procedures is MICE (multiple imputation by chained equations), which can be obtained, along with documentation and related reports, at *http://www.multiple-imputation.com*.

2.3.2 Some Details on Using SRMI in the Steps of the NHIS Imputation Procedure.

Imputing family income and family earnings. In the context of step 3 of the imputation procedure (described in Sec. 2.2), normal linear regression models with family as the unit of analysis were used in imputing family income and family earnings. To find a single, simple transformation for the two variables that would be consistent with the normality assumptions underlying the regression models, Box–Cox analyses (Box and Cox 1964) and examination of residual plots were performed using the complete cases from the 1997 NHIS. It was decided to use the cube root transformation, which is similar to the transformation (to the power .375) found by Paulin and Sweet (1996) to be optimal in modeling income data from the Consumer Expenditure Survey of the Bureau of Labor Statistics. The final imputed values of family income and family earnings were transformed back to the original scale.

Imputing personal earnings. In step 4 of the imputation procedure, for each family that had only one employed adult with personal earnings not reported, the person's earnings were determined, conditional on the imputed value of family earnings from step 3, by simply subtracting the total reported personal earnings for the other members of the family from the family earnings.

Again conditional on the imputed values of family earnings from step 3, let p denote the proportion of an employed adult's family earnings that were earned by the adult, and let $Y = \text{logit}(p)$. A normal linear regression model was used in imputing values of Y for all employed adults with missing personal earnings belonging to families with at least two such missing values. The imputed values of Y were then back-transformed to obtain imputed proportions. Within each family, the imputed proportions were then rescaled so that the sum of the observed proportions and the imputed proportions would be 1, and imputed personal earnings were calculated by multiplying each rescaled imputed proportion by the family earnings.

To illustrate the rescaling of imputed proportions, consider a family with four employed adults, two of whom are missing personal earnings. Let $p_{\text{obs}(1)}$ and $p_{\text{obs}(2)}$ denote the values of p for the two persons with reported personal earnings, and let $p_{\text{imp}(1)}$ and $p_{\text{imp}(2)}$ denote the values of p that were imputed (before rescaling) for the two persons with missing personal earnings. Each of the two imputed values of p was rescaled by multiplying by the constant $(1 - p_{\text{obs}(1)} - p_{\text{obs}(2)})/(p_{\text{imp}(1)} + p_{\text{imp}(2)})$.

Imputing covariates and employment status. For each covariate imputed in step 1 or step 3 and for employment status, the type of regression model used depended on the form of the variable being imputed (e.g., continuous, binary), as discussed in Section 2.3.1.

2.3.3 Comparison to Imputation Based on a Full Joint Model. Because SRMI requires only the specification of an individual regression model for each of the Y-variables, it does not necessarily imply a joint model for all of the Y-variables conditional on X. The decision to use SRMI and IVEware to create the imputations for the NHIS was influenced in large part by the complicating factors summarized in Section 2.1, specifically the structural dependencies, the bounds, and the large number of predictors of varying types that had small amounts of missing values. These complicating factors would be very difficult to handle using a method based on a full joint model. Without the complicating factors, however, the SRMI-based imputation procedure used in this project would actually be equivalent to the following two steps, corresponding to steps 3 and 4 in Section 2.2:

1. Impute the missing values of family income and family earnings based on a bivariate normal model (given predictors and transformations).
2. Impute the proportion of family earnings to be allocated to each employed adult with missing personal earnings (for each family with more than one missing value), using a normal linear regression model for the transformed proportion and rescaling of the imputed proportions, and then calculate the resulting personal earnings.

2.4 Predictors in the Imputation Models

When multiple imputations are being created, it is beneficial to include a large number of predictors in the imputation model, especially variables that will be used in subsequent analyses of the multiply imputed data (Meng 1994; Rubin 1996). In the NHIS multiple-imputation project, about 60 predictors were included in the models for person-level imputations and for family-level imputations (see the technical documentation that accompanies the multiply imputed data at *http://www.cdc.gov/nchs/data/nhis/techdoc.pdf*). The predictors included variables describing demographic characteristics, family structure, geography, education, employment status, hours worked per week, sources of income, limitations of activities, health conditions that caused limitations, overall health, use of health care, health insurance, indicators for the distinct combinations of stratum and primary sampling unit, survey weights, and SSU-level summaries of family income (as mentioned in Sec. 2.2).

The last three variables just listed were included in part to reflect the sample design of the NHIS in the imputations. As discussed by Rubin (1996), when using multiple imputation in the context of a sample survey with a complex design, it is important to include features of the design in the imputation model, so that approximately valid inferences will be obtained when the multiply imputed data are analyzed.

3. ANALYSES OF MULTIPLY IMPUTED DATA FROM THE 2001 NHIS

This section illustrates properties of the NHIS multiple imputations through analyses of relationships between health variables and other items, particularly poverty ratio, using data for the 2001 survey year. An emphasis is on comparing estimates and estimated standard errors obtained with multiple imputation to those obtained with no imputation and single imputation. Comparisons are also made with alternative analyses that involve poststratification reweighting of the complete cases from the NHIS to agree with control totals from the Annual Social and Economic Supplement to the CPS.

3.1 Summary of Methods for Analyzing Multiply Imputed Data

To begin, this section summarizes the methods for analyzing multiply imputed data that were used in the analyses of the multiply imputed NHIS data described in subsequent sections. Further details of these and other methods have been given by Rubin and Schenker (1986), Rubin (1987, secs. 3.3–3.5), Li, Meng, Raghunathan, and Rubin (1991), Li, Raghunathan, and Rubin (1991), Meng and Rubin (1992), and Barnard and Rubin (1999).

Suppose that there are m sets of imputations ($m = 5$ in the NHIS imputation project), which, when combined with the observed data, form m completed datasets, and let Q denote a scalar population quantity of interest. Application of the method of analysis that would be used with complete data to the lth completed dataset yields the point estimate $\hat{Q}_l$ and its estimated variance U_l, $l = 1, 2, \ldots, m$.

The combined multiple-imputation point estimate is $\bar{Q}_m = m^{-1}\sum_{l=1}^{m}\hat{Q}_l$. The estimated variance of this point estimate is $T_m = \bar{U}_m + [(m+1)/m]B_m$, where $\bar{U}_m = m^{-1}\sum_{l=1}^{m}U_l$ and $B_m = (m-1)^{-1}\sum_{l=1}^{m}(\hat{Q}_l - \bar{Q}_m)^2$. The quantity $\hat{\gamma}_m = [(m+1)/m]B_m/T_m$ is approximately the fraction of information about Q that is missing due to nonresponse (Rubin 1987, sec. 3.3), also known as the "fraction of missing information."

3.2 Analyses of the Relationship Between Poverty Ratio and Personal Health Status

Table 2 displays point estimates and estimated standard errors for the percentage of persons age 45–64 in fair or poor health, by category of poverty ratio, based on the NHIS with no imputation, with single imputation, and with multiple imputation. The single-imputation estimates were computed using the first set of imputations from the multiply imputed data, and thus correspond to $\hat{Q}_1$ and $U_1^{1/2}$ in the notation of Section 3.1. Table 2 also displays the approximate fractions of missing information ($\hat{\gamma}_m$) and the ratios of the estimated standard errors based on either no imputation or single imputation to those ($T_m^{1/2}$) based on multiple imputation.

Table 2. Point Estimates, Estimated Standard Errors (SEs), and Approximate Fractions of Missing Information ($\hat{\gamma}_m$) for the Percentage of Persons Age 45–64 in Fair or Poor Health, by the Ratio of Family Income to the Federal Poverty Threshold, 2001 NHIS

	No imputation (NI)		*Single imputation (SI)*		*Multiple imputation (MI)*				
Poverty ratio	*Point estimate*	*Estimated SE*	*Point estimate*	*Estimated SE*	*Point estimate*	*Estimated SE*	$\hat{\gamma}_m$	*Ratio of estimated SEs: NI/MI*	*Ratio of estimated SEs: SI/MI*
<1	45.6	1.68	39.4	1.34	39.9	1.54	.27	1.09	.87
[1, 2)	32.7	1.32	29.8	1.03	29.3	1.11	.19	1.19	.93
[2, 4)	16.1	.63	16.0	.51	15.9	.55	.05	1.15	.94
≥4	5.9	.34	6.1	.27	6.2	.30	.19	1.11	.90

Table 3. Estimated Percentage of Persons Age 45–64 in Fair or Poor Health, by Two-Category Family Income, for Reporters and Nonreporters of Exact Family Income, 2001 NHIS

Two-category family income	Exact family income reported	Exact family income not reported
<$20,000	41.6	33.5
≥$20,000	10.6	11.0

Table 5. Estimated Percentage of Persons Under Age 65 With no Health Insurance Coverage, by Two-Category Family Income, for Reporters and Nonreporters of Exact Family Income, 2001 NHIS

Two-category family income	Exact family income reported	Exact family income not reported
<$20,000	30.9	32.1
≥$20,000	11.5	15.3

Although all three analyses (no imputation, single imputation, and multiple imputation) indicate a negative association between poverty ratio and the percentage in fair or poor health, the point estimates based on single imputation and multiple imputation are very close to each other, whereas the point estimates based on no imputation (i.e., based on complete-case analysis) are quite different from those based on data with imputation for the two categories of poverty ratio indicating lower incomes. To help understand the differences between the point estimates with and without imputation, Table 3 displays estimates of the percentage of persons age 45–64 in fair or poor health, by two-category income (<$20,000 vs. ≥$20,000), both for those with an exact value of family income reported (the respondents) and those without an exact value of family income reported (the nonrespondents). (The estimates for the nonrespondents were calculated from the large proportion with categorical values reported.) For those with family incomes <$20,000, the estimated percentage in fair or poor health is substantially lower for nonrespondents than for respondents. It follows that compared with complete-case analysis, imputation should lower the estimated percentage in fair or poor health among those with low poverty ratios. The results in Tables 2 and 3 illustrate the possible biases associated with complete-case analysis that were mentioned in Section 1.2. They also illustrate the utility of incorporating information supplied by categorical reports of income into the imputation procedure.

In Table 2, the ratios of the estimated standard errors with no imputation to those with multiple imputation are all >1, suggesting that multiple imputation results in more precise point estimates than does complete-case analysis. These results need to be interpreted with caution, because the estimated standard errors for complete-case analysis can be biased when the missing data are not missing completely at random. (See Sec. 3.3 for a likely example of this phenomenon.) Nevertheless, the results conform with theory, in the sense that if the multiple-imputation analyses and complete-case analyses are both unbiased, and if there is additional information supplied through the imputation procedure, then the multiple-imputation analyses should be more precise than the complete-case analyses. The ratios of the estimated standard errors with single imputation to those with multiple imputation are all <1, reflecting the fact that the single-imputation analyses erroneously do not account for the extra uncertainty due to imputation.

The approximate fractions of missing information in Table 2, which range from 5% to 27%, are all smaller than the (weighted) nonresponse rate for poverty ratio, which is 34% for those age 45–64 with reported health statuses. In general, the fraction of missing information tends to be smaller than the nonresponse rate in practice, because the fraction of missing information accounts not only for the amount of nonresponse, but also for the information incorporated into the imputation model that is predictive of the variable subject to nonresponse. Moreover, the fraction of missing information is specific to the quantity being estimated. In Table 2, for example, the differing fractions of missing information across the categories of poverty ratio reflect the differing amounts of imputation in these categories and how well the imputation model predicts values in the different categories.

3.3 Analyses of the Relationship Between Poverty Ratio and Having Health Insurance

Table 4 displays results for the percentage of persons age <65 without health insurance, analogous to Table 2 for health status. As with health status, all three analyses indicate the same direction of association (negative) between poverty ratio and the percentage uninsured, although the point estimates from complete-case analysis are quite different than those based on the data with imputation. The largest differences occur for the two higher categories of poverty ratio. These differences are again explained, at least in part, by an analysis involving two-category income. Table 5 displays estimates of the percentage of persons age <65 who are uninsured, by two-category income, for both respondents and nonrespondents. The greatest difference between nonrespondents and respondents occurs for those with incomes of at least $20,000, with the percentage uninsured being larger among nonrespondents than among respondents. Once again, imputation appears to be correcting for

Table 4. Point Estimates, Estimated SEs, and Approximate Fractions of Missing Information ($\hat{\gamma}_m$) for the Percentage of Persons Under Age 65 With no Health Insurance Coverage, by the Ratio of Family Income to the Federal Poverty Threshold, 2001 NHIS

	No imputation (NI)		Single imputation (SI)		Multiple imputation (MI)				
Poverty ratio	Point estimate	Estimated SE	Point estimate	Estimated SE	Point estimate	Estimated SE	$\hat{\gamma}_m$	Ratio of estimated SEs: NI/MI	Ratio of estimated SEs: SI/MI
<1	31.4	1.06	32.6	.85	32.5	.85	.03	1.25	1.00
[1, 2)	28.7	.70	28.8	.57	28.7	.61	.04	1.15	.93
[2, 4)	13.0	.35	14.6	.33	14.7	.36	.12	.97	.90
≥4	4.6	.22	6.2	.23	6.1	.23	.15	.92	.98

biases that would occur if only the complete cases were analyzed, and the observed bounds on income appear to be contributing a substantial amount of information to the imputation model.

The ratios of estimated standard errors in Table 4 show the same patterns as they did for the analyses of health status (in Table 2), with two exceptions. For the two highest poverty ratio categories, the estimated standard errors from complete-case analysis are actually lower than those based on the multiply imputed data. This likely reflects the differences in the point estimates and the strong relationship of point estimates of percentages near 0 or 100 to their estimated standard errors. Because the point estimates from complete-case analysis are closer to 0 than those based on the multiply imputed data (which was shown earlier to be likely due to biases in the complete-case estimates), the estimated standard errors from complete-case analysis might be expected to be smaller, even after the extra information in the imputations is taken into account.

As was the case with health status, the approximate fractions of missing information for the analysis involving health insurance, which range from 3% to 15%, are much smaller than the simple (weighted) nonresponse rate for poverty ratio, which is 29% for those under age 65 with reported health insurance status.

3.4 Logistic Regression Analyses

Table 6 displays the results of logistic regression analyses in which the outcome is being in fair or poor health and the predictors include poverty ratio category and several other variables, under no imputation, single imputation, and multiple imputation. Table 7 displays similar analyses, for persons age <65, in which the outcome is being uninsured.

Several features of these estimates are noteworthy. First, the estimated coefficients of the indicator variables for poverty ratio category are smaller with imputation than when based on just the complete cases (no imputation). These differences are consistent with the results in Sections 3.2 and 3.3, given that the logistic regression coefficients for poverty ratio categories in Tables 6 and 7 reflect contrasts between each category and the highest category (the reference group) and the ranges of the logits of the estimated percentages when moving from the lowest to the highest category of poverty ratio in Tables 2 and 4 are smaller with imputation than without it. Note, however, that whereas the logistic regressions condition on a number of covariates, the estimates in Tables 2 and 4 are conditional only on the poverty ratio and thus could also be influenced by differences in the covariate distributions between the respondents and nonrespondents on family income.

Table 6. Point Estimates, Estimated SEs, and Approximate Fractions of Missing Information ($\hat{\gamma}_m$) for Coefficients of the Logistic Regression of Being in Fair or Poor Health on Selected Variables, 2001 NHIS

	No imputation (NI)		Single imputation (SI)		Multiple imputation (MI)			Ratio of estimated SEs: NI/MI	Ratio of estimated SEs: SI/MI
Variable	Point estimate	Estimated SE	Point estimate	Estimated SE	Point estimate	Estimated SE	$\hat{\gamma}_m$		
Constant	−2.16	.080	−1.97	.061	−1.98	.064	.100	1.26	.96
Poverty ratio									
<1.00	2.17	.067	1.90	.055	1.92	.060	.179	1.12	.92
[1, 2)	1.65	.059	1.42	.046	1.44	.051	.154	1.16	.92
[2, 4)	.94	.056	.85	.044	.85	.054	.339	1.03	.82
≥4 (reference)									
Age (years)									
<18	−3.23	.074	−3.26	.062	−3.26	.062	.008	1.19	1.00
18–24	−2.67	.096	−2.62	.078	−2.63	.078	.007	1.23	1.00
25–34	−2.03	.071	−2.11	.057	−2.11	.057	.003	1.25	1.01
35–44	−1.33	.059	−1.43	.049	−1.43	.049	.002	1.21	1.01
45–54	−.65	.055	−.76	.047	−.76	.046	.007	1.19	1.01
55–64	−.10	.055	−.23	.045	−.23	.045	.008	1.23	1.00
≥65 (reference)									
Race/ethnicity									
Hispanic	.24	.066	.30	.052	.30	.052	.002	1.27	1.00
Non-Hispanic black	.45	.059	.51	.048	.50	.049	.013	1.21	1.00
Non-Hispanic other	.21	.105	.21	.088	.21	.087	.007	1.20	1.01
Non-Hispanic white (reference)									
Born in the U.S.?									
No	−.30	.061	−.25	.049	−.25	.049	.003	1.24	.99
Yes (reference)									
Gender									
Male	.02	.030	.03	.022	.03	.022	.006	1.32	.99
Female (reference)									
Region of residence									
Northeast	−.12	.062	−.17	.054	−.16	.053	.009	1.17	1.01
South	.10	.056	.09	.047	.08	.047	.002	1.19	1.00
West	−.04	.066	−.05	.050	−.04	.050	.004	1.32	1.00
Midwest (reference)									
Resides in metropolitan area?									
Yes	−.08	.049	−.10	.039	−.10	.040	.006	1.24	.99
No (reference)									

Table 7. Point Estimates, Estimated SEs, and Approximate Fractions of Missing Information ($\hat{\gamma}_m$) for Coefficients of the Logistic Regression of Being Uninsured on Selected Variables for Persons Under Age 65, 2001 NHIS

	No imputation (NI)		*Single imputation (SI)*		*Multiple imputation (MI)*			*Ratio of estimated SEs:*	*Ratio of estimated SEs:*
Variable	*Point estimate*	*Estimated SE*	*Point estimate*	*Estimated SE*	*Point estimate*	*Estimated SE*	$\hat{\gamma}_m$	*NI/MI*	*SI/MI*
Constant	−3.67	.090	−3.35	.079	−3.37	.078	.027	1.15	1.02
Poverty ratio									
<1.00	2.09	.074	1.80	.059	1.82	.060	.094	1.23	.98
[1, 2)	1.98	.059	1.64	.049	1.66	.052	.156	1.14	.94
[2, 4)	1.05	.060	.86	.048	.89	.051	.166	1.17	.93
≥4 (reference)									
Age (years)									
<18	−.40	.065	−.29	.055	−.30	.055	.009	1.19	1.00
18–24	.97	.073	.96	.058	.96	.058	.006	1.25	1.00
25–34	.71	.062	.71	.050	.70	.050	.004	1.23	1.00
35–44	.42	.060	.43	.049	.42	.049	.004	1.23	1.00
45–54	.34	.059	.29	.050	.28	.050	.012	1.17	1.00
55–64 (reference)									
Race/ethnicity									
Hispanic	.65	.066	.64	.054	.64	.054	.011	1.22	1.00
Non-Hispanic black	.08	.058	.12	.049	.11	.049	.011	1.19	1.00
Non-Hispanic other	.11	.125	.08	.106	.08	.108	.017	1.16	.98
Non-Hispanic white (reference)									
Born in the U.S.?									
No	.67	.054	.77	.050	.77	.050	.010	1.09	1.00
Yes (reference)									
Gender									
Male	.25	.025	.23	.021	.23	.021	.001	1.20	1.00
Female (reference)									
Region of residence									
Northeast	−.08	.070	−.11	.059	−.10	.059	.010	1.18	.99
South	.43	.056	.44	.047	.45	.047	.001	1.20	1.01
West	.20	.076	.19	.063	.19	.063	.010	1.21	1.00
Midwest (reference)									
Resides in metropolitan area?									
Yes	−.19	.058	−.19	.057	−.19	.057	.003	1.03	1.00
No (reference)									

Second, the approximate fractions of missing information for the coefficients of variables other than the indicators for poverty ratio category are all near 0. Evidently, the high nonresponse on income does not result in much lost information about the coefficients of the other variables. Consistent with the low fractions of missing information for the coefficients of the nonincome variables, the ratios of the estimated standard errors with single imputation to those with multiple imputation are all near 1 for those variables. This is to be expected, because if there is little missing information, then there will be little variability across imputations (i.e., a small value of B_m), and the multiple-imputation analysis will not differ much from the single-imputation analysis.

Finally, the ratios of the estimated standard errors from complete-case analysis to those based on the multiply imputed data tend to be larger for the coefficients of variables other than the poverty ratio indicators, although this is not always the case for the logistic regression for health insurance. This is to be expected as well, because the analysis of just the complete cases excludes the observations with nonresponse on the poverty ratio and thus disregards a substantial number of observed values on other variables that are included in the analysis of the multiply imputed data. (Recall the caveat in Sec. 3.2 about interpreting estimated standard errors from complete-case analysis, however.)

3.5 Poststratification Reweighting of the NHIS Complete Cases to Agree With Estimated Totals From the CPS

Weighting adjustments are typically used by data producers for unit nonresponse, whereas imputation is typically used for item nonresponse, as occurs in the application presented in this article (see, e.g., Little and Rubin 2002, ex. 1.2). Nevertheless, as an alternative to imputation, a sophisticated analyst might consider a procedure such as reweighting the complete cases from the NHIS so that, within poststrata, the estimated population sizes from the NHIS agree with those from the CPS, which is considered a standard source of national income data (see, e.g., *http://www.census.gov/hhes/www/income/income.html*). This section presents such an analysis.

For poststratum k, $k = 1, \dots, K$, let $\hat{N}_k$ denote the NHIS complete-case estimate of the population size, and let $\hat{C}_k$ denote the CPS estimate. For each NHIS complete case within poststratum k, say case i, let w_i denote the original survey weight from the NHIS. Then the poststratification weighting adjustment changes the weight for case i to $w_i(\hat{C}_k/\hat{N}_k)$.

A poststratification weighting adjustment at the person level was carried out using the 96 poststrata defined by the cross-classification of age category (<18, 18–44, 45–64, or ≥65), gender (male or female), race/ethnicity category (Hispanic, non-Hispanic black, or non-Hispanic other), and poverty ratio (<1, [1, 2), [2, 4), or ≥4). Across the complete cases in the

Table 8. Point Estimates and Estimated SEs for the Percentage of Persons Age 45–64 in Fair or Poor Health and the Percentage of Persons Under Age 65 With no Health Insurance Coverage, by the Ratio of Family Income to the Federal Poverty Threshold, 2001 NHIS, With a Poststratification Weighting Adjustment Using Control Totals From the 2001 CPS

Poverty ratio	Percentage of persons age 45–64 in fair or poor health		Percentage of persons under age 65 with no health insurance coverage	
	Point estimate	Estimated SE	Point estimate	Estimated SE
<1	45.7	1.78	30.8	1.04
[1, 2)	33.0	1.34	28.6	.69
[2, 4)	16.1	.62	13.3	.35
≥4	5.9	.34	4.7	.22

NHIS, the poststratification reweighting factors ($\hat{C}_k/\hat{N}_k$) had a mean of 1.46 and a standard deviation of .172.

Table 8 presents the results of the poststratification adjustment for the analysis problems considered in Sections 3.2 and 3.3. A comparison with Tables 2 and 4 reveals that the point estimates and estimated standard errors following the poststratification adjustment are very similar to those obtained from complete-case analysis. (The estimation of standard errors for the poststratification analyses did not account for the fact that the CPS totals are estimates; accounting for this fact would likely increase the estimated standard errors somewhat.) The results (not shown here) of poststratification analyses for the logistic regressions discussed in Section 3.4 are also very similar to the results of the corresponding complete-case analyses displayed in Tables 6 and 7. Evidently, although the poststratification adjustment by definition forced the estimated population totals within poststrata to agree with those from the CPS, it had little effect on the analyses of the relationship between health variables and poverty status. This suggests that the additional data used in creating the imputations but not in performing the poststratification adjustment, such as the health-related variables and bounds on income for the nonrespondents, contain additional information about the differences between the nonrespondents and the respondents.

4. DISCUSSION

4.1 Summary of Results

Multiple imputation of missing income data in the NHIS has been carried out for the survey years 1997–2004 using an adaptation of SRMI (Raghunathan et al. 2001) that handles the hierarchical nature of the data, and there are plans to create imputations for later years. The imputed income files are available, with documentation, at the NHIS website (*http://www.cdc.gov/nchs/nhis.htm*).

Analyses of data from the 2001 NHIS indicate a negative association between poverty ratio and both being in fair or poor health and being uninsured. Although these relationships are found in analyses of the data both without and with imputation, estimates for some of the poverty ratio categories change substantially as a result of imputation. Examinations of observed data on two-category income suggest that imputation corrects for biases that occur in estimates based on the data without imputation. Moreover, multiple imputation usually results in lower estimated standard errors than analyses of the data without imputation.

The approximate fractions of missing information in the multiple-imputation analyses tend to be substantially smaller than the roughly 30% nonresponse rate for the exact value of family income, reflecting the information supplied by the imputation procedure.

4.2 Areas for Further Research

4.2.1 Iterating Between Imputing Income Items and Imputing Covariates. The NHIS imputation procedure went beyond the usual implementations of sequential regression methods by iterating between the imputation of person-level covariates (step 1 of Sec. 2.2) and the imputation of income items (steps 3 and 4 of Sec. 2.2), with SRMI used in each of the steps. Breaking the process up into steps facilitated handling the hierarchical nature of the data. The iterations were carried out to fully incorporate relationships between the covariates and the income items into the imputations.

The refinement of iterating between the tasks in the NHIS imputation project likely had a very minor effect, given that the missing-data rates for most of the covariates were very low. For this reason, during the development of the imputation procedure, omission of this refinement was considered. Including this refinement was straightforward, however, so it was carried out because it is preferable in principle. Research on the use of such iterations in other applications with hierarchical data structures and substantially higher rates of missing covariate data would be worthwhile.

4.2.2 Inconsistencies Between Family Income and Family Earnings. Because the items suggested for inclusion in family income in the NHIS questionnaire are all nonnegative and include the personal earnings of family members (see Sec. 1.1), it follows that family income ideally should be at least as large as family earnings. As noted in Section 1.1, however, family income in the NHIS is estimated by the respondent rather than constructed by summing responses to more detailed questions, such as the question about personal earnings of members of the family, and the questions about personal earnings and family income are located in different sections of the NHIS questionnaire. Thus some inconsistencies between family income and family earnings, in terms of the former being lower than the latter, might be expected.

In the 1997–2004 NHIS, 7–10% of responding families per year had reported family incomes lower than the reported family earnings. (The percentages presented in this section are weighted; as was the case in Sec. 1.2, the weighted percentages are close to their unweighted counterparts.) Moreover, the imputation procedure results in a larger percentage of families with family incomes lower than family earnings, with 16–19% of the families in a completed dataset (including both observed and imputed values) having such inconsistencies.

A reason for the higher rate of inconsistencies in the imputed data is that in addition to the 7–10% inconsistency rate in the reported data from which the imputation model is estimated, 36–43% of responding families had reported family incomes exactly equal to their reported family earnings. Because the imputation model does not force equality of family income and family earnings for any families, the imputation procedure tends to produce differences between family income and family earnings that are close to zero for a large percentage of families, with several positive differences and several negative differences.

As part of this project, research has been conducted on enforcing consistency between the imputed values of family income and family earnings, as well as on imputing new values of family income for those families whose reported family incomes and family earnings are inconsistent. The methods that have been developed to date tend to distort the marginal distributions of family income and family earnings. Given that the primary interest of data analysts is in the relationship of health variables to family income on its own, especially its ratio to the poverty threshold, it was decided that family income and family earnings would be imputed without imposing consistency. Research into methods for handling inconsistencies through imputation, as well as methods for decreasing the prevalence of inconsistencies in the reported data, would be worthwhile.

[Received September 2004. Revised October 2005.]

REFERENCES

Barnard, J., and Rubin, D. B. (1999), "Small-Sample Degrees of Freedom With Multiple Imputation," *Biometrika*, 86, 948–955.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Kennickell, A. B. (1991), "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 112–121.

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.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991), "Large-Sample Significance Levels From Multiply-Imputed Data Using Moment-Based Statistics and an *F* Reference Distribution," *Journal of the American Statistical Association*, 86, 1065–1073.

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley.

Meng, X.-L. (1994), "Multiple-Imputation Inferences With Uncongenial Sources of Input" (with discussion), *Statistical Science*, 9, 538–573.

Meng, X.-L., and Rubin, D. B. (1992), "Performing Likelihood Ratio Tests With Multiply-Imputed Data Sets," *Biometrika*, 79, 103–111.

National Center for Health Statistics (2005), "2004 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description," Division of Health Interview Statistics, National Center for Health Statistics, Centers for Disease Control and Prevention, available at *http://www.cdc.gov/nchs/nhis.htm.*

Paulin, G. D., and Sweet, E. M. (1996), "Modeling Income in the U.S. Consumer Expenditure Survey," *Journal of Official Statistics*, 12, 403–419.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, 85–95.

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

——— (1996), "Multiple Imputation After 18+ Years" (with discussion), *Journal of the American Statistical Association*, 91, 473–489.

Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.

Stata Corporation (2003), *Stata Statistical Software: Release 8.0*, College Station, TX: Author.