Multiple Imputation of Missing Household Poverty Level Values from the National Survey of Children with Special Health Care Needs, 2001, and the National Survey of Children's Health, 2003

Steven Pedlow, National Opinion Research Center

Julian V. Luke and Stephen J. Blumberg,
Centers for Disease Control and Prevention

The authors wish to thank Nathaniel Schenker and Pei-Lu Chiu, whose detailed reviews and comments helped us to make many improvements to this report. In addition, special thanks to Dr. Schenker for permission to borrow text from his report (with Trivellore E. Raghunathan, Pei-Lu Chiu, Diane M. Makuc, Guangyu Zhang, and Alan J. Cohen) on the multiple imputation of family income and personal earnings in the National Health Interview Survey (available on-line at http://www.cdc.gov/nchs/about/major/nhis/2005imputedincome.htm).

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics
Division of Health Interview Statistics
Survey Planning and Special Surveys Branch

June 22, 2007

SAFER · HEALTHIER · PEOPLE

Introduction

The 2001 National Survey of Children with Special Health Care Needs (NS-CSHCN) and the 2003 National Survey of Children's Health (NSCH) provide a rich source of data for studying the relationships between income and health and for monitoring health and health care for children at different income levels. However, as is common for most household interview surveys, nonresponse rates were high for the question on total combined household income for the previous calendar year. Answers to this question, along with answers to a question about the number of people living in the household, are used to create an index of income relative to the Department of Health and Human Services Federal Poverty Guidelines. If data for either of these two components were missing, refused, or had a "don't know" response, the household poverty status indicator was assigned a missing value code in the publicly released datasets. (Further details about the procedures for assigning household poverty status are available in Appendix IV of *Design and Operation of the National Survey of Children with Special Health Care Needs, 2001* and in Appendix V of the *Design and Operation of the National Survey of Children's Health, 2003*.)

Table 1 summarizes the amount of missing data in the variables for each of the two surveys' datasets. For the 2001 NS-CSHCN, poverty status is missing for 15.0% of the households (29,463 of 196,888 households). For the 2003 NSCH, poverty status is missing for 9.2% of the households (9,414 of 102,353 households). In both surveys, missing values for poverty status were predominately the result of missing data for income rather than missing data for household size.

There is evidence that the nonresponse on household income was related to several child-level characteristics, including items pertaining to health. Thus, the respondents 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" (also known as "listwise deletion") will generally be biased because this method deletes cases that are missing any of the variables

Table 1. Raw frequencies and percent of missing items needed to calculate the poverty level variable, by survey

Survey and missing data categories	Frequency	Percent (out of Records with Missing Poverty Status)	Percent (out of Total Survey Records)
National Survey of Children with Special Health Care Needs, 2001			
Missing income but not household size	26,601	90.3	13.5
Missing income and household size	2,773	9.4	1.4
Missing household size but not income	89	0.3	0.0
Total missing at least one of these variables	29,463	100.0	15.0
National Survey of Children's Health, 2003			
Missing income but not household size	9,232	98.1	9.0
Missing income and household size	102	1.1	0.1
Missing household size but not income	80	0.8	0.1
Total missing at least one of these variables	9,414	100.0	9.2

involved in the analysis. Moreover, since deletion of incomplete cases discards some of the observed data, complete-case analysis is generally inefficient as well; that is, it produces inferences that are less precise than those produced by methods that use all of the observed data. Imputation is a more appropriate approach to handling nonresponse on items in a survey for several reasons. First, imputation adjusts for observed differences between item nonrespondents and item respondents; such an adjustment is generally not made by complete-case analysis. Second, imputation results in a completed data set, so that the data can be analyzed using standard software packages without discarding any observed values. 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 will be consistent across users.

Although single imputation, that is, imputing one value for each missing datum, enjoys the positive attributes just mentioned, analysis of a singly imputed data set using standard software 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 too often when it is true.

Multiple imputation (Rubin, 1978, 1987, 1996) is a technique that seeks to retain the advantages of single 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. For public-use data, M is not usually larger than five, which is the value that has been used here in multiply imputing missing data for the NS-CSHCN and the NSCH.

This report describes the procedures used in multiply imputing household income and household size for the NS-CSHCN and the NSCH. Household poverty status is expressed as a percentage; households with income less than 100% of the federal poverty level (FPL) are considered to be living in poverty. For each of the multiply imputed data sets, household poverty status was derived from the imputed values for household income and household size.

Imputation Procedures

Income and household size were each imputed five times, creating five imputed datasets. The literature (e.g., Rubin, 1987) suggests that this is a sufficient number of imputations unless the amount of missing information is extreme. As noted earlier, the number of survey records with missing household size values was much smaller than the number of survey records with missing household income values. Since there is very little missingness in household size to explain, we did not feel a need to explore additional predictors for household size. Therefore, household size was imputed using the same predictors used for household income.

The imputation of household income and household size was complicated by two issues. First, neither household income nor household size was normally distributed. This is a disadvantage because linear regression modeling assumes that the dependent variable being modeled has a normal distribution. Therefore, we used transformed variables for modeling and imputation. Paulin and Sweet (1996) found the optimal transformation for income data in the U.S. Consumer Expenditure Survey to be the three-eighths root (income to the power p = .375). We used the Box-Cox transformation algorithm (Box and Cox, 1964) to determine the optimal transformation for the 2001 NS-CSHCN and 2003 NSCH distributions of income and household size. For income, the optimal transformation was p = .23 for the 2001 NS-CSHCN and p = .28 for the 2003 NSCH; we have used the quarter-root (p = .25) for both surveys. For household size, the optimal transformation was p = 0 for both surveys; the natural logarithm was therefore used.

Second, in some cases, the imputed values of household income and household size needed to be constrained within certain bounds. Household respondents were asked to provide an exact household income. However, when respondents did not provide an exact household income, a series (i.e., cascade) of questions asking whether the household income was below, exactly at, or above threshold amounts were then asked. The multiple imputation procedures employed for the NS-CSHCN and the NSCH needed to impute the income value so that it was consistent with any information gathered from the cascade questions. For households with missing data on household size, we also needed to restrict the imputed values so that they were consistent with other information provided in the survey (e.g., household size is greater than the number of children in the household).

Fortunately, there is software that allows constrained multiple imputation. IVEware, described in Raghunathan et al. (2002) and available online at http://www.isr.umich.edu/src/smp/ive, allows the user to specify lower and upper limits of imputed values, constraining the imputation distribution from which draws are made. This software was used by Schenker et al. (2006) to impute family income for the National Health Interview Survey (NHIS).

IVEware uses the sequential regression multivariate imputation (SRMI) of Raghunathan et al. (2001), which does not necessarily imply a joint model (see Schafer, 1997) for both income and household size conditional on the predictor variables. Since the software uses sequential regression imputations, income and household size will have separate models (using the same covariates, including each other). However, we concluded that the ability to constrain the imputed values outweighed this slight disadvantage. IVEware is a generally accepted multiple imputation program since it does impute the variables simultaneously.

IVEware builds regression models, and then multiply imputes variables based on the models built. For understanding model relationships, parsimony is desired, but in prediction (imputation can be thought of as "predicting" the missing values), more complicated models are often better for two reasons. First, using more variables leads to a higher correlation between the observed and predicted values for a model. Second, the validity of analyses conducted on multiply-imputed datasets is broader when more variables are included in the model (see Meng, 1995). Of course, larger models can lead to overfitting and also can be a drain on computer resources, time, the software and the number of degrees of freedom in the data. The size of the data sets involved is substantial (the 2001 NS-CSHCN has 196,888 households while the 2003 NSCH has 102,353 households), so this last concern is a minor one. Nevertheless, the numbers of covariates in the regression models were reduced based on analyses conducted outside of IVEware.

Candidate covariates considered for the imputation models were all available household-level variables from the surveys, household-level variables created from child-level variables (combining multiple children in a household for the 2001 NS-CSHCN), sampling design variables, and information about the household's telephone exchange provided by the GENESYS Sampling System database. (GENESYS is a proprietary product of Marketing Systems Group; available covariates are listed in Appendix A.) As noted earlier, any variables that were included in the imputation models but were missing in some survey records were imputed in IVEware simultaneous to the imputation of income and household size. These additional variables and imputed values were not retained in the final public-use data base for the NS-CSHCN or NSCH multiply imputed data.

Child-level variables from the 2001 NS-CSHCN

While the 2003 NSCH collected data on exactly one child per household, the 2001 NS-CSHCN potentially has data on multiple children per household. More specifically, the 2001 NS-CSHCN has screening and basic demographic data for all children in the household, with detailed health information for one child with special needs in nearly every household that has at least one such child (except for households with nonresponse following the screener) and detailed health insurance information for at least one child without special health care needs in nearly every household with such children (except for households with nonresponse following the screener). In other words, the 2001 NS-CSHCN has some data for every child in every household and detailed data for one or two children from nearly every household. Because income was assessed at the household level, we considered only household summaries of these child-level data. For example, we created race/ethnicity variables indicating if <u>any</u> child was Hispanic, if <u>any</u> child was African-American, and so on. As another example, mother's education differed for some children in the same 2001 NS-CSHCN household. We created two variables: mother's education for the child with the most educated mother and mother's education for the child with the lowest educated mother. We did not expect interview variables (such as whether any child in the household has an emotional/developmental/behavioral problem) to have much predictive power for income, especially for the 2001 NS-CSHCN because the 2001 NS-CSHCN interview was only completed in households with at least one child with special health care needs. Nevertheless, as part of our thorough search for possible predictors, we considered interview variables as possible predictors, and some are in the final models.

Sampling design variables for 2001 NS-CSHCN and 2003 NSCH

For the two surveys, a random-digit-dial sample of households with children under 18 years of age was selected from each of the 50 states and the District of Columbia. The sample designs generally yield stratified simple random samples of telephone numbers within Immunization Action Plan (IAP) areas. These IAP areas are whole states or portions of states (a city, county, or remainder). To represent the sampling information during imputation, then, we considered the state (some IAPs have very few cases) as a possible covariate, and we also included IAP-level and state-level income summary variables (mean and standard deviation of the quarter-root-transformed reported values) as possible covariates. To account for the differential sampling weights, we also considered the sampling weight as a possible covariate.

Results of Modeling

For the regression modeling, we wanted to use as many cases as possible, including those with known interval data. Therefore, we created a separate (transformed) income variable for modeling. To determine a value for those cases with some, but incomplete, income data, we used a simple "median imputation" technique. When the range was known, the midpoint was used; if only one constraint (upper or lower bound) was known, the median for fully observed data fulfilling this condition was used (e.g., the median income for respondents with income less than $20,000 is $12,000). None of the predictor variables were imputed prior to this modeling.

2001 NS-CSHCN Modeling

Appendix B shows all of the 156 variables considered for the 2001 NS-CSHCN multiple imputation model. Most variables were built from only one questionnaire item; one exception was the creation of three indicator variables for specific reasons why a child did not get needed care in the C4Q05 and C4Q06 series of questions: "any unmet need because service costs too much" (NOAFFORD); "any unmet need because of a health care problem" (HCPROB); and "any unmet need because of no insurance" (NOINS). Many of the questionnaire items were recoded before modeling. Most of the recoding was simply to recode "don't know" and "refused" responses to missing. There were some variables for which a logical skip was a yes or no, and we used top-coding or bottom-coding to merge levels with a few cases.

The table in Appendix B describes each variable considered for the model: whether it came from the 2001 NS-CSHCN household, screener, interview, or insurance datasets, the GENESYS data, or the sample design information; the type of bivariate analysis (with income) done; the F- or t-statistic; the degrees of freedom; the P-value; and the R-squared value. Categorical variables were analyzed using ANOVA (F-statistic), while continuous variables were analyzed using regression (t-statistic). Because the amount of data is large, even small effects are quite significant; only 19 of the 156 variables have a P-value greater than .0001. Therefore, a more useful tool for comparing the significance of different variables is the R-square statistic, which shows how much of the variability in household income (as transformed) is explained. The four variables most related to household income are all related to insurance and education and were derived from other variables in the survey: "none of the children with completed interviews had employer-based health insurance," "at least one child with completed interviewes received Medicaid benefits," mother's education for the child with the most educated mother, and mother's education for the child with the lowest educated mother. These four variables all have R-squared values above 0.18 (correlations of at least 0.42). The next ten variables, which all have an R-squared above 0.07 (correlations of at least 0.26), are all GENESYS variables. These

GENESYS variables are all related to income and education, except for one, which was the percentage of 45-54-year-old adults in the population served by the telephone exchange. Appendix B is sorted by R-squared, in descending order. It should be noted that the bivariate analyses are shown only for completeness; they are not central to the task since the regression model chooses the best predictors on a multivariate rather than bivariate basis.

The final model was not just made up of the variables with the highest R-squared values. In fact, three (GENESYS variables PHI2, PHI3, and PHI4) of the fourteen variables mentioned above with R-squared values of at least 0.07 are not in the final model. The final model is the best set of combined variables to explain the variability in household income. Some of the variables with the strongest bivariate relationships with household income were themselves related and/or did not add much explanatory power when combined.

Table 2 shows the variables chosen for the model by stepwise regression within SAS. The R-squared value of 0.4861 suggests a strong relationship between the observed (transformed) values of household income and the values predicted by the model. Note that the values predicted by the stepwise regression model were not the values that were imputed; the imputed values were drawn from the posterior distribution of household income based on the model derived from this regression.

2003 NSCH Modeling

Appendix C shows all of the 152 variables considered for the 2003 NSCH multiple imputation model. Most variables were built from only one questionnaire item; the exceptions were age of the oldest child (in four categories) and a variable indicating that the child did not receive at least some needed care (built from S4Q07, S4Q23, and S4Q17). Questionnaire items were excluded only if they were asked for a small sample or all levels but one were sparse (e.g., the "what specific teeth problems" questions satisfy both of these conditions). Questions from Sections 6 and 7 were excluded because the questions in these sections were asked only for children of particular ages. Many of the questionnaire items were recoded before modeling. Most of the recoding was simply to recode "don't know" and "refused" responses to missing. There were some variables for which a logical skip was a yes or no, and we used top-coding or bottom-coding to merge levels with few cases.

The table in Appendix C describes each variable considered for the model: whether it came from the 2003 NSCH questionnaire, the GENESYS data, or the sample design; the type of bivariate analysis (with income) done; the F- or t-statistic; the degrees of freedom; the P-value; and the R-squared value. Categorical variables were analyzed using ANOVA (F-statistic), while continuous variables were analyzed using regression (t-statistic). Because the amount of data is large, even small effects are quite significant; only 11 of the 152 variables have a P-value greater than .0001. Therefore, a more useful tool for comparing the significance of different variables is the R-squared statistic, which shows how much of the variability in household income (as transformed) is explained. The four variables most related to household income are all questionnaire items: S3Q02 (child enrolled in Medicaid or SCHIP), C11Q11B (any child in household receives free or reduced cost lunches at school), S1Q05A (highest level of education achieved by anyone in household), and C11Q11A (any child in household received food stamps). These four variables all have R-squared values above 0.18 (correlations of at least 0.42). Seven of the eight other

Table 2. Covariates used in multiple imputation for 2001 NS-CSHCN household income

Source	Covariate[a]
Design	IAP_MEAN
Design	IAP_STD
Design	STATE_STD
Design	WEIGHT_H
GENESYS	AVGRENT
GENESYS	CENSDIV1
GENESYS	CENSDIV2
GENESYS	CENSDIV3
GENESYS	CENSDIV4
GENESYS	CENSDIV7
GENESYS	DAYSAV
GENESYS	HOMEVAL
GENESYS	IAP03
GENESYS	IAP06
GENESYS	IAP10
GENESYS	IAP12
GENESYS	IAP13
GENESYS	IAP28
GENESYS	IAP29
GENESYS	IAP32
GENESYS	IAP35
GENESYS	IAP36
GENESYS	IAP43
GENESYS	IAP48
GENESYS	IAP52
GENESYS	IAP54
GENESYS	IAP62
GENESYS	IAP63
GENESYS	IAP67
GENESYS	IAP71
GENESYS	IAP73
GENESYS	MDYEDUC
GENESYS	MEDINC
GENESYS	MET2
GENESYS	MET3
GENESYS	PAGE2
GENESYS	PAGE4
GENESYS	PAGE5
GENESYS	PASIAN
GENESYS	PCOLGRAD
GENESYS	PHI1
GENESYS	PHI6
GENESYS	PHI7
GENESYS	PHI8
GENESYS	PWHITE
GENESYS	STATE03
GENESYS	STATE05
GENESYS	STATE06
GENESYS	STATE10
GENESYS	STATE15
GENESYS	STATE20
GENESYS	STATE21
GENESYS	STATE31
GENESYS	STATE32
GENESYS	STATE36
GENESYS	STATE38
GENESYS	STATE39
GENESYS	STATE41
GENESYS	STATE43
GENESYS	STATE44
GENESYS	STATE45
GENESYS	STATE46
GENESYS	STATE51
GENESYS	TIMEZ
GENESYS	TOTALHH
GENESYS	TOTALPOP
Household	C11Q14
Household	C11Q20
Household	SPANISH
Household	INT_LANG
Household	TOTKIDS
Household	TOTPERS
Household	C11Q11
Insurance	CHIPNAME
Insurance	MEDICAID
Insurance	MILITARY
Insurance	MOTHER_EDUCR (MOTHEDH2)
Insurance	MOTHER_EDUCR (MOTHEDH3)
Insurance	MOTHER_EDUCR (MOTHEDH4)
Insurance	MOTHER_EDUCR (MOTHEDL3)
Insurance	NATIVINS
Insurance	OTHERINS
Insurance	OTHERPUB
Insurance	PRIVATE
Insurance	SCHIP
Insurance	UNINS
Insurance	UNINS_YR
Insurance	WEIGHT_I
Insurance	YS_UNINS
Interview	C11Q12
Interview	C3Q10
Interview	C3Q11
Interview	C4Q03
Interview	C4Q05_01
Interview	C4Q05_02
Interview	C4Q05_03
Interview	C4Q05_09
Interview	C4Q06_01
Interview	C4Q06_0A
Interview	C5Q08
Interview	C6Q05
Interview	C8Q01_B
Interview	C8Q02
Interview	C8Q05
Interview	C9Q05
Interview	C9Q06
Interview	C9Q07
Interview	NOAFFORD
Screener	AGE_YEARS (OLDEST)
Screener	AGE_YEARS (YOUNG)
Screener	C1001_01
Screener	C1001_04
Screener	C1001_05
Screener	C1001_08
Screener	C1002_01
Screener	C1002_02
Screener	C1002_03
Screener	C1002_05
Screener	C1002_08
Screener	CALLYRF01 (2000 INTERVIEW)
Screener	CALLYRL02 (2002 INTERVIEW)
Screener	FACCT1
Screener	FACCT2
Screener	FACCT3
Screener	FACCT5
Screener	SEX (ALLFEM)
Screener	SEX (MIXGEND)

[a] See Appendix A for a description of the telephone exchange-level covariates and Blumberg et al. (2003) for a description of other covariates.

variables with an R-squared value above 0.10 (correlations of at least 0.32) are GENESYS variables; the one questionnaire item is IN_HH (number of adults living in household). The seven GENESYS variables are all related to income and education. Appendix C is sorted by R-squared, in descending order. Again, it should be noted that the bivariate analyses are shown only for completeness; they are not central to the task since the regression model chooses the best predictors on a multivariate rather than bivariate basis.

The final model was not just made up of the variables with the highest R-squared values. In fact, two (GENESYS variables PHI2 and PHI3) of the twelve variables mentioned above with R-squared values of at least 0.10 were not in the final model. The final model is the best set of combined variables to explain the variability in household income. Some of the variables with the strongest bivariate relationships with household income were themselves related and/or did not add much explanatory power when combined.

Table 3 below shows the variables chosen for the model by stepwise regression within SAS. The R-squared value for this model is 0.5415. The R-square value for this model is greater than the R-squared for the 2001 NS-CSHCN model (0.4861). Note that the values predicted by the stepwise regression model were not the values that were imputed; the imputed values were drawn from the posterior distribution of income based on the model derived from this regression.

Table 3. Covariates used in multiple imputation for 2003 NSCH household income

Source	Covariate[a]
Design	STATE_STD
GENESYS	AVGRENT
GENESYS	DMACNTY
GENESYS	HOMEVAL
GENESYS	MDYEDUC
GENESYS	MEDINC
GENESYS	NWBANKSN
GENESYS	PAGE1
GENESYS	PAGE4
GENESYS	PASIAN
GENESYS	PBLACK
GENESYS	PCOLGRAD
GENESYS	PERRENT
GENESYS	PHI4
GENESYS	PHI8
GENESYS	PHISP
GENESYS	PWHITE
GENESYS	STATE
GENESYS	TIMEZB
GENESYS	TOTALPOP
NSCH	AGE GRID (OLDESTCH)
NSCH	C11Q11
NSCH	C11Q11A
NSCH	C11Q11B
NSCH	C11Q20
NSCH	INCENTIVE_ PROTOCOL
NSCH	IN_HH (# PARENTS)
NSCH	NUM_PHON
NSCH	OUT_HH (# PARENTS)
NSCH	RACE
NSCH	S_UNDR18
NSCH	S10Q03
NSCH	S10Q04
NSCH	S10Q05
NSCH	S10Q06
NSCH	S11Q01
NSCH	S11Q02X01
NSCH	S11Q02X02
NSCH	S11Q02X03
NSCH	S11Q02X06
NSCH	S11Q05
NSCH	S11Q06
NSCH	S11Q08
NSCH	S1Q02
NSCH	S1Q05
NSCH	S1Q05A
NSCH	S1Q06
NSCH	S2Q01
NSCH	S2Q07
NSCH	S2Q13
NSCH	S2Q18
NSCH	S2Q19
NSCH	S2Q24
NSCH	S2Q40
NSCH	S2Q54
NSCH	S2Q56
NSCH	S3Q01
NSCH	S3Q02
NSCH	S3Q03
NSCH	S3Q04
NSCH	S4Q03
NSCH	S4Q07, S4Q23, S4Q17
NSCH	S4Q09
NSCH	S4Q15
NSCH	S4Q27
NSCH	S5Q08A
NSCH	S8Q03
NSCH	S8Q06
NSCH	S8Q08
NSCH	S8Q09
NSCH	S8Q10
NSCH	S8Q12
NSCH	S8Q13
NSCH	S8Q15
NSCH	S9Q00
NSCH	S9Q08
NSCH	S9Q15
NSCH	S9Q15C
NSCH	S9Q18
NSCH	S9Q34
NSCH	SPANISH

[a] See Appendix A for a description of the telephone exchange-level covariates and Blumberg et al. (2005) for a description of other covariates.

User's Guide

This section comprises a user's guide to using the multiply-imputed values derived using the above procedures. The two datasets contain the same variables and are structured the same way, so procedures for processing and analyzing the data will be the same regardless of which data set is used. This user's guide is split into three sections: general guidelines for using the data, general guidelines for analyzing the data using SAS, and specific guidelines with SAS and SUDAAN code examples for analyzing the data using SUDAAN.

The derived imputed poverty level (POVLEVEL_I) variable that is available for public use was calculated from the imputed household income and household size.[1] The household income and household size have been imputed five times, so the resulting imputed data set contains five times as many observations as were in the original data set. For the 2001 NS-CSHCN, the datasets have 5(196,888) = 984,440 records, while the 2003 NSCH datasets have 5(102,353) = 511,765. Each imputation is distinguished by the SAS variable IMPUTATION. Therefore, each IDNUMR appears five times in the file, with IMPUTATION having values of 1, 2, 3, 4, and 5 corresponding to the five separate imputations.

General Guidelines

There are three possible ways to analyze the data, and <u>we will also describe one invalid way to use the data that should not be attempted</u>.

Taking the possible ways first, a complete-case (only) analysis is the simplest, which uses only the cases with observed values. This can be done by using the poverty level variables (POVLEVEL in the NS-CSHCN file and POVERTY_LEVELR in the NSCH file) in the public use files. Any analysis using these variables could be biased due to nonresponse, and the variability will be larger because of the missing values.

The second possible way of using the data is to use only a single imputation from the multiple imputation files. Each of the five imputations has been drawn from a valid distribution based on a regression model, but this model and the distribution are slightly different for each imputation. To analyze only one imputation, choose only the subset of cases with IMPUTATION = *c*, where *c* is 1, 2, 3, 4, or 5. Single imputation analyses result in estimated standard errors that are too small because the imputed values are treated as if they were observed. This ignores the inherent uncertainty resulting from lack of knowledge about the true (unobserved) value, but is superior to the complete-case analysis. It should be noted that slightly different results will be obtained depending on which subset of cases is chosen, but no subset is superior to another.

The statistically valid way to analyze the data is to analyze all five imputed datasets together. To do this, five separate analyses are conducted; one on each of the five imputed datasets. These analyses are then combined following the standard multiple imputation combining rules (Rubin,

---

[1] The public use data files for the NS-CSHCN and the NSCH do not include household income, to protect against inadvertent disclosure of survey subjects' identities. Only poverty level is reported on the public use data files. Similarly, imputed household income will not be released as public use data. Researchers interested in accessing the original and imputed household income data may access the data through the NCHS Research Data Center.

1987). This is superior to the previous two methods. Since this is more complex, brief instructions for analyzing the data using SAS are given first. Following that, a more detailed explanation with sample code provides instructions for analyzing the data using SAS-callable SUDAAN.

It is very important to note that it is invalid to combine the five imputed values into one analysis. For example, taking the average poverty level (which might not be an integer) to derive one "average" poverty status value per case is invalid. Poverty status must be analyzed as a multiply imputed variable with SAS, SUDAAN, IVEware, or another appropriate statistical software package to make use of the multiply-imputed data.

Regardless of the statistical software used to analyze the data, one must merge the survey data from the public use analysis files (the household, screener, interview, or insurance files if using the NS-CSHCN data or the single interview file if using the NSCH data) with the data from the multiple imputation file by the unique household identifier (IDNUMR). To combine these files, we first need to sort by IDNUMR and then merge using this identifier as our merge variable. To improve the efficiency and speed at which the files are processed, it is important to subset the analysis files by keeping only the variables we are interested in analyzing. We do this by using a KEEP statement as part of our data set options or within the data step.

Analyzing the data with SAS

Prior to running any SAS procedures to analyze the combined file, it is very important to have the dataset sorted by IMPUTATION since analyses of the multiply imputed data need to be done separately by IMPUTATION. Separate analyses are specified in SAS by using the procedure option keyword BY ("BY IMPUTATION;" should be one line within the analysis).

The two basic steps to using the multiply-imputed data are to 1) analyze the data separately by IMPUTATION as if each were a separate data set, and 2) combine the results from the different imputed data sets using PROC MIANALYZE. In the first step, separate analyses are done with options set to keep the covariances that are needed to combine the analyses. Then, PROC MIANALYZE combines these different analyses using the standard multiple imputation combining rules (Rubin, 1987). For more information on the use of PROC MIANALYZE, please refer to Yuan (undated) or see the *SAS/STAT User's Guide*.

Analyzing the data with SAS-Callable SUDAAN

Starting with SUDAAN version 9, one of the new features incorporated into various SUDAAN procedures is the ability to analyze multiply imputed data sets. Sample programs to analyze both survey data sets using SUDAAN are available online:

> http://www.cdc.gov/nchs/about/major/slaits/cshcn.htm (for NS-CSHCN)
> http://www.cdc.gov/nchs/about/major/slaits/nsch.htm (for NSCH)

The following instructions will highlight important concepts and syntax from these programs.

Step 1: Input and Subset Interview File

The first step is to select the variables from the survey analysis files we are interested in analyzing in conjunction with the poverty level data. Using data from the NSCH as an example, we subset our data using a KEEP statement:

```
data ansch;
 set data1.nschpuf3(keep = idnumr state racer s11q01 poverty_levelr s3q01
 s3q02 weight_i);
```

Step 2: Recode and prepare analytical variables

To process categorical variables in SUDAAN, variable levels must not contain zero and must increase in consecutive integers. Variables with “no” and “yes” responses in the NSCH are coded as “0” and “1” respectively and must be recoded. We also need to code “don’t know” and “refused” responses to missing. The following code demonstrates one way to do this:

```
if s3q01 in (.L,.M,.P,6,7) then s3q01 = .;
 s3q01 = s3q01 + 1;

poverty200 = .;
if 1 <= poverty_levelr <= 5 then poverty200 = 1;
 else if 6 <= poverty_levelr <= 8 then poverty200 = 2;
```

The variable S3Q01 is a NO = 0 and YES = 1 variable that we have coded to NO = 1 and YES = 2. In addition to the recode, we created a derived variable using the original poverty level variable with the missing values. This variable will take a value of 1 for households with poverty levels less than 200% FPL and a value of 2 for households with poverty levels greater than or equal to 200% FPL.

Step 3: Sort the survey data set by IDNUMR

To merge the interview file to the multiple imputation file, we must first sort the interview file by the unique household identifier.

```
proc sort data = ansch;
 by IDNUMR;
run;
```

Step 4: Input the multiple imputation poverty level file

```
data imp;
 set data2.nsch03mimp;
   if 1 <= povlevel_i <= 5 then poverty200i = 1;
    else if 6 <= povlevel_i <= 8 then poverty200i = 2;
format povlevel_f yn. povlevel_i pov. poverty200i povb.;
run;
```

After inputting this file, we do not need to sort by IDNUMR because the data set is already sorted by this variable. We created a derived variable using the imputed income variable. This variable collapses the original poverty level variable to households with poverty levels less than 200% FPL and households with poverty levels greater than or equal to 200% FPL. This variable will be used in the next section where we compare the results between the three valid methods for analyzing the data.

Step 5: Merge the interview file with the imputation file and create five output files

For SUDAAN to process the files correctly, it expects to have a separate analytical file for each of the five imputations. To merge and create these files in one data step, we can use the following code:

```
DATA
 ansch_mimp1
 ansch_mimp2
 ansch_mimp3
 ansch_mimp4
 ansch_mimp5;
  MERGE ansch(in = one) imp(in = two);
   BY idnumr;
    IF one and two;
    if imputation = 1 then output ansch_mimp1;
    if imputation = 2 then output ansch_mimp2;
    if imputation = 3 then output ansch_mimp3;
    if imputation = 4 then output ansch_mimp4;
    if imputation = 5 then output ansch_mimp5;
run;
```

The DATA statement creates five output data sets (*ansch_mimp1-ansch_mimp5*). We need to be sure to use a naming convention for our output files that uses a numeral at the end of the filename to specify the imputation number because SUDAAN will need this later to know which data sets to input in our procedure. The MERGE statement identifies the interview data set (*ansch*) and the imputation data set (*imp*) as the files to merge, and the BY statement identifies the unique household identifier we want to merge by (IDNUMR). The IF statement makes sure we select records that are contained in both the interview file and the imputation file (which will be all records in this merge). Once the merge is completed, we use the values from the IMPUTATION variable to separate our combined data set into 5 smaller data sets. Each data set has one record for each record in the interview file, so each data set has 102,353 records.

Step 6: Sort all five data sets by STATE and IDNUMR

Prior to analyzing data using any of the SUDAAN procedures, all of our data sets must be sorted by the stratum (which is the STATE variable) and the primary sampling unit (which is the unique household identifier or the IDNUMR variable).

```
proc sort data = ansch_mimp1 out = data3.ansch_mimp1;
 by state idnumr;
run;

proc sort data = ansch_mimp2 out = data3.ansch_mimp2;
 by state idnumr;
```

```
run;

proc sort data = ansch_mimp3 out = data3.ansch_mimp3;
 by state idnumr;
run;

proc sort data = ansch_mimp4 out = data3.ansch_mimp4;
 by state idnumr;
run;

proc sort data = ansch_mimp5 out = data3.ansch_mimp5;
 by state idnumr;
run;
```

These statements sort all of the temporary analytical files by STATE and IDNUMR and then output permanent data sets using the same naming convention.

Step 7: Analyze the five data sets

To analyze our data using the five imputation files, we need to add the MI_COUNT command to our SUDAAN procedure call. For example,

```
proc crosstab data = data3.ansch_mimp1 design=wr mi_count=5;
 nest state idnumr;
 weight weight_i;
 subgroup eth_race poverty_levelr s3q01 s3q02;
 levels         5              8     2     2;
 tables poverty_levelr * (eth_race s3q01 s3q02);
run;
```

The MI_COUNT command tells SUDAAN how many imputation files to expect. In our case, we have 5. In the data statement, we identify our first permanent data set (*ansch_mimp1*). By identifying the first data set along with the number of imputation data sets, we are instructing SUDAAN to use the data sets *ansch_mimp1-ansch_mimp5* in the CROSSTAB analysis.

Multiple Imputation Diagnostics

In this section, we compare the multiple imputation output with similar analyses using only the complete cases (no imputation) and a single imputation. Using data from the 2003 NSCH, crosstab comparisons are made using data from the poverty status variable (collapsed to groups with income less than 200% FPL and greater than or equal to 200% FPL) along with the variable assessing whether the child currently receives insurance through Medicaid or the State Children's Health Insurance Program (S3Q02). To carry out the complete-case analysis, the poverty status variable from the original public use file (POVLEVEL in the CSHCN files and POVERTY_LEVELR in the NSCH file) was used to create a collapsed poverty level variable (POVERTY200). For the single and multiple imputation analysis, we used the imputed poverty level variable (POVLEVEL_I) to create a collapsed poverty level variable (POVERTY200I; see Step 4). To carry out the single imputation analysis, any of the individual imputation files created in the previous program could have been used (though the results may differ across the five imputation files).

To generate crosstabs with standard errors, the following code was used for the complete cases.

```
proc crosstab data = data3.public_use_file design=wr;
 nest state idnumr;
 weight weight_i;
 subgroup  poverty200 s3q02;
 levels             2     2;
 tables poverty200 * s3q01;
 rtitle “Poverty Level by Public Insurance Status”;
run;
```

To generate crosstabs with standard errors, the following code was used for the single imputation cases.

```
proc crosstab data = data3.ansch_mimp1 design=wr;
 nest state idnumr;
 weight weight_i;
 subgroup  poverty200i s3q02;
 levels              2     2;
 tables poverty200i * s3q01;
 rtitle “Poverty Level by Public Insurance Status”;
run;
```

To generate crosstabs with standard errors, the following code was used for the multiple imputation cases.

```
proc crosstab data = data3.ansch_mimp1 design=wr mi_count=5;
 nest state idnumr;
 weight weight_i;
 subgroup  poverty200i s3q02;
 levels              2     2;
 tables poverty200i * s3q01;
 rtitle “Poverty Level by Public Insurance Status”;
run;
```

Table 4 compares crosstabs and standard errors for each data set using each of the three analysis methods.

Table 4. Percents and Standard Errors for Three Analysis Methods of Household Income Relative to Poverty and Children's Health Insurance Status, 2003 NSCH

Household poverty level and child health insurance status	Complete Cases Only % (SE)	Single Imputation % (SE)	Multiple Imputation % (SE)
Less than 200% FPL			
Does not receive public insurance	35.2 (0.52)	35.3 (0.49)	35.3 (0.50)
Receives public insurance	64.8 (0.52)	64.7 (0.49)	64.7 (0.50)
200% FPL or greater			
Does not receive public insurance	92.1 (0.20)	91.0 (0.21)	91.0 (0.23)
Receives public insurance	7.9 (0.20)	9.0 (0.21)	9.0 (0.23)

For estimates about children living in households with income less than 200% FPL, the standard errors based on multiple imputation are less than the standard errors based on only complete cases (because we have obtained additional information out of variable relationships and this information reduces some of the uncertainty in the estimate), but greater than the standard errors based on single imputation (because single imputation ignores one component of the estimate's variability). For estimates about children living in households with income 200% FPL or greater, the standard errors based on multiple imputation are actually greater than the standard errors based on only complete cases, but this is due to the magnitude of the estimated percentages. Standard errors for proportions are related to the estimated proportion,[2] and are smaller when the proportion is closer to 0% or 100%. In this example, 91.0% is further from 100% than is 92.1%; this increase leads to a larger standard error and outweighs the gain in efficiency by using the cases with missing values.

References

Blumberg, S. J., Olson, L., Frankel, M., Osborn, L., Becker, C. J., Srinath, K. P., and Giambo, P. (2003). "The Design and Operation of the National Survey of Children with Special Health Care Needs, 2001," Vital and Health Statistics, Series 1, Number 41.

Blumberg, S. J., Olson, L., Frankel, M., Osborn, L., Srinath, K. P., and Giambo, P. (2005). "The Design and Operation of the National Survey of Children's Health, 2003," Vital and Health Statistics, Series 1, Number 43.

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

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

Paulin, G. D. and Sweet, E. M. (1996). "Modeling Income in the U.S. Consumer Expenditure Survey," Journal of Official Statistics, Vol. 12, pp. 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, Vol. 27, pp. 75-85.

Raghunathan, T. E., Solenberger, P., and Van Hoewyk, J. (2002). "IVEware: Imputation and Variance Estimation Software User's Guide," available on the web at: ftp://ftp.isr.umich.edu/pub/src/smp/ive/ive_user.pdf

Rubin, D. B. (1978). "Multiple Imputation in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse," 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, Vol. 91, pp. 473-489.

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

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, Vol. 101, pp. 924-933.

[2] Under simple random sampling, the standard error is the square root of p(1-p)/n, where p is the proportion and n is the sample size. Values of p closer to 0% or 100% lead to smaller standard errors.

Yuan, Y. (undated). “Multiple Imputation for Missing Data: Concepts and New Development,” SAS Institute document P267-25, available online at: http://support.sas.com/rnd/app/papers/multipleimputation.pdf

APPENDIX A. Telephone Exchange-Level Covariates Available for Imputation Modeling Through GENESYS Sampling System

Covariate Name	Covariate Description
ACTUALHH	Listed Number of Households in Exchange
AVGRENT	Median Rent
BELLTYPE	Exchange Type (residential only or shared)
CENSDIV	Census Division (9)
DAYSAV	Daylight Saving Code
DMACNTY	Designated Market Area County Size (4 levels)
HH_DENS	Household Density (Persons per HH)
HOMEVAL	Median Home Value
IAPNUM	Immunization Action Plan Area Number
MDYEDUC	Median Years Education
MEDINC	Median Household Income
MEST_STATUS	Alternative Metropolitan Status Code
MET	Metropolitan Status Code (5 levels)
MET_STATUS	Alternative Metropolitan Status Code
NWBANKS	Number of Active (>0 listed households) working banks in Exchange
NXXCNT	Number of Exchanges Assigned to County
PAGE1	Percent Aged 0-17
PAGE2	Percent Aged 18-24
PAGE3	Percent Aged 25-34
PAGE4	Percent Aged 35-44
PAGE5	Percent Aged 45-54
PAGE6	Percent Aged 55-64
PAGE7	Percent Aged 65+
PASIAN	Percent Asian Population
PBLACK	Percent African-American Population
PCOLGRAD	Percent College Graduates
PEROWNER	Percent Owners
PERRENT	Percent Renters/Others
PHI1	Percent with Household Income between $0-$10,000
PHI2	Percent with Household Income between $10,000-$15,000
PHI3	Percent with Household Income between $15,000-$25,000
PHI4	Percent with Household Income between $25,000-$35,000
PHI5	Percent with Household Income between $35,000-$50,000
PHI6	Percent with Household Income between $50,000-$75,000
PHI7	Percent with Household Income between $75,000-$100,000
PHI8	Percent with Household Income above $100,000
PHISP	Percent Hispanic
PWHITE	Percent White
STATE	State Abbreviation
TIMEZ	Time Zone
TOTALHH	Estimated Total Households in Exchange (including unlisted)
TOTALPOP	Estimated Total Population in Exchange

APPENDIX B. Covariates Considered for Imputation of Household Income, 2001 NS-CSHCN

Covariate[1]	Source	Test	F-Statistic	t-Statistic	DF[2]	P-value	R-squared
PRIVATE	Insurance	ANOVA	52157.20		1	<.0001	0.2250
MEDICAID	Insurance	ANOVA	42443.50		1	<.0001	0.1911
MOTHER_EDUCR (MOTHEDH)	Insurance	ANOVA	9802.78		4	<.0001	0.1872
MOTHER_EDUCR (MOTHEDL)	Insurance	ANOVA	9790.71		4	<.0001	0.1870
MEDINC	GENESYS	Regress		160.03	1	<.0001	0.1248
PHI8	GENESYS	Regress		149.23	1	<.0001	0.1103
PHI3	GENESYS	Regress		-143.83	1	<.0001	0.1033
PHI1	GENESYS	Regress		-141.36	1	<.0001	0.1001
PHI2	GENESYS	Regress		-140.50	1	<.0001	0.0990
MDYEDUC	GENESYS	Regress		140.40	1	<.0001	0.0989
PHI7	GENESYS	Regress		139.12	1	<.0001	0.0973
PCOLGRAD	GENESYS	Regress		133.37	1	<.0001	0.0901
PHI4	GENESYS	Regress		-125.12	1	<.0001	0.0802
PAGE5	GENESYS	Regress		119.00	1	<.0001	0.0731
AVGRENT	GENESYS	Regress		116.1	1	<.0001	0.0698
C11Q11	Household	ANOVA	13171.80		1	<.0001	0.0685
HOMEVAL	GENESYS	Regress		110.22	1	<.0001	0.0633
C11Q14	Household	ANOVA	10607.00		1	<.0001	0.0559
SPANISH	Household	ANOVA	9301.98		1	<.0001	0.0492
NOT ENGLISH	Household	ANOVA	9123.31		1	<.0001	0.0483
SCHIP	Insurance	ANOVA	8639.22		1	<.0001	0.0459
UNINS_YR	Insurance	ANOVA	8413.55		1	<.0001	0.0447
C1002_01	Screener	ANOVA	8288.81		1	<.0001	0.0441
PHI6	GENESYS	Regress		89.26	1	<.0001	0.0425
C1001_01	Screener	ANOVA	7039.84		1	<.0001	0.0377
IAP_MEAN	Design	Regress		81.51	1	<.0001	0.0357
IAPNUM	GENESYS	ANOVA	86.25		77	<.0001	0.0357
C1002_02	Screener	ANOVA	5799.93		1	<.0001	0.0313
MET	GENESYS	ANOVA	1424.54		4	<.0001	0.0307
C11Q20	Household	ANOVA	5638.24		1	<.0001	0.0305
NOPHONE (in HH)	Household	ANOVA	5629.48		1	<.0001	0.0305
UNINS	Insurance	ANOVA	5306.36		1	<.0001	0.0287
PWHITE	GENESYS	Regress		71.60	1	<.0001	0.0277
STATE_MEAN	Design	Regress		68.55	1	<.0001	0.0255
STATE	GENESYS	ANOVA	93.95		50	<.0001	0.0255
YS_UNINS	Insurance	Regress		-66.72	1	<.0001	0.0245
PAGE4	GENESYS	Regress		64.45	1	<.0001	0.0226
PHI5	GENESYS	Regress		-62.96	1	<.0001	0.0216
DMACNTY	GENESYS	ANOVA	1294.01		3	<.0001	0.0212
C1001_02	Screener	ANOVA	3736.71		1	<.0001	0.0204
C6Q04	Interview	ANOVA	695.04		1	<.0001	0.0194
C1002_08	Screener	ANOVA	3501.73		1	<.0001	0.0191
PBLACK	GENESYS	Regress		-58.37	1	<.0001	0.0186
PAGE1	GENESYS	Regress		-56.53	1	<.0001	0.0175
PERRENT	GENESYS	Regress		-51.96	1	<.0001	0.0148
PEROWNER	GENESYS	Regress		51.96	1	<.0001	0.0148
PHISP	GENESYS	Regress		-50.76	1	<.0001	0.0141
PAGE6	GENESYS	Regress		47.66	1	<.0001	0.0125
CENSDIV	GENESYS	ANOVA	252.28		9	<.0001	0.0125
C6Q06	Interview	ANOVA	380.92		1	<.0001	0.0106
C1001_03	Screener	ANOVA	1921.34		1	<.0001	0.0106
C6Q05	Interview	ANOVA	379.14		1	<.0001	0.0106
C11Q12	Interview	ANOVA	1843.84		1	<.0001	0.0102
AGE_YEARS (YOUNG)	Screener	Regress		38.52	1	<.0001	0.0082
ACTUALHH	GENESYS	Regress		37.38	1	<.0001	0.0077

Covariate[1]	Source	Test	F-Statistic	t-Statistic	DF[2]	P-value	R-squared
TOTPERS	Household	Regress		33.23	1	<.0001	0.0061
AGE_YEARS (OLDEST)	Screener	Regress		32.25	1	<.0001	0.0058
C9Q07	Interview	ANOVA	1028.49		1	<.0001	0.0057
C1002_03	Screener	ANOVA	1005.53		1	<.0001	0.0056
FACCT3	Screener	ANOVA	943.23		1	<.0001	0.0052
C12Q2	Interview	ANOVA	744.09		1	<.0001	0.0052
TOTALHH	GENESYS	Regress		29.29	1	<.0001	0.0047
C9Q10	Interview	ANOVA	807.87		1	<.0001	0.0045
C9Q05	Interview	ANOVA	792.49		1	<.0001	0.0044
TIMEZ	GENESYS	ANOVA	150.52		5	<.0001	0.0042
PAGE7	GENESYS	Regress		-26.71	1	<.0001	0.0040
C1001_10	Screener	ANOVA	714.94		1	<.0001	0.0040
C3Q11	Interview	ANOVA	236.63		3	<.0001	0.0039
FACCT5	Screener	ANOVA	698.31		1	<.0001	0.0039
TOTALPOP	GENESYS	Regress		26.23	1	<.0001	0.0038
C4Q03	Interview	ANOVA	675.10		1	<.0001	0.0037
C3Q10	Interview	Regress		-24.61	1	<.0001	0.0034
C1001_04	Screener	ANOVA	606.89		1	<.0001	0.0034
TITLEV	Insurance	ANOVA	596.24		1	<.0001	0.0033
C8Q05	Interview	ANOVA	575.70		1	<.0001	0.0032
C12Q3	Interview	ANOVA	459.37		1	<.0001	0.0032
FACCT2	Screener	ANOVA	568.50		1	<.0001	0.0032
C9Q06	Interview	ANOVA	506.24		1	<.0001	0.0028
WEIGHT_H	Design	Regress		22.35	1	<.0001	0.0028
C1001_07	Screener	ANOVA	480.10		1	<.0001	0.0027
C3Q02	Interview	ANOVA	478.79		1	<.0001	0.0027
C1002_05	Screener	ANOVA	451.45		1	<.0001	0.0025
CHIPNAME	Insurance	ANOVA	208.82		2	<.0001	0.0023
NOAFFORD	Interview	ANOVA	399.50		1	<.0001	0.0022
C1001_06	Screener	ANOVA	386.19		1	<.0001	0.0021
FACCT4	Screener	ANOVA	367.42		1	<.0001	0.0020
C4Q06_01	Interview	ANOVA	356.27		1	<.0001	0.0020
C4Q06_02	Interview	ANOVA	314.95		1	<.0001	0.0018
PASIAN	GENESYS	Regress		16.85	1	<.0001	0.0016
C4Q07	Interview	ANOVA	279.61		1	<.0001	0.0016
C4Q05_06	Interview	ANOVA	276.73		1	<.0001	0.0015
C3Q13	Interview	ANOVA	261.32		1	<.0001	0.0015
C5Q08	Interview	ANOVA	254.54		1	<.0001	0.0014
C4Q05_03	Interview	ANOVA	248.89		1	<.0001	0.0014
HCPROB	Interview	ANOVA	245.80		1	<.0001	0.0014
PAGE2	GENESYS	Regress		-14.22	1	<.0001	0.0011
C4Q05_08	Interview	ANOVA	190.01		1	<.0001	0.0011
C8Q04	Interview	ANOVA	169.66		1	<.0001	0.0010
C1001_09	Screener	ANOVA	168.98		1	<.0001	0.0009
WEIGHT_I	Insurance	Regress		12.88	1	<.0001	0.0009
C4Q05_05	Interview	ANOVA	155.87		1	<.0001	0.0009
C8Q01_B	Interview	ANOVA	150.45		1	<.0001	0.0009
INT_LANG	Household	ANOVA	144.32		1	<.0001	0.0008
NOINS	Interview	ANOVA	141.62		1	<.0001	0.0008
C4Q05_01	Interview	ANOVA	137.86		1	<.0001	0.0008
C4Q06_03	Interview	ANOVA	138.05		1	<.0001	0.0008
C8Q02	Interview	ANOVA	126.30		1	<.0001	0.0007
C1001_05	Screener	ANOVA	101.42		1	<.0001	0.0006
CALLYRF	Screener	ANOVA	48.74		2	<.0001	0.0005
CALLYRL	Screener	ANOVA	46.57		2	<.0001	0.0005
C4Q05_14	Interview	ANOVA	79.82		1	<.0001	0.0004
C6Q03	Interview	ANOVA	78.53		1	<.0001	0.0004
OTHERINS	Insurance	ANOVA	75.41		1	<.0001	0.0004
C4Q05_02	Interview	ANOVA	75.28		1	<.0001	0.0004

Covariate[1]	Source	Test	F-Statistic	t-Statistic	DF[2]	P-value	R-squared
PAGE3	GENESYS	Regress		-8.38	1	<.0001	0.0004
NATIVINS	Insurance	ANOVA	71.25		1	<.0001	0.0004
C4Q05_10	Interview	ANOVA	65.07		1	<.0001	0.0004
OTHERPUB	Insurance	ANOVA	64.41		1	<.0001	0.0004
C8Q06	Interview	ANOVA	60.34		1	<.0001	0.0003
C1002_04	Screener	ANOVA	58.69		1	<.0001	0.0003
SEX (MIXGEND)	Screener	ANOVA	55.87		1	<.0001	0.0003
TOTKIDS	Household	Regress		-7.72	1	<.0001	0.0003
C9Q02	Interview	ANOVA	48.32		1	<.0001	0.0003
C4Q06_0A	Interview	ANOVA	37.12		1	<.0001	0.0002
C4Q05_13	Interview	ANOVA	33.24		1	<.0001	0.0002
C5Q07	Interview	ANOVA	31.47		1	<.0001	0.0002
C4Q05_04	Interview	ANOVA	28.24		1	<.0001	0.0002
C1002_06	Screener	ANOVA	26.25		1	<.0001	0.0001
NM_SP	Household	Regress		-5.08	1	<.0001	0.0001
NM_NSP	Household	Regress		-4.65	1	<.0001	0.0001
TOTKIDSM	Household	Regress		-4.13	1	<.0001	0.0001
TOTKIDSF	Household	Regress		-4.51	1	<.0001	0.0001
NM_SPM	Household	Regress		-4.56	1	<.0001	0.0001
IAP_STD	Design	Regress		-4.71	1	<.0001	0.0001
SINGLINS	Insurance	ANOVA	17.10		1	<.0001	0.0001
FACCT1	Screener	ANOVA	15.57		1	<.0001	0.0001
SEX (ALLFEM)	Screener	ANOVA	14.92		1	0.0001	0.0001
NM_NSPF	Household	Regress		-3.48	1	0.0005	0.0001
SEX (ALLMALE)	Screener	ANOVA	11.01		1	0.0009	0.0001
C8Q01_C	Interview	ANOVA	10.49		1	0.0012	0.0001
C4Q05_11	Interview	ANOVA	8.09		1	0.0045	<0.0001
FLAGSEC8	Interview	ANOVA	6.55		1	0.0105	<0.0001
NEEDTYPE	Screener	ANOVA	4.75		1	0.0293	<0.0001
NEEDTYPE	Interview	ANOVA	4.65		1	0.0311	<0.0001
C1002_07	Screener	ANOVA	1.99		1	0.1587	<0.0001
C1001_08	Screener	ANOVA	1.84		1	0.1744	<0.0001
MILITARY	Insurance	ANOVA	0.55		1	0.4565	<0.0001
DAYSAV	GENESYS	ANOVA	0.40		1	0.5267	<0.0001
C4Q05_09	Interview	ANOVA	0.33		1	0.5649	<0.0001
MO_FLAG	Interview	ANOVA	0.29		1	0.5918	<0.0001
UNKINS	Insurance	ANOVA	0.28		1	0.5977	<0.0001
NM_SPF	Household	Regress		-2.79	1	0.0052	<0.0001
NM_NSPM	Household	Regress		-2.01	1	0.0442	<0.0001
STATE_STD	Design	Regress		-1.02	1	0.3088	<0.0001
C8Q01_A	Interview	ANOVA	0.07		1	0.7973	<0.0001
HH_STATUS	Household	ANOVA	0.02		1	0.8821	<0.0001

[1] See Appendix A for a description of the telephone exchange-level covariates and Blumberg et al. (2003) for a description of other covariates.

[2] The degrees of freedom associated with the denominator of the F-ratio was greater than or equal to 179,500.

APPENDIX C. Covariates Considered for Imputation of Household Income, 2003 NSCH

Covariate[1]	Source	Test	F-Statistic	t-Statistic	DF[2]	P-value	R-squared
S3Q02	NSCH	ANOVA	28686.40		1	<.0001	0.2298
C11Q11B	NSCH	ANOVA	26353.00		1	<.0001	0.2159
S1Q05A	NSCH	ANOVA	6632.05		4	<.0001	0.2152
C11Q11A	NSCH	ANOVA	22116.40		1	<.0001	0.1859
MDYEDUC	GENESYS	Regress		121.73	1	<.0001	0.1325
MEDINC	GENESYS	Regress		118.89	1	<.0001	0.1271
PCOLGRAD	GENESYS	Regress		117.06	1	<.0001	0.1237
PHI8	GENESYS	Regress		116.16	1	<.0001	0.1221
PHI3	GENESYS	Regress		-112.5	1	<.0001	0.1154
IN_HH (# PARENTS)	NSCH	ANOVA	6124.59		2	<.0001	0.1122
PHI2	GENESYS	Regress		-106.56	1	<.0001	0.1047
PHI4	GENESYS	Regress		-104.09	1	<.0001	0.1004
S9Q15C	NSCH	ANOVA	10090.30		1	<.0001	0.0997
PHI1	GENESYS	Regress		-103.48	1	<.0001	0.0994
S9Q34	NSCH	ANOVA	10408.40		1	<.0001	0.0970
AVGRENT	GENESYS	Regress		98.49	1	<.0001	0.0909
PHI7	GENESYS	Regress		98.29	1	<.0001	0.0905
HOMEVAL	GENESYS	Regress		96.36	1	<.0001	0.0873
S11Q08	NSCH	ANOVA	9223.92		1	<.0001	0.0870
S9Q08	NSCH	ANOVA	2697.53		3	<.0001	0.0815
C11Q11	NSCH	ANOVA	7316.57		1	<.0001	0.0703
S9Q18	NSCH	ANOVA	2141.55		3	<.0001	0.0658
PAGE4	GENESYS	Regress		79.28	1	<.0001	0.0608
S1Q06	NSCH	ANOVA	3131.78		2	<.0001	0.0607
SPANISH	NSCH	ANOVA	6064.43		1	<.0001	0.0588
S2Q01	NSCH	ANOVA	1905.62		3	<.0001	0.0556
S2Q54	NSCH	ANOVA	1782.69		3	<.0001	0.0556
PHI5	GENESYS	Regress		-75.08	1	<.0001	0.0549
S11Q02X01	NSCH	ANOVA	5338.09		1	<.0001	0.0524
PAGE5	GENESYS	Regress		68.58	1	<.0001	0.0462
S10Q01	NSCH	ANOVA	2298.99		2	<.0001	0.0461
OUT_HH (# PARENTS)	NSCH	ANOVA	2320.53		2	<.0001	0.0457
S11Q01	NSCH	ANOVA	4607.79		1	<.0001	0.0454
S1Q05	NSCH	ANOVA	875.72		5	<.0001	0.0432
S10Q06	NSCH	ANOVA	2142.39		2	<.0001	0.0425
S10Q03	NSCH	ANOVA	2042.91		2	<.0001	0.0410
S2Q56	NSCH	ANOVA	1317.93		3	<.0001	0.0393
S9Q01	NSCH	ANOVA	3439.40		1	<.0001	0.0390
IAP_MEAN	Design	Regress		62.30	1	<.0001	0.0385
S8Q13	NSCH	ANOVA	1256.85		3	<.0001	0.0375
S4Q09	NSCH	ANOVA	3560.96		1	<.0001	0.0355
S10Q07	NSCH	ANOVA	1126.32		2	<.0001	0.0341
S8Q11	NSCH	ANOVA	3395.31		1	<.0001	0.0339
MET	GENESYS	ANOVA	834.09		4	<.0001	0.0332
S11Q02X02	NSCH	ANOVA	2937.77		1	<.0001	0.0295
STATE	GENESYS	ANOVA	58.98		50	<.0001	0.0295
STATE_MEAN	Design	Regress		54.32	1	<.0001	0.0295
S5Q01	NSCH	ANOVA	2925.25		1	<.0001	0.0293
DMACNTY	GENESYS	ANOVA	945.58		3	<.0001	0.0284
C11Q20	NSCH	ANOVA	2706.75		1	<.0001	0.0272
PWHITE	GENESYS	Regress		51.69	1	<.0001	0.0268
NOPHONE	NSCH	ANOVA	2640.11		1	<.0001	0.0265
S10Q04	NSCH	ANOVA	844.43		3	<.0001	0.0263
S9Q11B	NSCH	ANOVA	2290.77		1	<.0001	0.0262
S10Q05	NSCH	ANOVA	1199.59		2	<.0001	0.0245
S8Q03	NSCH	ANOVA	343.93		7	<.0001	0.0242
S5Q08A	NSCH	ANOVA	2392.88		1	<.0001	0.0241

Covariate[1]	Source	Test	F-Statistic	t-Statistic	DF[2]	P-value	R-squared
MEST_STATUS	GENESYS	ANOVA	702.97		3	<.0001	0.0213
S1Q02	NSCH	ANOVA	694.15		3	<.0001	0.0210
S5Q02	NSCH	ANOVA	878.70		2	<.0001	0.0208
S3Q01	NSCH	ANOVA	2034.58		1	<.0001	0.0206
S10Q02	NSCH	ANOVA	971.39		2	<.0001	0.0201
S11Q03	NSCH	ANOVA	1819.44		1	<.0001	0.0196
PERRENT	GENESYS	Regress		-43.72	1	<.0001	0.0193
PEROWNER	GENESYS	Regress		43.72	1	<.0001	0.0193
S8Q09	NSCH	ANOVA	884.83		2	<.0001	0.0180
PHISP	GENESYS	Regress		-41.94	1	<.0001	0.0178
PBLACK	GENESYS	Regress		-41.07	1	<.0001	0.0171
S8Q12	NSCH	ANOVA	555.29		3	<.0001	0.0169
NUM_PHON	NSCH	ANOVA	810.91		2	<.0001	0.0164
PHI6	GENESYS	Regress		38.35	1	<.0001	0.0149
CENSDIV	GENESYS	ANOVA	158.35		9	<.0001	0.0145
S3Q04	NSCH	ANOVA	1392.41		1	<.0001	0.0142
S9Q15	NSCH	ANOVA	1283.27		1	<.0001	0.0139
S11Q06	NSCH	ANOVA	337.81		4	<.0001	0.0138
S5Q06	NSCH	ANOVA	1326.41		1	<.0001	0.0135
S8Q07	NSCH	ANOVA	634.00		2	<.0001	0.0129
PAGE2	GENESYS	Regress		-33.26	1	<.0001	0.0113
S8Q14	NSCH	ANOVA	356.83		3	<.0001	0.0109
S4Q01	NSCH	ANOVA	1022.99		1	<.0001	0.0105
S5Q04	NSCH	ANOVA	436.62		2	<.0001	0.0104
ACTUALHH	GENESYS	Regress		31.01	1	<.0001	0.0098
S2Q59	NSCH	ANOVA	743.14		1	<.0001	0.0076
S4Q03	NSCH	ANOVA	687.85		1	<.0001	0.0071
NWBANKS	GENESYS	Regress		26.28	1	<.0001	0.0071
S2Q10	NSCH	ANOVA	668.15		1	<.0001	0.0068
S2Q23	NSCH	ANOVA	651.23		1	<.0001	0.0067
TIMEZ	GENESYS	ANOVA	130.18		5	<.0001	0.0067
S8Q15	NSCH	ANOVA	296.10		2	<.0001	0.0061
S5Q07	NSCH	ANOVA	589.53		1	<.0001	0.0060
MET_STATUS	GENESYS	ANOVA	293.18		2	<.0001	0.0060
S_UNDR18	NSCH	ANOVA	192.79		3	<.0001	0.0059
S11Q02X03	NSCH	ANOVA	570.43		1	<.0001	0.0059
PAGE1	GENESYS	Regress		-24.06	1	<.0001	0.0059
S4Q04	NSCH	ANOVA	560.93		1	<.0001	0.0058
S3Q03	NSCH	ANOVA	508.02		1	<.0001	0.0053
S4Q06	NSCH	ANOVA	512.55		1	<.0001	0.0053
S8Q10	NSCH	ANOVA	243.87		2	<.0001	0.0050
S8Q08	NSCH	ANOVA	236.76		2	<.0001	0.0049
S2Q16	NSCH	ANOVA	470.33		1	<.0001	0.0048
AGE GRID (OLDESTCH)	NSCH	ANOVA	154.29		3	<.0001	0.0047
S11Q02X05	NSCH	ANOVA	434.14		1	<.0001	0.0045
PASIAN	GENESYS	Regress		20.85	1	<.0001	0.0045
WEIGHT_I	Design	Regress		21.04	1	<.0001	0.0045
S11Q05	NSCH	ANOVA	392.99		1	<.0001	0.0040
S2Q18	NSCH	ANOVA	380.72		1	<.0001	0.0039
TOTALHH	GENESYS	Regress		19.34	1	<.0001	0.0038
PAGE7	GENESYS	Regress		-18.94	1	<.0001	0.0037
S4Q07	NSCH	ANOVA	293.72		1	<.0001	0.0030
TOTALPOP	GENESYS	Regress		16.70	1	<.0001	0.0029
S2Q07	NSCH	ANOVA	255.51		1	<.0001	0.0026
S2Q44	NSCH	ANOVA	254.62		1	<.0001	0.0026
S2Q13	NSCH	ANOVA	233.20		1	<.0001	0.0024
S5Q09	NSCH	ANOVA	180.97		1	<.0001	0.0022
S2Q42	NSCH	ANOVA	200.27		1	<.0001	0.0021
S4Q07, S4Q23, S4Q17	NSCH	ANOVA	208.41		1	<.0001	0.0021

Covariate[1]	Source	Test	F-Statistic	t-Statistic	DF[2]	P-value	R-squared
S4Q27	NSCH	ANOVA	164.53		1	<.0001	0.0018
S8Q06	NSCH	ANOVA	88.86		2	<.0001	0.0018
S9Q00	NSCH	ANOVA	160.64		1	<.0001	0.0017
INCENTIVE_ PROTOCOL	NSCH	ANOVA	76.12		2	<.0001	0.0016
NXXCNT	GENESYS	Regress		12.37	1	<.0001	0.0016
INCENTIVE_ GROUP	NSCH	ANOVA	145.28		1	<.0001	0.0015
S2Q41	NSCH	ANOVA	137.57		1	<.0001	0.0014
S4Q23	NSCH	ANOVA	125.24		1	<.0001	0.0013
CALLDATE	NSCH	ANOVA	114.22		1	<.0001	0.0012
S2Q19	NSCH	ANOVA	107.84		1	<.0001	0.0011
S2Q22	NSCH	ANOVA	107.15		1	<.0001	0.0011
S4Q15	NSCH	ANOVA	105.36		1	<.0001	0.0011
S10Q08	NSCH	ANOVA	74.64		1	<.0001	0.0008
PAGE6	GENESYS	Regress		8.93	1	<.0001	0.0008
STATE_STD	Design	Regress		9.04	1	<.0001	0.0008
S2Q20	NSCH	ANOVA	67.57		1	<.0001	0.0007
RACE	NSCH	ANOVA	59.89		1	<.0001	0.0006
S2Q21	NSCH	ANOVA	52.69		1	<.0001	0.0005
S2Q37	NSCH	ANOVA	50.58		1	<.0001	0.0005
S11Q02X04	NSCH	ANOVA	36.63		1	<.0001	0.0004
BELLTYPE	GENESYS	ANOVA	9.41		4	<.0001	0.0004
S2Q38	NSCH	ANOVA	33.00		1	<.0001	0.0003
S2Q24	NSCH	ANOVA	21.91		1	<.0001	0.0002
S11Q02X06	NSCH	ANOVA	17.32		1	<.0001	0.0002
HH_DENS	GENESYS	Regress		-4.20	1	<.0001	0.0002
S2Q26	NSCH	ANOVA	6.61		1	0.0101	0.0001
S11Q02X07	NSCH	ANOVA	5.16		1	0.0231	0.0001
DAYSAV	GENESYS	ANOVA	3.04		1	0.0813	<0.0001
S5Q10	NSCH	ANOVA	2.86		1	0.0909	<0.0001
S2Q40	NSCH	ANOVA	1.61		1	0.2048	<0.0001
S2Q35	NSCH	ANOVA	0.72		1	0.3972	<0.0001
S2Q39	NSCH	ANOVA	0.51		1	0.4746	<0.0001
PAGE3	GENESYS	Regress		-0.35	1	0.7274	<0.0001
S1Q01	NSCH	ANOVA	0.02		1	0.8935	<0.0001
IAP_STD	Design	Regress		0.10	1	0.9204	<0.0001
S2Q04	NSCH	ANOVA	<0.01		1	0.9773	<0.0001

[1] See Appendix A for a description of the telephone exchange-level covariates and Blumberg et al. (2005) for a description of other covariates.

[2] The degrees of freedom associated with the denominator of the F-ratio was greater than or equal to 63,851.