Multiple Imputation of Family Income in 2020 National Health Interview Survey: Methods

September 2021

Contents

Abstract

The National Health Interview Survey (NHIS) provides a rich source of data for studying relationships between income and health and for monitoring the health and health care for persons at different income levels. However, the nonresponse rates are high for income variables. Through the survey year 2018, NHIS collected two key income items, total family income in the previous calendar year and personal earnings from employment in the previous calendar year. Multiple imputation methods were used to address high item nonresponse rates for these two income variables for the survey years 1997-2018. Starting in 2019, the NHIS questionnaire was redesigned and the personal earnings from employment in the previous calendar year was no longer collected. A new imputation model was developed to impute total family income in 2019 NHIS. Datasets containing the imputed values, along with related documentation, can be obtained from the NHIS Web site (http://www.cdc.gov/nchs/nhis.htm). The 2020 NHIS income imputation follows the same imputation procedure of the 2019 NHIS. This report describes how total family income was imputed in 2020 NHIS and methods for analyzing the multiply imputed data.

1. Introduction

The National Health Interview Survey (NHIS) is a multi-purpose health survey and is the principal source of information on the health of the civilian, noninstitutionalized household population of the United States (National Center for Health Statistics, 2021). The NHIS provides a rich source of data for studying relationships between income and health and for monitoring the 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 their access to and use of health care services. However, the nonresponse rates are high for income variables. Through the survey year 2018, NHIS collected two key income measures, total family income in the previous calendar and personal earnings from employment in the previous calendar year. Starting in 2019, the NHIS questionnaire was redesigned and only total family income in the previous calendar was collected.

1.1. Questions on Family Income in the 2020 NHIS

In 2019, the NHIS questionnaire underwent a major revision. The 2020 NHIS used the same income questions as the 2019 NHIS. Some of the relevant changes starting from 2019 are as follows:

- The 2018 and prior NHIS questionnaires collected detailed information on all family members through the Family Core component. The Family Core component collected information on everyone in the family and included sections on family relationships, health status, limitations of activities, health care access and utilization, and health insurance. Since 2019, the NHIS questionnaire no longer has a Family Core component, and does not collect as much information on every family member. The 2020 NHIS, like the 2019 NHIS, collected basic demographics for all persons in the household, such as age, sex, race and ethnicity, and employment status and education for all the adults in the household.
- Additionally, through 2018, all families within a household were included in the survey, and within each family, one adult (sample adult) and one child (if any; sample child) were randomly selected for sample adult and sample child interviews. Starting in 2019, the NHIS no longer sampled respondents separately from each family, but instead randomly sampled one adult (sample adult) and one child (if any; the sample child) within a household without considering whether they were from the same family. For households with multiple families, the sample adult and sample child could potentially be from different families.

- Lastly, as mentioned above, the personal earnings from employment in the previous calendar year was no longer collected in 2019 NHIS and forward; instead only total family income in the previous calendar was collected. The respondent was asked about total combined family income for all family members including children as follows: "What is your best estimate of {your total income/the total income of all family members} from all sources, before taxes, in {last calendar year}?" If the respondent refused or did not know the amount, a series of income bracketing questions were asked starting with the question: "Was your total family income from all sources less than <<250% of poverty threshold>> or <<250% of poverty threshold>> or more?" (see the Survey Description Document's Family Income section for the exact fill values used for the poverty threshold in the instrument). The poverty threshold was based on the size of their family, the number of children and the presence of a person aged 66 years or over. The poverty threshold dollar amounts are adjusted each year. Depending on the response to the 250% of poverty question, different follow-up questions were asked such as those about incomes greater than or less than 138%, 100%, 200% of poverty threshold. In addition, follow-up questions of income ranges were asked based on the respondent's answers including "was your total family income from all sources less than $75,000 or $75,000 or more?", "less than $100,000 or $100,000 or more", "less than 400% of poverty threshold or 400% of poverty threshold or more" and "less than $150,000 or $150,000 or more" (see the questionnaire for the exact questions and skip patterns).

1.2. Missing Data on Family Income

For the years 1997 – 2006, the weighted percentages of persons with unknown exact family income ranged between 24% and 34%. For the years 2007 – 2019, the weighted percentages of persons with unknown exact family income ranged between 19% and 33% for the exact value and between 4% and 15% for the family income bracketing questions. In 2020, the weighted percentages of families with unknown exact family income was 24% for the exact value and 9% for any of the family income bracketing questions. There is evidence that the nonresponse on family income was related to several person-level and family-level characteristics (Schenker *et al.,* 2006). Thus, the respondents cannot be treated as a random subset of the original sample. One common method for handling missing data in software packages is "complete-case analysis" (Little and Rubin 2002, Section 3.2), also known as "listwise deletion," which deletes cases that are missing any of the variables involved in the analysis. However, since item nonresponse is not completely random, simply deleting cases with missing data can result in biased analyses. Moreover, since deletion of incomplete cases discards some of the observed data, complete-case analysis generally produces estimates that are less precise than those produced by methods that use all of the observed data.

1.3. Multiple Imputation of Income of 2020 NHIS

Multiple imputation was used to impute missing data on the family income variable. This document describes how the family income variable was imputed and provides guidance on using multiply imputed data in analyses.

Similar to the multiple imputation procedure in survey years 1997 – 2018 (Schenker *et al.,* 2006) and in 2019, the imputation procedure on family income in 2020 NHIS incorporated a large number of predictors, including demographic and health-related variables. For each year in the range 1997 to 2018, there were five multiply imputed income datasets, one for each imputation. Starting in 2019, 10 multiply imputed family income values were created to more precisely assess the variability due to imputation. Increasing the number of imputations (e.g. to 10 or higher) produces more efficient estimates for a wide variety of analyses (Van Buuren 2018). Analysts are recommended to use all 10 imputed files.

1.4. Objective and Contents of this Report

The objective of this report is to describe the approach used to multiply impute missing family income in the 2020 NHIS and methods for analyzing the multiply imputed data. Section 2 provides an overview of multiple imputation and how the 2020 NHIS family income variable was imputed. Section 3 describes how multiply imputed data should be analyzed, including which software packages which can be used to analyze multiply imputed data.

2. Multiple Imputation

2.1. Overview of Multiple Imputation

Imputation is a popular approach to handling nonresponse on items in a survey for several reasons. First, imputation is intended to reduce bias that can be introduced in analyses by discarding all cases with item nonresponse due to differences between item nonrespondents and respondents. Second, imputation results in a greater number of cases being available in the analysis. Third, when a data set is being produced for analysis by the public, imputation by the data producer allows the incorporation of specialized knowledge about the reasons for missing data in the imputation procedure, including confidential information that cannot be released to the public. Moreover, the nonresponse problem is addressed in the same way for all users, so that analyses can be consistent across users.

Although single imputation, that is, imputing one value for each record with item nonresponse, has the positive attributes just mentioned, analysis of a singly imputed data set fails to reflect the uncertainty stemming from the fact that the imputed values are plausible replacements for the missing values but are not the true values themselves. As a result, analyses of singly imputed data tend to produce estimated standard errors that are too small, confidence intervals that are too narrow, and significance tests that reject the null hypothesis more frequently when it is true. For example, large-sample results reported in Rubin and Schenker (1986) suggest that when the item nonresponse is between 20% to 30%, nominal 95% confidence intervals computed from singly imputed data have actual coverage rates between 85% and 90%. Moreover, the performance of single imputation can be even worse when inferences are desired for a multi-dimensional quantity, such as a k-component regression coefficient vector θ, k>1. For example, large-sample results reported in Li, Raghunathan, and Rubin (1991) demonstrate that for testing hypotheses about multi-dimensional quantities, the actual rejection rate under the null hypothesis increases as the number of components being tested increases, and the actual rejection rate can be much larger than the nominal rate.

Multiple imputation (Rubin 1978, 1987) is a technique that seeks to retain the general advantages associated with imputation while also allowing the uncertainty due to imputation to be reflected in the analysis. The idea is to simulate M > 1 plausible sets of replacements for the missing values, thereby generating M completed data sets. The M completed data sets are analyzed separately using a standard method for analyzing complete data, and then the results of the M analyses are combined in a way that reflects the uncertainty due to imputation. Details of how to analyze multiply imputed data are provided in Section 3.

With multiple imputation, the M sets of imputations for the missing values are ideally independent draws from the predictive distribution of the missing values conditional on the observed values. Consider, for example, the simple case in which there are two variables, X and Y, with Y subject to nonresponse and X fully observed. Suppose further that the imputation model specifies that: Y has a normal linear regression on X, that is, $Y=\beta_0+$

$\beta_1 X+\varepsilon$, where ε has a normal distribution with mean 0 and variance σ^2; and given X, the missing values of Y are only randomly different from the observed values of Y. After the regression of Y on X is fitted to the complete cases, a single set of imputations for the missing Y-values can be generated in two steps. First, values of β_0, β_1, and σ^2 are drawn randomly from the joint posterior distribution of the regression parameters. For example, the appropriately scaled chi-square distribution could be used for drawing σ^2, and the appropriate bivariate normal distribution could be used for drawing β_0 and β_1 given σ^2. Second, for each nonrespondent, say nonrespondent *i*, the missing value of Y is drawn randomly as $\beta_0+ \beta_1 X_i+\varepsilon$, where X_i is the X-value for nonrespondent *i*, and ε is a value drawn from a normal distribution with mean 0 and variance σ^2. The first step reflects the uncertainty due to the fact that the regression model was fitted to just a sample of data, and the second step reflects the variability of the Y-values about the regression line. Multiple imputations of the missing Y-values are generated by repeating the two steps independently M times. Although most imputation problems, including the imputation of missing data in the NHIS, are much more complicated than the simple example just presented, the basic principle illustrated by the simple example, reflecting all of the sources of variability across the M sets of imputations, still applies.

2.2. Procedure for Creating Imputations for the 2020 NHIS

There are a few issues with the imputation of family income in 2020 NHIS. First, variables used in the imputation models are hierarchical in nature, i.e., a few are reported for all members of the household, some variables are reported at the family level, and, and most variables are reported on the sample adult and sample child level. Second, for some respondents, no income information was collected while for other respondents, even though their exact family income was not obtained, some information on what the upper and/or lower bound of their family income was collected. This information needed to be incorporated in the imputation algorithm. For example, as discussed in Section 1.2, some respondents did not report exact income values but did report coarser income ranges; such ranges were used to form bounds for imputing exact income. Third, a sample adult and a sample child (if children are present) are randomly selected from the same household; for households with multiple families, the selected sample adult and sample child may be from different families and therefore their family incomes would need to be imputed separately. Fourth, the variables used as predictors in the imputation procedure often had small percentages of missing values that themselves needed to be imputed. Finally, due to the COVID-19 pandemic, the 2020 NHIS included a follow back (FB) sample, i.e., a subset of respondents interviewed in the 2019 NHIS were invited to participate in the 2020 NHIS. To account for the inclusion of the FB sample in the 2020 NHIS datafile, a follow back indicator was included in the imputation model. In addition, a sensitivity analysis was conducted to compare the impact of including only the FB indicator with including the 2019 family income variable in the model. The inclusion of the 2019 family income variable in the model was evaluated by including the interaction of 2019 NHIS family income and the follow back indicator into the imputation model as an additional predictor (the interaction term was used since the 2019 NHIS income was unavailable for the 2020 NHIS regular samples). While the inclusion of the 2019 income variable for FB cases yielded similar means and quantiles of family income but it did yield slightly smaller variances than the model with just the FB indicator. However, to be consistent with the 2019 and the future NHIS income imputation procedures, the 2020 NHIS public-use imputed income was based on the imputation model with the follow back indicator included, but without 2019 family income, in the model.

The following two sections describe the imputation procedure. Section 2.2.1 provides an overview of the steps in the procedure, the general algorithm used, and how features of the sample design were incorporated into the procedure. In Section 2.2.2, some additional details of the steps in the imputation procedure are described.

Note that in the process of imputing family income, missing values of several additional variables were imputed, and several new variables were created as well. These additional variables and imputed values were not retained in the final NHIS public-use multiply imputed income datafile.

2.2.1. Overview of the 2020 NHIS Imputation Procedure

2.2.1.1 Steps in the Imputation Procedure

In the imputation of family income, several family-level covariates were used, including some summaries based on data collected about each family member. Most of the variables formed by summarizing data on each adult family member had very low rates of missingness. To facilitate their use, their missing values were imputed prior to the imputation of family income. Any remaining missing values in the family-level and sample adult-level covariates, were imputed together with family income.

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

1. Impute missing values of education and working status covariates (education, working for pay at a job or business, working 35 hours or more in total) based on all adult family members
2. Create family-level covariates
3. Impute missing values of family income, and re-impute any missing values of family-level covariates for use in the next iteration of step 1.

The family income variable was used in the initial imputation of covariates in step 1. After steps 2 and 3 were carried out, the procedure cycled through steps 1 – 3 five more times, with the imputed income and family-level variables included as predictors in step 1. To create multiple imputations, the entire imputation process described above was repeated independently 10 times.

2.2.1.2 Sequential Regression Multivariate Imputation

The imputations in steps 1 and 3 described in Section 2.2.1.1 were created using sequential regression multivariate imputation (SRMI) (Raghunathan *et al*. 2001), as implemented by SAS proc MI procedure.

A brief description of SRMI is as follows; see Raghunathan *et al*. (2001), He *et al*. (2021) for details. Let X denote the fully-observed variables, and let $Y_1,Y_2,...,Y_k$ denote k variables with missing values, ordered by the amount of missingness, from least to most. The imputation process for $Y_1,Y_2,...,Y_k$ proceeds in c rounds. In the first round: Y_1 is regressed on X, and the missing values of Y_1 are imputed (using a process analogous to that described in the simple example of Section 2.1); then Y_2 is regressed on X and Y_1 (including the imputed values of Y_1), and the missing values of Y_2 are imputed; and so on, until Y_k is regressed on X, $Y_1,Y_2,...,Y_{k-1}$, 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 now, in each regression, all variables except for the variable to be imputed are included as predictors. Thus: Y_1 is regressed on X, $Y_2,Y_3,...,Y_k$, and the missing values of Y_1 are re-imputed; then Y_2 is regressed on $X,Y_1,Y_3,...,Y_k$, and the missing values of Y_2 are re-imputed; and so on. After c rounds, the final imputations of the missing values in $Y_1,Y_2,...,Y_k$ are used. A value of c = 5 was used in the NHIS income imputation.

The SAS FCS procedure specifies a multivariate imputation by fully conditional specification methods, it allows the following models:

- A normal linear regression model or predictive mean matching model if the Y-variable is continuous;

- A logistic regression model or discriminate function if the Y-variable is binary;
- A cumulative logit model if the Y-variable is ordinal;
- A generalized logit regression model or discriminate function if the Y-variable is categorical with more than two categories.

Because SRMI requires only the specification of individual regression models 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 to create the imputations for the NHIS was influenced in large part by the complicating factors summarized at the beginning of Section 2.2, specifically, the large number of predictors of varying types that had missing values. These complicating factors would be very difficult to handle using a method based on a full joint model.

2.2.1.3 Reflecting the Sample Design in Creating the Imputations

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 (Rubin 1996). The sample design of the NHIS was reflected in the imputations for this project via the inclusion of the indicators for the distinct combinations of stratum (strata_ER) and primary sampling unit (PSU_ER). The household weights were also used in the imputation of family income.

2.2.2. Further Details of the Imputation Procedure

Additional details of the steps outlined in Section 2.2.1.1 are now described.

Step 1: Imputing Person-Level Education and Employment Status Covariates for Adults

The variables included in the imputation of education and employment status for all adult family members are listed in Appendix A. The 2019 NHIS collects age, sex, race and ethnicity information of all persons in a family, as well as education and employment status (works for pay at a job or business, usually works 35 hours or more per week in total) for all adults in a family. Since the information for multiple members of a family was used to impute family income, missing values in education and employment status for adult persons were imputed. Design information (strata_ER, PSU_ER) was included in the imputation. Family-level covariates (such as family size, house/apartment owned or rented), geographic information (such as region, MSA status, urban/rural) and US census information (such as percentage of families with annual income less than $15,000 in the block group, median family income within a block group) were also included in the imputation of education and employment status covariates based on all adult family members.

Step 2: Creating Family-Level Covariates

The variables imputed for separate adult family members in step 1 were summarized, by family, to create family-level covariates for use in imputing family income. These family-level covariates are included in the listing in Appendix B. Examples include the total number of male and female earners in a family (M_EARN, F_EARN), total number of male and female adults who have college degrees (M_BA, F_BA), etc. Summaries of all persons within a family are also created, such as total size of family, number of adult persons in a family, percent of White, Black, Asian persons in a family, etc.

Step 3: Imputing Family Income (and Family-Level and Sample Adult-Level Covariates)

The variables included in the imputation of family income are listed in Appendix C. Family-level covariates (such as family size, house/apartment owned or rented, total number of male and female earners in a family),

geographic information (such as region, MSA status, urban/rural), US census information (such as percentage of families with annual income less than $15,000 in the block group, median family income within a block group), and sample adult-level covariates (such as age, sex, race and ethnicity) were included in the imputation model.

To determine a transformation for family income to conform to the normality assumption in the imputation model, Box-Cox transformations (Box and Cox 1964) were estimated from the complete cases for the regressions predicting family income. The closest simple transformation suggested by the Box-Cox analysis was the cube-root transformation, which is also close to and consistent with the optimal transformation (the power 0.375) found by Paulin and Sweet (1996) in modeling income data from the Consumer Expenditure Survey of the Bureau of Labor Statistics. After the imputation procedure was completed, the variables were transformed back to their original scale.

There were several families for whom an exact income was not reported, but an income category was reported. In each such case, the bounds specified by the reported category were used in imputing the family income. In addition to the bounds just described, a reported family income value was top-coded at the top-code value ($999,995). Family income was imputed using a linear regression model with predictive mean matching procedure (SAS proc MI). After family income was imputed, the imputed values were compared to the reported income bounds. For families with imputed values outside of the reported bounds, their family income values were re-drawn from truncated normal distributions (Thomopoulos 2018), where the mean and the variance of the truncated normal distribution for each family were derived from the linear regression model of family income given the covariates in the current imputation step, the lower and upper bounds of the truncated normal distribution for each family were based on their reported income bounds.

Missing values of family-level and sample adult-level variables (if any) were also imputed during this imputation step and the imputed family-level values were used in the imputation of person-level covariates in step 1 of the next iteration.

3. Analyzing Multiply Imputed Data

3.1. General Procedures

Suppose that the primary interest is in estimating a scalar population quantity, such as a mean, a proportion, or a regression coefficient. The analysis of the M completed data sets resulting from multiple imputation proceeds as follows:

- Analyze each of the M completed data sets separately using a suitable software package designed for complete data (for example, SUDAAN or Stata or SAS).
- Extract the point estimate and the estimated standard error from each analysis.
- Combine the point estimates and the estimated standard errors to arrive at a single point estimate, its estimated standard error, and the associated confidence interval or significance test.

Technical details of how to analyze multiply imputed data are given in Section 3.2. Briefly, however, the combined point estimate is the average of the point estimates obtained from the M completed data sets. The estimated variance of the combined point estimate is computed by adding two components. The first component is the average of the estimated variances obtained from the M completed data sets. The second component is the variation among the point estimates obtained from the M completed data sets. The latter component represents the uncertainty due to imputing for the missing values. Confidence intervals and significance tests are constructed using a t reference distribution.

3.2. Technical Details for Analyzing Multiply Imputed Data

Suppose that M completed data sets have been generated via multiple imputation, and let Q denote the scalar population quantity of interest. Application of the chosen method of analysis to the i^{th} completed data set yields the point estimate $\hat{Q}_i$ and its estimated variance (square of the estimated standard error) U_i, where i =1,2,...,M. It is important to analyze each data set separately to derive the M point estimates and estimated variances.

The combined multiple-imputation point estimate is

$$\bar{Q}_M = \frac{1}{M}\sum_{i=1}^{M}\hat{Q}_i \qquad (1)$$

The estimated variance of this point estimate consists of two components. The first component, the "within-imputation variance"

$$\bar{U}_M = \frac{1}{M}\sum_{i=1}^{M}U_i$$

is, approximately, the variance that one would have obtained had there been no missing data. The second component, the "between-imputation variance"

$$B_M = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{Q}_i - \bar{Q}_M)^2,$$

is the component of variation due to differences across the M sets of imputations.

The total estimated variance of the multiple-imputation point estimate $\bar{Q}_M$ is

$$T_M = \bar{U}_M + \frac{M+1}{M}B_M. \qquad (2)$$

The factor (M+1)/M is a correction for small M. Furthermore, it is shown in Rubin and Schenker (1986) and Rubin (1987, Section 3.3) that, approximately,

$$T_M^{-1/2}(Q - \bar{Q}_M) \sim t_v$$

where the degrees of freedom v for the t distribution are given by

$$v = (M-1)\hat{\gamma}_M^{-2}$$

with

$$\hat{\gamma}_M = \frac{M+1}{M}\frac{B_M}{T_M}.$$

The quantity $\hat{\gamma}_M$ measures the proportionate share of T_M that is due to between-imputation variability; it is also approximately the fraction of information about Q that is missing due to nonresponse (Rubin 1987, p. 93).

For a multi-dimensional population quantity Q, such as a k-component regression coefficient vector θ, k>1, Li, Raghunathan, and Rubin (1991) developed multiple-imputation procedures for significance testing when the hypothesis to be tested involves several parameter estimates simultaneously. In addition, Li, Meng, Raghunathan, and Rubin (1991) developed procedures for combining test statistics and p-values (rather than point estimates and estimated variances) computed from multiply imputed data.

The procedures described above assume that the degrees of freedom that would be used for analyzing the complete data if there were no missing values, i.e., the "complete-data degrees of freedom," are large (or infinite); that is, a large-sample normal approximation would be valid for constructing confidence intervals or performing significance tests if there were no missing data. This is clearly not true in many survey settings, where the number of sampled PSUs may be small, and a t reference distribution would be used if there were no missing data. For example, for a survey involving H strata with 2 PSUs selected from each stratum, the complete-data degrees of freedom for inferences about the population mean are H.

Barnard and Rubin (1999) relaxed the assumption of large complete-data degrees of freedom and suggested the use of

$$v' = (\frac{1}{v} + \frac{1}{k})^{-1}$$

for the multiple-imputation analysis, where

$$k = \frac{df\,(df+1)}{df+3}(1-\hat{\gamma}_M)\,,$$

and *df* are the complete-data degrees of freedom.

For the NHIS multiply imputed data, depending on the survey year there are 5 or 10 versions of the imputed variables (M=5 or 10), and the complete-data degrees of freedom, *df*, are 300 or more for many analyses. For v' or v greater than 100, the normal approximation is generally valid. When v' and v are small, for many analyses of the NHIS data, use of either v or v' should give similar results, although use of v' will be slightly more conservative (smaller degrees of freedom) (National Center for Health Statistics, 2018).

3.3. Software Packages for Analyzing Multiply Imputed Data

After analyzing each of the M completed data sets resulting from multiple imputation, one can combine the results of the M analyses using software packages. SAS-callable SUDAAN is a software package for analyzing data from complex surveys, which includes a built-in option for analyzing multiply imputed data (Research Triangle Institute, 2012). IVEware is a free SAS-callable software package, which has different modules for performing various multiple-imputation analyses incorporating complex sample designs. IVEware can be downloaded from the Web site https://www.src.isr.umich.edu/software/. SAS users can also use SAS proc procedures to conduct statistical analysis on each imputed data and then use SAS proc MIanalyze procedure to combine results of analyses of multiply imputed data (SAS, 2016). Stata procedures (StataCorp LP 2009) for performing multiple-imputation analyses and the mice (Multiple Imputation with Chained Equations) package in R (van Buuren S,

Groothuis-Oudshoorn K, 2011) are also widely used to perform multiple imputation and the subsequent analysis.

3.4. Combining Data Across Years of the NHIS

A common practice with the NHIS, especially when rare events or small subsets of the population are being studied, is to combine more than one year of data in order to increase the sample size. For analyses of the combined data, the data files are typically concatenated and the analysis weights adjusted accordingly. Botman and Jack (1995) and the 2020 NHIS Survey Description Document available at https://www.cdc.gov/nchs/nhis/2020nhis.htm provide further information on how to conduct such analyses.

With the NHIS multiply imputed data, there are M=5 completed data sets for each year from 1997-2018, and M=10 completed data sets for each year from 2019 and forward. To combine more than one year of data, the corresponding completed data sets from the years in question can be concatenated to obtain concatenated completed data sets. Suppose, for example, that the data from 1999 and 2000 were to be combined. Then the first completed data set from 1999 and the first completed data set from 2000 would be concatenated to create the first concatenated completed data set for 1999 – 2000. The analogous concatenations would be carried out for the second through fifth completed data sets, with the end result being M=5 concatenated completed data sets for 1999 – 2000.

To combine the 2018 NHIS (which has M=5 imputed data) and the 2019 NHIS (which has M=10 imputed data), a user can combine the first 5 imputed data from NHIS 2019, with the corresponding completed data sets from 2018 NHIS to obtain M=5 concatenated completed data sets.

After concatenated completed data sets have been created by combining data across years, each of the concatenated completed data sets is analyzed using the standard techniques for concatenated data from multiple years of the NHIS, as described by Botman and Jack (1995) and the 2020 NHIS Survey Description Document available at https://www.cdc.gov/nchs/nhis/2020nhis.htm. The results of these analyses are then combined using the rules given in Section 3.2.

3.5. Analyzing Only a Single Completed Data Set

Users of the multiply imputed NHIS data who are unfamiliar with multiple imputation or who find the analysis of multiply imputed data cumbersome might be tempted to analyze only a single completed data set, such as the first imputed data. Such an analysis, which is equivalent to using single imputation, would produce point estimates that are unbiased (under the assumption that the imputation model is correct). However, as discussed in Section 2.1, it would produce underestimates of variability and resultant inferences that may be inaccurate, since it would not account for the additional variability due to imputation.

When applying a model-selection procedure such as stepwise regression, it is not clear how to formally combine the results from M completed data sets. Therefore, an analyst might decide to apply the model-selection procedure to, for example, just the first completed data set. Since variability would be underestimated, such an approach would tend to judge more variables as “statistically significant” than would be the case if variability were estimated correctly. Thus, fewer variables would tend to be eliminated from the model under single imputation. Recent developments on variable selection on multiply imputed data can be found in Chen and Wang (2013), Geronimi and Saporta (2017).

References

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

Botman, S.L., and Jack, S.S. (1995), "Combining National Health Interview Survey Datasets: Issues and Approaches," *Statistics in Medicine*, 14, 669-677.

Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 26, 211-243.

Chen, Q. and S. Wang, S. (2013), "Variable selection for multiply-imputed data with application to dioxin exposure study," Stat. Med., 32 (21) 3646-3659.

Geronimi, J. and Saporta G. (2017), "Variable selection for multiply-imputed data with penalized generalized estimating equations," Computational Statistics & Data Analysis, Volume 110, 103-114.

He, Y., Zhang, G., Hsu, C.H. (2021), *Multiple Imputation of Missing Data in Practice: Basic Theory and Analysis Strategies,* CRC Press (in press).

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 edition, Hoboken: Wiley.

National Center for Health Statistics (2021), "2020 National Health Interview Survey (NHIS) Public Use Data Release: Survey Description," Division of Health Interview Statistics, National Center for Health Statistics. Available from the NHIS Web site (http://www.cdc.gov/nchs/nhis.htm).

National Center for Health Statistics (2018). "Multiple Imputation of Family Income and Personal Earnings in the National Health Interview Survey: Methods and Examples". Hyattsville, Maryland.

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.

Research Triangle Institute (2012). SUDAAN Language Manual, Volumes 1 and 2, Release 11.
Research Triangle Park, NC: Research Triangle Institute.

Rubin, D.B. (1978), "Multiple Imputation in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 20-34.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley. Rubin, D.B. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473-489.

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

SAS Institute Inc. 2016. SAS/STAT 14.2 User's Guide. Cary, NC: SAS Institute Inc.

Schenker, N., Raghunathan T.E., Chiu, P.-L., Makuc D.M., Zhang G., and Cohen A.J. (2006), "Multiple Imputation of Missing Income Data in the National Health Interview Survey," *Journal of the American Statistical Association*, 101, 924-933.

StataCorp LP. (2009), *Stata Multiple Imputation Reference Manual: Release 11*, College Station: Stata Press.

Thomopoulos NT (2018). Probability Distributions: With Truncated, Log and Bivariate Extensions. Springer.

Van Buuren S, Groothuis-Oudshoorn K (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **45**(3), 1-67. https://www.jstatsoft.org/v45/i03/

Van Buuren S. (2018). Flexible Imputation of Missing Data, 2nd Edition, Chapman and Hall/CRC

Appendix A. Variables in the imputation of person-level education and employment status for adults (step1)

Variable name	Label and Code values
AGE_P	Age of person (from the roster – an restricted use variable)
SEX_FINAL	Sex 1 = male 2 = female
RACRECI4	Race recode 1= White 2= Black 3= Asian 4= all other race groups
HISP_FINAL	Ethnic origin 1 = Hispanic 2 = Non-Hispanic
EDUC_R	Education recode 1=high school or less than high school 2=some college 3=college 4=master and above
FEMWORK	Person works for pay at a job or business 1=yes 2=no
FEMWKFT	Person usually works 35 hours or more per week in total 1=yes 2=no
CUBROOT_INCOME	Family income, cubic root transformed
UCF_UASIAN	% Asian in a census block group
UCF_UBLACK	% Black in a census block group
UCF_UHISPAN	% Hispanic in a census block group
UCF_UINCOME	% families with annual income < $15K in a census block group
UCF_INCOME_MEDIAN	Median family income within a census block group
UCF_POVRATE	Poverty rate % by tract
MSA	MSA residence 1=in MSA and in the principal city 2=in MSA and not in the principal city 3= in MSA but cannot determine if in principal city 4= not in an MSA city
REGION	1 = Northeast 2 = South 3 = West 4 = Midwest
URB_RRL	Urban/Rural 1 = Urban 2 = Rural
INCWRKO	Any family members 18 or older received income from wages, salaries, commissions, bonuses, tips, or self-employment last year 1=yes 2=no
INCINTER	Any family members received income from interest-bearing accounts or investments, dividends from stocks or mutual funds, net rental income, royalty income, or income from estates and trusts last year 1=yes

	2=no
INCSSRR	Any family members received income from Social Security or Railroad Retirement last year 1=yes 2=no
INCSSISSDI	Any family members received supplemental Security Income, SSI, or Social Security Disability Income, SSDI, which are different from Social Security last year 1=yes 2=no
INCWELF	Any family members received any public assistance or welfare payments from the state or local welfare office last year 1=yes 2=no
FSNAP12M	Any family members received food stamp benefits last year 1=yes 2=no
HOUYRSLIV	About how long has person lived in this house/apartment 1 Less than 1 year 2 1-3 years 3 4-10 years 4 More than 10 years
HOUTENURE	House/apartment owned or rented 1=Owned or being bought 2=rented 3=other arrangement
HOUGVASST	Paying lower rent because the Federal, State, or local government is paying part of the cost 1=yes 2=no
TELCURWRK	At least one telephone is currently working and is not a cell phone? 1=yes 2=no
PAYBLL12M	Anyone in the family have problems paying or were unable to pay any medical bills 1=yes 2=no
FM_TOTAL	Family size
STRATA_PSU	Stratum and PSU combination recoded based on STRAT_ER, PSU_ER from the inhouse NHIS data file

Appendix B. Family-Level covariates created from all persons within a family (step 2)

Variable name	Label and Code values
FM_TOTAL	Size of family
FM_ADULT	Number of adults 18 and older in family
P_BLACK	% of Black in family
P_WHITE	% of White in family
P_ASIAN	% of Asian in family
M_EARN	Number of males who work for pay last year in family
F_EARN	Number of females who work for pay last year in family
M_35HRABOVE	Number of males had job and worked 35+ hours last week in family
F_35HRABOVE	Number of females had job and worked 35+ hours last week in family
M_ERNAGE	Mean age of male earners in family
F_ERNAGE	Mean age of female earners in family
M_BA	Number of male adults who have college degrees in family
F_BA	Number of female adults who have college degrees in family
M_MSABOVE	Number of male adults who have master and above degrees in family
F_MSABOVE	Number of female adults who have master and above degrees in family

Appendix C. Variables included in imputation of family income (step 3)

Variable name	Label and Code values
FM_TOTAL	Size of family
FM_ADULT	Number of adults 18 and older in family
P_BLACK	% of Black in family
P_WHITE	% of White in family
P_ASIAN	% of Asian in family
M_EARN	Number of males who work for pay last year in family
F_EARN	Number of females who work for pay last year in family
M_35HRABOVE	Number of males had job and worked 35+ hours last week in family
F_35HRABOVE	Number of females had job and worked 35+ hours last week in family
M_ERNAGE	Mean age of male earners in family
F_ERNAGE	Mean age of female earners in family
M_BA	Number of male adults who have college degrees in family
F_BA	Number of female adults who have college degrees in family
M_MSABOVE	Number of male adults who have master and above degrees in family
F_MSABOVE	Number of female adults who have master and above degrees in family
UCF_UASIAN	% Asian in a census block group
UCF_UBLACK	% Black in a census block group
UCF_UHISPAN	% Hispanic in a census block group
UCF_UINCOME	% families with annual income < $15K in a census block group
UCF_INCOME_MEDIAN	Median family income within a census block group
UCF_POVRATE	Poverty rate % by tract
HOUSEHOULD_WEIGHT_FINAL	Internal household sampling weights
STRATA_PSU	Stratum and PSU combination recoded based on STRAT_ER, PSU_ER from the inhouse NHIS data file
MSA	MSA residence 1= in MSA and in the principal city 2= in MSA and not in the principal city 3= in MSA but cannot determine if in principal city 4= not in an MSA city
REGION	1 = Northeast 2 = South 3 = West 4 = Midwest
URB_RRL	Urban/Rural 1 = Urban 2 = Rural
INCWRKO	Any family members 18 or older received income from wages, salaries, commissions, bonuses, tips, or self-employment last year 1=yes 2=no
INCINTER	Any family members received income from interest-bearing accounts or investments, dividends from stocks or mutual funds, net rental income, royalty income, or income from estates and trusts last year 1=yes 2=no
INCSSRR	Any family members received income from Social Security or Railroad Retirement last year 1=yes 2=no

INCSSISSDI	Any family members received supplemental Security Income, SSI, or Social Security Disability Income, SSDI, which are different from Social Security last year 1=yes 2=no
INCWELF	Any family members received any public assistance or welfare payments from the state or local welfare office last year 1=yes 2=no
FSNAP12M	Any family members received food stamp benefits last year 1=yes 2=no
HOUYRSLIV	About how long has the sample adult (if not available then use sample child's value) lived in this house/apartment 1 Less than 1 year 2 1-3 years 3 4-10 years 4 More than 10 years
HOUTENURE	House/apartment owned or rented 1=Owned or being bought 2=rented 3=other arrangement
HOUGVASST	Paying lower rent because the Federal, State, or local government is paying part of the cost 1=yes 2=no
TELCURWRK	At least one telephone is currently working and is not a cell phone? 1=yes 2=no
PAYBLL12M	Anyone in the family have problems paying or were unable to pay any medical bills 1=yes 2=no
AGE_A	Age of sample adult
HISP_A	Ethnic origin of sample adult 1 = Hispanic 2 = Non-Hispanic
RACRECI4_A	Race recode of sample adult 1= White 2= Black 3= Asian 4= All other race groups
SEXWT_A	Sex of sample adult 1 = male 2 = female
EDUC_A	Education recode of sample adult 1=high school or less than high school 2=some college 3=college 4=master and above
MARITAL_A	Marital status of sample adult 1= married or partner 2= widowed/devoiced/ separated 3= never married
HICOV_A	Health insurance of sample adult 1=yes

	2=no
HIKIND01_A	Private health insurance (sample adult) 1=yes 2=no
HIKIND02_A	Medicare (sample adult) 1=yes 2=no
HIKIND03_A	Medigap (sample adult) 1=yes 2=no
HIKIND04_A	Medicaid (sample adult) 1=yes 2=no
HIKIND08_A	State-sponsored health plan (sample adult) 1=yes 2=no
PHSTAT_A	Reported health status of sample adult 1= Excellent 2 = Very good 3 = Good 4 =Fair 5 = Poor
HYPEV_A	Hypertension status of sample adult 1=yes 2=no
CHLEV_A	Had high cholesterol? (sample adult) 1=yes 2=no
CHDEV_A	Coronary heart disease? (sample adult) 1=yes 2=no
PREDIB_A	Has a doctor or other health professional EVER told you that you had prediabetes or borderline diabetes? (sample adult) 1=yes 2=no
DIBEV_A	A doctor or other health professional ever told you that you had diabetes? (sample adult) 1=yes 2=no
COPDEV_A	Chronic Obstructive Pulmonary Disease, C.O.P.D., emphysema, or chronic bronchitis? (sample adult) 1=yes 2=no
ARTHEV_A	Some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia (sample adult) 1=yes 2=no
ANXEV_A	Any type of anxiety disorder? (sample adult) 1=yes 2=no
DEPEV_A	Any type of depression? (sample adult) 1=yes 2=no
HEARAID_A	Do you use a hearing aid? (sample adult) 1=yes 2=no

EQUIP_A	Do you use any equipment or receive help for getting around? (sample adult) 1=yes 2=no
UPPSLFCR_A	Difficulty with selfcare (sample adult) 1 = no difficulty 2 = some/a lot/can't do at all
SOCWRKLIM_A	Limited in the kind OR amount of work (sample adult) 1=yes 2=no
SINCOVDE_A	Covered by a SEPARATE plan that only pays for dental services (sample adult) 1=yes 2=no
SINCOVVS_A	Covered by a SEPARATE plan that only pays for vision services? (sample adult) 1=yes 2=no
SINCOVRX_A	SEPARATE plan that only pays for prescriptions? (sample adult) 1=yes 2=no
PAYWORRY_A	How worried are you that you will be able to pay your medical bills (sample adult)? 1 = very worried 2 = some 3 = not at all
USUALPL_A	Is there a place that you USUALLY go to if you are sick and need health care? (sample adult) 1=yes 2=no 3 =more than 1 place
HOSPONGT_A	Have you been hospitalized overnight? (sample adult) 1=yes 2=no
MEDDL12M_A	Have you DELAYED getting medical care because of the cost? (sample adult) 1=yes 2=no
SHTFLU12M_A	During the past 12 months, have you had a flu vaccination (sample adult) 1=yes 2=no
SMKEV_A	Have you smoked at least 100 cigarettes in your entire life (sample adult) 1=yes 2=no
SMKNOW_A	Now smoke cigarettes every day, some days or not at all? (sample adult) 1 = every day 2= some days 3= not at all
CITIZEN_A	citizen of the United States? (sample adult) 1=yes 2=no
SCHCURENR_A	Are you currently enrolled in or attending school? (sample adult) 1=yes 2=no
EMPLASTWK_A_R	Last week, did you work for pay at a job or business or did you have a job or business last week, but were temporarily absent due to illness, vacation, family or maternity leave, or some other reason? (sample adult)
EMPWRKHRS_A_R	How many hours did you work LAST WEEK/do you USUALLY work per week (sample adult) 0= not employed

	1= less than 35 hr 2 =more than 35 hr
EMPPDSKLV_A	Is paid sick leave available if you need it? (sample adult) 0= not employed 1= yes 2 =no
EMPOFFHI_A	Was health insurance offered to you through your workplace? (sample adult) 1=yes 2=no
FB_INDICATOR	Follow back indicator 1=2019 follow back sample 0= not 2019 follow back sample