The Pennsylvania State University

The Graduate School

The College of the Liberal Arts

THE PRACTICAL CONSEQUENCES OF IMPUTATION STRATEGY ON CHILDREN'S HEALTH INSURANCE COVERAGE ESTIMATES IN THE 2007 CURRENT POPULATION SURVEY ANNUAL SOCIAL AND ECONOMIC SUPPLEMENT

A Thesis In

Sociology and Demography

by

Rebekah Lynn Young

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Arts

August 2009

The thesis of Rebekah Lynn Young was reviewed and approved* by the following:

David R. Johnson
Associate Professor of Sociology, Demography, and Human Development and Family Studies
Thesis Advisor

Melissa Hardy
Professor of Human Development and Family Studies and Sociology

Paul Amato
Distinguished Professor of Sociology and Demography

Barrett Lee
Professor of Sociology and Demography, Program Chair

*Signatures are on file in the Graduate School

Abstract

For decades the Census Bureau has handled item-level incomplete data by imputing the missing values using hot deck procedures. These procedures have come under increasing criticism for yielding biased population and subpopulation estimates and for underestimating the amount of uncertainty in the imputed values. In this paper I compare estimates based on the Census Bureau's hot deck imputation to estimates from a multiple imputation procedure using data on children's health insurance coverage from the 2007 Current Population Survey Annual Social and Economic Supplement. This comparative analysis addresses three questions. First, what are the theoretical advantages and disadvantages of HD and MI for this particular data? Second, does the choice of imputation procedure change the state-level estimates of the number of uninsured children and how might this difference impact social policy? Finally, what are the potential substantive consequences of different imputation strategies for social science researchers who use the ASEC's children's health insurance variables in multivariate analyses? I find that while HD and MI produce different point estimates of the number of uninsured children by state, the consequences of imputation strategy for researchers who rely on these data to answer substantive questions may be minimal.

Table of Contents

List of Tables

List of Figures

Introduction

The presence of missing data due to item level non-response is a common problem in survey research. For decades the Census Bureau has handled item-level incomplete data by imputing the missing values using hot deck procedures. These procedures have come under increasing criticism for yielding biased population and subpopulation estimates and for underestimating the amount of uncertainty in the imputed values. When used to make policy decisions or allocate federal funds reliance on these potentially biased estimates has important consequences. Alternative ways of accounting for the effect of the missing data that could yield less biased estimates have been proposed but have not been rigorously tested. In this paper I compare estimates based on the Census Bureau's hot deck (HD) imputation to estimates from a multiple imputation (MI) procedure using data on children's health insurance coverage from the 2007 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC). This comparative analysis addresses three questions. First, what are the theoretical advantages and disadvantages of HD and MI for this particular data? Second, does the choice of imputation procedure change the state-level estimates of the number of uninsured children, and how might these differences impact social policy? Finally, what are the broader potential substantive consequences of using different imputation strategies for social science researchers who rely on the ASEC's children's health insurance variables for multivariate analyses?

Background

Recent declines in health insurance coverage have generated broad public and political interest. The quality of data on health insurance coverage has important implications for research, policy, and social welfare. The ASEC is the most widely cited source of health

insurance statistics in the U.S. (Fisher and Turner 2003). For researchers, the ASEC is often used as the "gold standard" by which investigators measure the quality of their data and develop population estimates for applying weights to survey data (Groves 2006). For example, the National Survey of American Families uses ASEC estimates as external validation for their sample distribution of earnings (NSAF). Researchers also use ASEC data to produce state and county level estimates of various social and economic characteristics such as health insurance coverage to examine a wide range of substantive questions.

ASEC data are the official source used to estimate the number of uninsured children and number of children living in poverty in each state (Davern, Beebe, Blewett, and Call 2003). The state level estimates of the number of uninsured children is a key component of federal allocation formulas that distribute $3-4 billion of federal funds to the State Children's Health Insurance Program (SCHIP) (Census Bureau 2005). The Centers for Medicare and Medicaid Services report that more than 6.6 million children were enrolled in the SCHIP programs at some point during 2006 (SCHIP). On February 4, 2009, President Obama signed into law the Children's Health Insurance Program Reauthorization Act of 2009 (CHIRPA), a new law that will allocate $32.8 billion to states over the next four and a half years to cover an additional 4 million uninsured children, illustrating the increasingly important role that estimates of the uninsured will play in allocation of federal dollars during the next few years.

In the 2007 ASEC 13.2% of the items comprising the variables indicating whether or not children have health insurance contain missing data. Item non-response is a common cause of missing data in survey research and researchers have implemented a variety of imputation strategies to deal with this issue (Allison 2000; Schafer 1997; Schafer and Graham 2002). Imputing missing values is a process of replacing a missing (unknown) value with a plausible

estimate. This method of dealing with missing data is generally regarded as preferable to options such as complete case analysis, which limits analysis to the subset of cases with no missing information, or mean substitution, which assigns to each missing case the average value observed from complete cases (Allison 2000).

The ASEC employs three principal imputation methods, relational imputation, longitudinal edits, and HD allocation (CPS 2003). Relational imputation assigns values for blank or inconsistent responses on the basis of other characteristics on the person's record or information from other members of the household. Longitudinal edits (primarily used for labor force edits) look at a previous month's data to replace the missing value. Finally, the method used to replace the health insurance variables of primary interest for this paper, HD allocation assigns responses for missing data to sample persons with information from matched sample persons with similar demographic and economic information who answered the same questions. HD imputation techniques assign actually observed values from a non-missing record, called the donor, to a record with a missing value, called the recipient. Donors and recipients are matched on key demographic variables such as age, sex, employment status and other characteristics of the household. Replacing a missing value with a value that actually occurred in the dataset is generally considered an advantage of the method. All missing data within the health insurance variables are replaced with HD imputation.

There are many types of HD procedures, which are differentiated by how the donor case is chosen (Huisman 2000). The HD imputation used by the ASEC is conducted in a logical and deliberate sequence. Data are sorted by state and primary sampling unit (PSU) such that missing values are typically allocated from geographically related areas. For example, missing values for records in Oregon are not likely to be matched to observed records for Pennsylvania. This

distinction is critical due to the geographic clustering of labor force and industry and occupation characteristics. Values are not imputed for inappropriate or illogical entries and out-of-range values are not permitted. For a more detailed description of the ASEC HD procedures used by the Census Bureau for health insurance coverage see Davern et. al 2007.

The HD procedure used by the Census Bureau has long been a subject of criticism. For example, Rubin shows that the CPS HD underestimates income by 7 percent and Lillard et al. suggest that the CPS HD underestimates wages and salary by 73 percent (Lillard, Smith, and Welch 1986). The ASEC health insurance coverage estimates have continued to be scrutinized through recent years, with many researchers finding that the ASEC estimates of people without health insurance coverage are higher than those found in other large surveys (Bennefield 1996; Fronstin 2000; Lewis, Elwood, and Czajka 1998). Focusing on HD imputation methodology as a cause of this difference, research by Davern et al. (2003) show that the ASEC imputation method leads to bias in estimating health insurance coverage for subpopulations at the state level, such as low-income uninsured children (Davern, Beebe, Blewett, and Call 2003). This bias is not surprising given that when missing values are replaced by imputed values a single imputation will generally underestimate variability (Rubin 1987). In addition to this weakness, a reasonably large sample size is required for a HD method to work properly. Small scale estimates from the ASEC HD imputation, such as those used to generate estimates for a state-level sub-population of uninsured children for example, may be especially problematic due to the small sample size of available donors. Although the ASEC is a large scale survey of more than 78,300 households nationwide, the state-specific sample sizes vary widely from approximately 900 interviewed households in Arkansas to 5,600 in California.

The ubiquity and influence of the ASEC estimates illustrate the importance of having

accurate estimates of health insurance coverage. Davern et al. show that the HD imputation is a source of bias for the health insurance coverage estimates (2004). Little and Rubin argue that an MI method produces a more accurate imputation of missing data and Davern et al. suggest that MI may be a preferable method that the Census Bureau should explore (2002; 2004). The Census Bureau recognizes that the state-level estimates are insufficient for many policy purposes and has recently made a number of revisions in an attempt to improve the quality of the estimates for insured and uninsured data. The 2007 March ASEC (with results for health insurance data for calendar year 2006) reflects these changes. The Census Bureau has called for external research proposals to provide insights on the best way to impute missing items for surveys to help improve their results. Specifically, they are interested in studies that discuss imputation techniques and how they can be applied to Census Bureau data and an assessment of MI (Census Bureau 2006).

There are some reasons to believe that MI will offer better estimates than a HD procedure. Research on the theoretical properties of HD methods is sparse, especially compared to MI which is strongly theoretically grounded (Little and Rubin 2002; Marker, Judkins, and Winglee 2002). Whereas a single imputation is likely to underestimate the error variances and provide biased significance tests, MI does not (Little and Rubin 2002; Schafer and Graham 2002). The HD procedure is severely limited by the number of categories and variables that can be used to create the deck. A complete listing of the variables used in the ASEC HD are included in Table 1, though it is important to realize that only a subset of these variables are included for the imputation of each variable. For example, the imputation of whether or not a person has group health coverage uses two decks. First, people are divided into groups of workers and non-workers. Workers are allocated based on Age1, Family Relation, Class of Worker, Earnings

Level and Firm Size. Non-workers are allocated based on Age1, Government Health Coverage and Family Relation. The maximum number of variables included in any particular ASEC deck is six and the minimum is two. Effective HD imputations should match the donor on as many characteristics as possible, but reliance on too many characteristics may result in too few matches and donors must be used who are less similar. The constraints imposed by sample size also lead to arbitrary categorization of variables and recoding of the informing variables so that much of the information and variance is lost. For example, the categorization of variables such as age is necessary for enough matches to occur in the deck, but this approach reduces age variance in the absence of any clear rationale. Further, if the matching categories used do not represent all the important correlates of the variable being imputed then the relationship between the imputed values and other variables in the data can be distorted. For instance, in the example above, marital status may be an important correlate of whether or not a person is covered by group health insurance but is not included in the deck. The Census does not state a logical reason as to why particular sets of variables are included in each allocation specification. MI does not operate under the same constraints as a HD procedure because hundreds of variables can theoretically be taken into consideration in the imputation process. A more informed imputation may provide more accurate estimates.

On the other hand, MI may not provide as good of an estimate as a HD procedure as it involves model assumptions about the distribution of the variables. Since a specific statistical model is used to generate the estimates, the quality of the estimates is dependent on the accuracy of the model and validity of underlying assumptions. For a correct MI, the model used to generate the estimates must be at least as complete as the multivariate models used to analyze the data (Allison 2000; Johnson 2007; Schafer and Graham 2002). For example, if the data analyst

tests for statistical interactions or non-linear relationships among the variables, failure to take these multiplicative and non-linear relationships into account in the imputation model can lead to attenuation of the estimates of these effects. HD imputation does not suffer from this limitation. Although MI can contain a greater range of variables, present technology does not allow it to incorporate the range of interactions that are accounted for by a HD procedure. For example, the ASEC HD used to impute health insurance contains an inherent interaction between age, employment, class of work, earnings, firm size, and spouse's labor force status due to the cross-categorization among these variables (Census Bureau 2007). With a sample this large, current software such as SPSS and Stata ICE cannot accommodate an interaction this complex and still converge. In this respect, the HD strategy may contain more information but on a smaller set of variables with a smaller number of categories than the MI procedure.

An important consideration in selecting a strategy to impute missing data is how consistently the method can yield plausible estimates that do not bias results gleaned from the data. All imputed variables in the ASEC data have been "flagged" such that imputed values can be distinguished from reported values. This advantage allows me to address the following research questions: (1) Is there a statistically significant difference between HD and MI state-level estimates of children's health insurance coverage? How might this impact SCHIP funding? (2) Is the difference between imputation methods sufficiently large to potentially cause different substantive conclusions for social science researchers?

Data and Methods

The CPS is a monthly survey of 50,000 or more households conducted by the Census Bureau for the Bureau of Labor Statistics mainly to estimate the unemployment rate (Census

Bureau 2005). The ASEC is a supplement to the CPS that is conducted annually in the month of March. The March CPS supplement contains approximately 78,000 households and includes detailed health insurance questions asked of the household respondent for every household resident (Census Bureau 2005). Respondents are asked about health insurance coverage for the previous calendar year for themselves and for all other household members. I use the 2007 file which describes health insurance coverage for all or part of 2006. The Census Bureau distinguishes between private and government health insurance. Private health insurance is provided by an employer or union or can be privately purchased and unrelated to employment. Government health insurance includes Medicare, Medicaid, military health insurance, health insurance from somebody outside the household and "other". Respondents are asked separate questions about each type of health insurance and asked to answer yes or no for each type. Those who answer "no" have their answers verified. People are considered insured if they were covered at any time during the year.

I construct a variable indicating whether or not children in the household were covered by health insurance by combining three questions asking whether or not children in the household were covered by health insurance of someone in the household, covered by health insurance of someone outside the household, or covered by any other type of health insurance. Davern et al. point out that the documentation of the children's health insurance coverage is unclear about the imputation of coverage for all dependents in the family and has alerted the Census Bureau to this problem (2004). I follow the logic used by Davern et al. to deal with this limitation and only treat those who were allocated to have a family policy as missing cases. I create a single flag indicating whether or not each case included an imputed value on health insurance and use the flag to set the values imputed by the HD to missing. Overall, 13.2% of the responses indicating

whether or not children were covered by health insurance contained imputed values. Percentage of data imputed by state is included in Table 3. Connecticut, Florida, New York, Vermont and Nebraska had the highest percent of imputed data, with each of these states having more than 17 percent missing responses. Montana, Oklahoma, Alabama, Idaho and Arkansas had the lowest percent of imputed data, with each of these states having less than 9 percent missing responses.

The MI model I construct was designed to maximize the amount of information included in the estimation within the limits imposed by available software. The model included all household members and every question the ASEC survey asked regarding details of health insurance coverage. In addition to health insurance coverage variables I included 124 auxiliary variables. This auxiliary information included all variables that the Census Bureau used in their HD matrices, all variables that were included in the example multivariate model in Table 4, and a set of variables that were chosen because they were highly correlated with children's health insurance coverage. I used Stata ICE to perform the imputations (Royston 2005). The model was constructed under the fully normal assumption except for the three variables regarding children's health insurance, which were imputed using a logistic model. I ran 200 burn-in iterations, 100 between-dataset iterations and used 5 datasets.

Results

After removing the values allocated by the HD and replacing them with values imputed with MI, I compared the results in two ways. First, I compared state level estimates of insurance rates. The need for accurate state level estimates reaches far beyond SCHIP funding (Blewett and Davern 2006). Still, there are direct policy implications of the estimate of uninsured children per state for SCHIP funding because this number is a key component of the formula used to allocate

money to states. Second, I explored whether differences in these estimates had substantive consequences for social science researchers who use these data in multivariate models. The model I selected was used for derivative theoretical reasons, simply as a model representing variables that are commonly used when modeling determinants of children's health insurance coverage, assuming researchers' choice of these variables was theoretically informed (Huang 1997; Sommers 2005). The purpose of this model was to see if the differences in the parameters generated when using data imputed by HD versus MI would lead the researcher to different substantive conclusions.

Estimates of children's insurance coverage

A comparison of the estimates of uninsured children by state based on HD and MI is presented in Table 3 and Figure 1. I used a two-tailed t-test of difference in proportions, testing the difference between the values imputed using the HD compared to those imputed with MI. MI produced statistically significant different values ($p < .001$) compared to the HD estimates for 33 states. The difference was not significant for 18 states. The MI estimates almost universally estimated a smaller percentage of children being uninsured, with the exception of Hawaii, Tennessee, Arkansas, South Carolina, Montana and Michigan. Of these states, only the difference between Montana and Hawaii was statistically significant. For example, the HD method estimates 28.0% of the children living in Hawaii do not have health insurance where MI estimates that the number of uninsured children in Hawaii is 28.8%.

The largest observed differences occurred for Virginia, Nebraska, Mississippi, Florida, Utah and Arizona, where MI produced estimates between 2 and 3 percentage points lower than HD estimates of uninsured children. In Arizona, for example, the HD method estimates that 45.6% of children are uninsured where MI estimates 42.8%. Based only on these differences in

percentages the consequences of the imputation method may seem relatively minor. Small differences in percentage points, however, may have large dollar implications for federal funding that uses these estimates in the allocation formulas. Figure 1 shows states that would have received more, less, and similar funding as a result of the strategy for handling missing data. The total number of uninsured children shown in Figure 1 is based on the state level estimates of the population less than 18 yeas of age in 2000 (Meyer 2001). The number of children that each percentage point comprises is important when thinking about health care services. Applying these estimates to California, for example, using the MI approach over 70,000 fewer children would be estimated to be uninsured when compared to estimates from the data imputed by the HD method.

Multivariate Analysis

The model presented in Table 4 shows that in a logistic regression model of child health insurance coverage the differences between HD and MI estimates of the odds ratios were negligible. The dependent variable was whether or not a child has any form of health insurance (0=not covered; 1=covered). The independent variables were selected because these variables have been frequently used in conjunction with models examining children's health care by social science researchers. Age represents the age of the child and ranges from 0-17 years of age. Males were compared to females as the omitted category. All race categories were compared to whites and indicate the race of the child. Hispanic was included as a mutually exclusive category where any respondent who replied 'yes' to having Hispanic or Latino ethnicity was coded as Hispanic. Any respondent with one or more race marked was placed into the mixed race category. Region was based on Census Bureau definitions. Family characteristics included family income, whether

or not the family lives in public housing, number of months the family received food stamps, and number of people in the family.

Providing one example with negligible difference between the approaches is insufficient evidence to conclude the choice of imputation strategy does not matter for multivariate analyses. To explore the differences more broadly, I tested 24 different versions of this model (not shown here), changing how the variable for children's health care was used (e.g. dependent variable versus independent variable, restricting the analysis to specific type of insurance coverage, such as Medicaid) as well as a trying a variety of predictor variables. In all models tested, the regression coefficients from multivariate analyses of the HD and MI imputed data sets were so similar that substantive conclusions would not have differed.

Discussion

The research I present here contributes to filling a demand for a clear comparison of HD methods to MI methods using the ASEC data. The application of the ASEC imputation of health insurance coverage has potential policy implications particularly for SCHIP funding. At a national level, MI estimates just over 3 million more children to be insured than HD estimates, consistent with Davern et al.'s (2003) finding that the HD leads to an overestimation of uninsured children in the ASEC data.

My research compared estimates based on two different imputation approaches but the findings do not necessarily support an argument that the MI approach produces more or less "correct" estimates than the HD. Statisticians such as Little and Rubin (2002) have shown that MI is more likely to yield estimates that more accurately take into account the uncertainty introduced by the imputation than those from a single imputation method. Because the true value

in the population is not known, however, it is not possible to say which is the more accurate. Further, both missing data techniques compared here handle only item level missingness. Nonresponse to the whole survey may be as high as 17 percent of people in the CPS, typically adjusted by weighting the data. No rigorous comparisons have been made of weighting and MI approaches to respondents missing completely from the survey.

There are several important limitations of this comparison. The set of variables I include in the MI model are not ideal. First, I was unable to include state of residence in the model. The amount of computer time it takes to run an imputation model increases exponentially as the number of variables increases. I estimated that including 50 additional parameters to incorporate the states would have caused the imputation to take approximately six weeks to run with no guarantee that the estimates would converge. Although Davern et al. (2003) argue that part of the reason the Census HD is biased is due to its inability to account for geographic location, this could not be completely overcome with the MI approach. This may be a limitation because we would expect that state policy would impact the proportion of children who have health insurance. A compromise was to include region of the country to help capture some of the geographic location correlates. Region could easily be included because only four parameters needed to be added to the model. The current software limitations, however, may not pose the same dilemmas in the future. The second limitation is that although all the individual variables from the HD are included in the MI model, MI software cannot support the range of interactions between these variables that is an inherent part of the HD. These two limitations mean that the MI model is more informed than the HD in some respects but not in others. On the other hand, the MI model does not require that continuous variables such as age, education, income, etc. are represented by a small number of categories as is required in an HD approach. Another possible

limitation relates to the way I handled missing data in variables other than those that made up the health insurance measures. Many of these other variables had their missing values imputed with the HD approach. Because some of the variables I used in the MI imputation had imputed HD values, this may have lead to different estimates than might have been obtained had I removed HD imputed values for all variables. Because the MI needs to be at least as informed as the HD imputation, I would have had to use all the variables included in the hot deck models for each of the variables which would likely have taxed the capacity of the MI software.

The real goal of the research was to establish whether or not a MI technique gives significantly different estimates than the currently used HD technique. The strategy used by the Census Bureau to handle missing data has important policy consequences, clearly illustrated by the difference in the estimates of uninsured children. However, substantive researchers should feel safe using the HD estimates in multivariate models as we cannot expect that MI will produce different results.

References

Allison, Paul D. 2000. *Missing Data*: Thousand Oaks.

Bennefield, R. L. 1996. "A Comparative Analysis of Health Insurance: Data from CPS and SIPP." in *Joint Statistical Meetings of the American Statistical Association*. Chicago: U.S. Census Bureau.

Blewett, Lynn A. and Michael Davern. 2006. "Meeting the need for state-level estimates of health insurance coverage: Use of state and federal survey data." *Health Services Research* 41:946-975.

Census Bureau, U.S. 2005. "Health Insurance Overview." vol. 2007. Washington D.C. : U.S. Census Bureay.

—. 2006. "Potential Research Data Center Methodological Topics." Washington, D.C. : U.S. Census Bureau.

—. 2007. *Health Insurance Edit and Imputation Specifications for the Current Population Survey*. U.S. Census Bureau, Housing and Household Economic Statistics Division, Washington, D.C. .

CPS. 2003. "Current Population Survey Technical Paper 63: Design and Methodology." edited by U. S. C. B. f. B. o. L. Statistics. Washington: U.S. Census Bureau.

Davern, Michael, Timothy J. Beebe, Lynn A. Blewett, and Kathleen Thiede Call. 2003. "Recent changes to the Current Population Survey: Sample expansion, health insurance verification, and state health insurance coverage estimates." *Public Opinion Quarterly* 67:603-626.

Fisher, Robin and Joanna Turner. 2003. "Health Insurance Estimates for Counties." in *2003 Joint Statistical meetings - Section on Survey Research Methods*: U.S. Census Bureau.

Fronstin, P. 2000. *Counting the Uninsured: A comparison of National Surveys. Employee Benefit Research Institute Brief No. 225*. Washington, D.C.: The Employee Benefit Research Institute.

Huang, Fung-Yea. 1997. "Health insurance coverage of the children of immigrants in the United States." *Maternal and Child Health* 1:69-80.

Huisman, Mark. 2000. "Imputation of Missing Item Responses: Some simple techniques." *Quality and Quantity* 34:355-351.

Johnson, David R. 2007. "Infertility: Pathways and psychosocial outcomes." Grant Application submitted to Department of Health and Human Services.

Lewis, K., M. Elwood, and J. Czajka. 1998. *Counting the Uninsured: A review of the literature.* Washington, D.C.: The Urban Institute.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. "What do we really know about wages? The importance of nonreporting and census imputation." *The Journal of Political Economy* 94:489-506.

Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. Hoboken, New Jersey: John Wiley & Sons, Inc. .

Marker, David A., David R. Judkins, and Marianne Winglee. 2002. "Large scale imputation for complex surveys." in *Survey Non-Response*. New York: Wiley.

Meyer, Julie. 2001. *Age: 2000* Washington, D.C. : U.S. Census Bureau, Economics and Statistics Administration.

NSAF. 1997 and 1999. "National Survey of America's Families Methodology Reports." The Urban Institute Web site.

Royston, Patrick. 2005. "Multiple imputation of missing values." *Stata Journal* 4:227-241.

Rubin, Donald B. 1983. "Imputing income in the CPS: Comments on "Measures of aggregate labor cost in the United States." Pp. 333-344 in *The measurement of labor cost*, vol. 48, *Studies in income and wealth*, edited by J. E. Triplett. Chicago: University of Chicago Press.

Rubin, Donald B. 1987. *Multiple imputation for nonresponse in surveys*. New York: Chichester.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate data*. New York: Chapman and Hall.

Schafer, Joseph L and John W. Graham. 2002. "Missing Data: Our view of the state of the art." *Psychological Methods* 7:147-177.

SCHIP. 2007. *FY 2007 Annual Enrollment Report*. Baltimore, MD: Centers for Medicare and Medicaid Services.

Sommers, Benjamin. 2005. "From Medicaid to uninsured: Drop-out among children in public insurance programs." *Health Services Research* 40:59-78.

Appendix

Table 1. Range of Variables Included in the CPS ASEC Health Insurance Edit and Imputation Specifications

Age1	
	15-24
	25-34
	35-44
	45-64
	65+
Age2	
	Less than 15
	15-24
	25-44
	45-64
	65+
Children	
	One or more own children under 18
	All others
Class of Worker	
	Self-employed
	All others
Earnings Level	
	Under $2000
	$2000-14999
	$15000-29999
	$30,000 or more
Family Relation	
	Reference person (w/relatives) or spouse
	Child or other relative
	Unrelated individual
Government Health Coverage	
	Covered by Medicare, Medicaid, or CHAMPUS
	All others
Group Health Coverage	
	Covered by employers-provided health plan
	All others
Marital Status	
	Married
	Never Married
	Divorced or Separated
	Widowed
Poverty Status	
	Received public assistance or SSI
	All others
Privately Purchased Health Coverage	
	Covered by privately purchased health plan
	All others

Table 1. (continued)

Size of Employer Firm	
	Under 25 employees
	25-499 employees
	500-999 employees
	1000 or more employees
Social Security Income	
	Received Social Security
	All others
Spouse Employment Status	
	NIU (non married)
	Spouse worked last year
	Spouse did not work last year
Veteran Status	
	Veteran
	Non-Veteran
	Current Armed Forces, or longest job last year was AF
Work/Disability Status	
	Worked last year
	Did not work last year - disabled
	All others

Table 2. Comprehensive List of Variables Included in MI Procedure

Age in years
Armed Forces, ever served (yes/no)
Children's health insurance (yes/no), # children covered by
    Insurance of someone not in household
    Medicare
    Other health insurance
Discouraged worker (yes/no)
Educational Attainment
    Children, Less than 1st Grade, 1st through 4th grade, 5th or 6th grade, 7th or 8th grade, 9th grade, 10th grade, 11th grade, 12th grade (no diploma), High school graduate, Some college but no degree, Associates degree in college, Bachelor's degree, Master's degree, Professional school degree, Doctorate degree
Ethnicity
    Hispanic, Spanish or Latino, yes or no
Family Income
Family size (number of persons)
Family Type
    Primary family, nonfamily householder, related subfamily, unrelated subfamily, secondary individual
Food; number of children who ate hot lunch at school
Full/Part-time Status
    Children or Armed Forces, Full-time schedules, Part-time for economic reasons usually FT, Part-time for non-economic reasons usually FT, Part-time for economic reasons usually PT, Unemployed, Not in labor force
Health care coverage (was anyone in the household covered by)
    Employment or union based health insurance coverage (yes/no)
    Private health insurance (yes/no)
    Coverage from outside the household (yes/no)
    Medicare (yes/no)
    Medicaid (yes/no)
    State sponsored health insurance plan (yes/no)
    Other health insurance including CHAMPUS, CHAMPVA, VA or military health care (yes/no)
Health Status
    Self-rated, 1-5
Hours worked last week
Immigrant nativity (in years)
Immigrant Status (yes/no)
Labor Force Status
    Children, Armed Forces, Working, With job not at work, Unemployed looking for work, Unemployed on layoff, Not in labor force
Marital Status
    Married, Never married, Divorced or separated, Widowed
Metropolitan Status
Number of people in the family
Number of own children
    Less than 6 years of age
    Less than 18 years of age
Number of people employeed by employer
Number of weeks spouse worked last year
Poverty Level (ratio of family income to poverty level)

Table 2. (continued)

Poverty Status
Public assistance
- Education assistance (yes/no)
- Transportation (yes/no)
- Child care services (yes/no)
- Public housing (yes/no)
- Job assistance (yes/no)
- Food stamps
  - Recipient (yes/no)
  - Value in dollars
  - Number of children covered
  - Number of months covered
- Food programs
  - Free lunch (yes/no)
  - Reduced lunch (yes/no)
- WIC program benefits (yes/no)
- Energy assistance (yes/no)

Race
- White only, Black only, American Indian or Alaskan Native only, Asian only Hawaiian/Pacific Islander only, mixed races

Reason not working
- Not in labor force, Ill or disabled, Taking care of home or family, Going to school, Could not find work, Other

Region
- Northeast, Midwest, South, West

Residential mobility; moved since last year (yes/no)
Sex
- Male, Female

Source of Income
- Unemployment compensation (yes/no)
- Worker's compensation (yes/no)
- Social Security Income (yes/no)
- Supplemental Security Income (yes/no)
- Public Assistence or Welfare (yes/no)
- Veterans' Administration benefits (yes/no)
- Disability income (yes/no)

Total person income in $2,500 increments

Table 3. Estimates of Children's Health Insurance Coverage Using HD or MI Imputation

	% imputed	Percent of children uninsured			Number of children uninsured		
State		*HD*	*MI*	*Difference**	*HD*	*MI*	*Difference*
Alabama	7.4	38.7	37.0	**1.686**	210,261	201,096	9,165
Alaska	14.6	34.8	33.6	**1.224**	53,940	52,042	1,898
Arizona	13.6	45.6	42.8	**2.770**	318,515	299,153	19,362
Arkansas	8.5	42.3	42.8	-0.455	129,677	131,069	(1,393)
California	12.0	43.1	41.8	**1.256**	2,436,505	2,365,507	70,998
Colorado	13.5	33.3	33.2	0.085	227,753	227,168	585
Connecticut	20.9	28.9	28.6	0.313	107,410	106,247	1,163
D_C	15.2	45.9	44.1	**1.755**	20,692	19,901	791
Deleware	14.2	33.4	31.2	2.125	30,987	29,013	1,974
Florida	20.2	40.1	37.5	**2.578**	336,347	314,722	21,625
Georgia	12.4	38.7	37.0	**1.762**	535,759	511,379	24,381
Hawaii	14.3	28.0	28.8	**-0.723**	37,900	38,877	(977)
Idaho	7.9	34.0	32.5	**1.518**	75,844	72,457	3,387
Illinois	14.9	31.6	29.6	**2.003**	551,581	516,616	34,965
Indiana	11.8	29.2	27.7	**1.493**	239,988	227,726	12,262
Iowa	11.9	28.5	27.1	**1.419**	84,908	80,688	4,221
Kansas	9.1	33.7	33.6	0.095	120,187	119,849	337
Kentucky	13.7	36.2	34.8	**1.449**	173,881	166,926	6,955
Louisiana	15.8	41.6	40.3	**1.298**	292,440	283,318	9,122
Maine	12.6	35.3	34.7	0.589	41,554	40,859	695
Maryland	10.0	25.7	25.7	-0.014	194,389	194,498	(109)
Massachusetts	12.9	26.6	26.1	0.563	170,364	166,762	3,601
Michigan	11.5	28.6	28.7	-0.123	393,962	395,657	(1,695)
Minnesota	10.2	28.1	27.3	**0.806**	194,465	188,881	5,584
Mississippi	15.9	46.0	43.4	**2.576**	198,457	187,338	11,120
Missouri	15.1	30.2	29.9	0.338	203,366	201,092	2,275
Montana	5.4	39.0	39.2	**-0.208**	42,548	42,775	(227)
N_Carolina	12.1	42.6	40.7	**1.871**	423,570	404,958	18,613
N_Dakota	9.2	33.4	32.7	0.684	22,154	21,700	454
Nebraska	17.3	32.2	29.6	**2.526**	70,132	64,624	5,509
Nevada	9.7	31.9	30.5	**1.408**	93,539	89,415	4,124
New_Hampshire	13.2	23.7	23.2	0.497	38,356	37,552	803
New_Jersey	15.1	32.7	30.8	**1.906**	317,336	298,819	18,518
New_Mexico	11.3	46.9	45.0	**1.845**	172,579	165,784	6,795
New_York	19.2	33.2	32.1	**1.126**	745,206	719,957	25,249
Ohio	13.1	33.5	32.0	**1.561**	462,836	441,279	21,557
Oklahoma	7.1	43.1	42.5	**0.614**	188,182	185,501	2,681
Oregon	11.4	34.2	33.1	**1.132**	139,668	135,045	4,624
Pennsylvania	15.0	29.6	28.9	**0.757**	297,179	289,581	7,598
Rhode_Island	15.6	32.4	31.7	0.645	30,883	30,268	616
S_Carolina	12.3	34.2	34.5	**-0.291**	179,387	180,911	(1,524)
S_Dakota	9.2	33.0	31.6	1.421	31,210	29,867	1,343
Tennessee	16.3	33.7	34.1	-0.472	234,040	237,321	(3,282)
Texas	10.9	46.8	46.1	**0.689**	1,785,159	1,758,876	26,283
Utah	14.6	29.9	27.2	**2.692**	157,791	143,565	14,226
Vermont	19.0	35.8	34.7	**1.098**	25,077	24,308	769
Virginia	14.0	32.4	30.0	**2.349**	306,294	284,071	22,223
W_Virginia	12.0	37.6	37.1	0.451	47,180	46,614	566
Washington	12.8	34.2	32.9	**1.353**	291,525	280,000	11,525
Wisconsin	12.9	28.0	27.4	0.591	186,375	182,439	3,936
Wyoming	10.8	32.0	31.8	0.207	22,774	22,626	148

*Bold numbers indicate that difference was statistically signifcant at .001 level.

Figure 1. Funding Changes with Multiple Imputation Estimates instead of HD Estimates of % of Children Uninsured

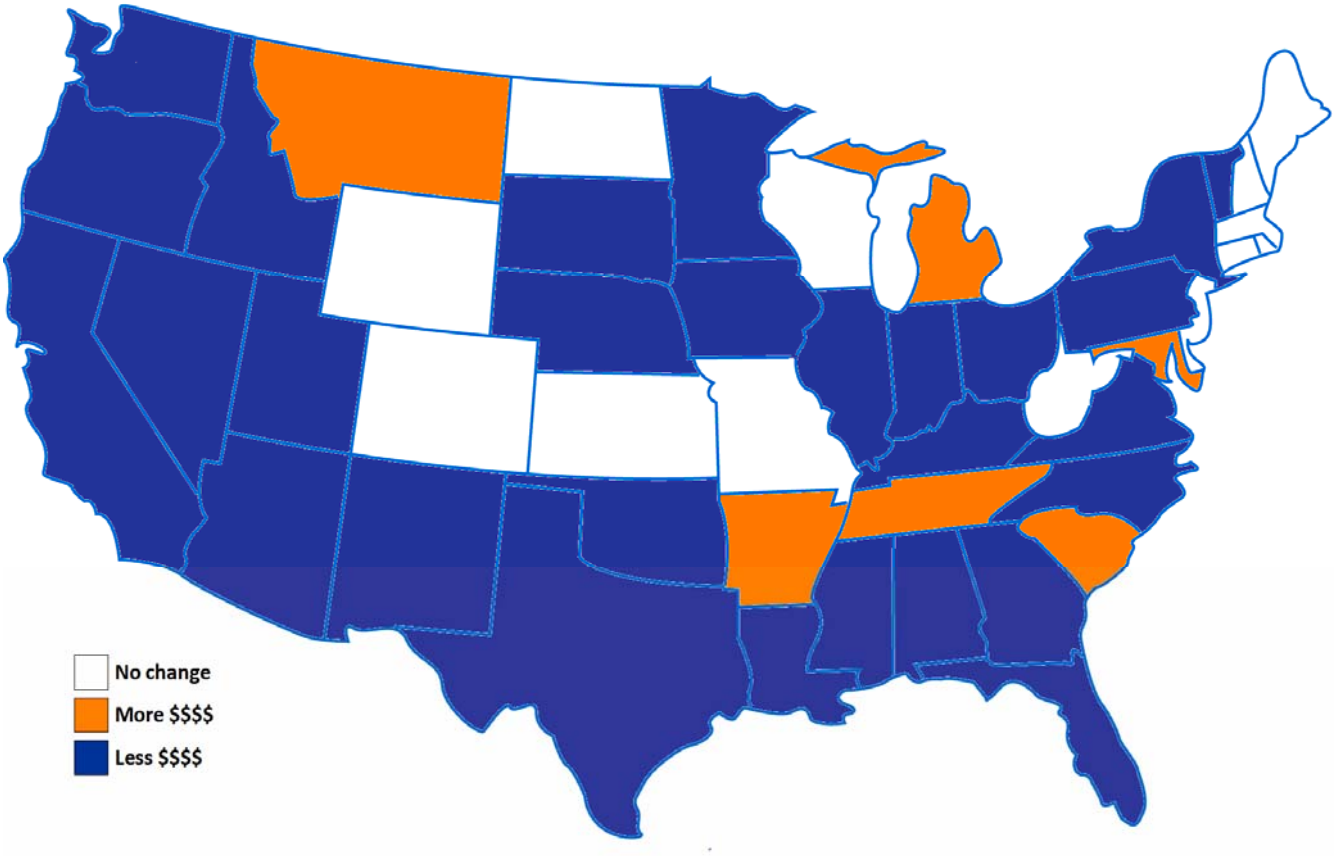

Table 4. Logistic Regression Predicting Whether or not Child has Health Insurance: Comparing HD and MI imputation results

	Hot Deck				Multiple Imputation				Difference (HD - MI)	
	OR	Std. Err.	z	P>z	OR	Std. Err.	z	P>z	OR	Std. Err.
Demographics										
Age	0.930	0.002	-38.730	0.000	0.928	0.002	-36.270	0.000	0.002	0.000
Male	0.992	0.019	-0.420	0.675	0.989	0.020	-0.530	0.594	0.003	-0.002
Black	0.834	0.026	-5.830	0.000	0.800	0.028	-6.410	0.000	0.033	-0.002
Am. Ind.	0.351	0.031	-12.030	0.000	0.300	0.028	-12.720	0.000	0.051	0.002
Asian	1.184	0.062	3.220	0.001	1.194	0.064	0.063	0.001	-0.010	-0.002
Hispanic	0.378	0.009	-39.370	0.000	0.357	0.010	-38.310	0.000	0.021	0.000
Mixed Race	0.853	0.041	-3.300	0.001	0.815	0.042	-3.920	0.000	0.038	-0.002
Region										
Northeast	1.207	0.035	6.450	0.000	1.186	0.038	5.290	0.000	0.022	-0.003
Midwest	1.162	0.032	5.440	0.000	1.173	0.035	5.300	0.000	-0.011	-0.003
South	0.928	0.023	-2.980	0.003	0.916	0.025	-3.200	0.001	0.012	-0.002
Family Characteristics										
Family Inc.	1.052	0.002	30.820	0.000	1.073	0.002	35.770	0.000	-0.021	0.000
Public Housing	0.670	0.007	-36.430	0.000	0.656	0.008	-35.020	0.000	0.014	-0.001
Food Stamps	4.055	0.125	45.280	0.000	4.072	0.138	41.300	0.000	-0.017	-0.013
Family Size	1.115	0.007	16.250	0.000	1.090	0.008	11.760	0.000	0.025	0.000

n = 18,332