Multiple Imputation to Enhance the CMS-Medicaid Linked Data

Jennifer Rammon[1] M.S., Yulei He[2] Ph.D., and Jennifer D. Parker[2] Ph.D.

[1]Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, U.S.A.

[2]Division of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, U.S.A.

Disclaimer: The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Abstract

Data from the National Health and Nutrition Examination Survey (NHANES) have been linked to the Center for Medicare and Medicaid Services' Medicaid Enrollment and Claims Files. As not all survey participants provide sufficient information to be eligible for record linkage, linked data often includes fewer records than the original survey data. This project presents an application of multiple imputation (MI) for handling missing Medicaid status due to linkage refusals in linked NHANES-Medicaid data using linkages of 1999-2004 NHANES survey years. By examining multiple outcomes and subgroups among children, the analyses compares the utility of a multi-purpose dataset from a single MI model to that of individualized models. Outcomes examined here include obesity, untreated dental caries, attention deficit hyperactivity disorder (ADHD), and exposure to second hand smoke.

Keywords: Multiple Imputation, Data Linkage, NHANES, Medicaid, children

Introduction

The Medicaid program is the largest health insurance program in the United States. Together with the Children's Health Insurance Program (CHIP), Medicaid covers over thirty percent of all children, over fifty percent of low-income children, and over forty percent of all births in the United States (Kaiser 2016). In 2014, children represented 43% of overall Medicaid enrollment and 17% of all Medicaid expenditures (Burwell et al. 2015). Given that such a large number of children rely on Medicaid and CHIP coverage for their health care, understanding the health status of these enrollees is important. Future assessments of the Medicaid and CHIP program rely on a clear evaluation of the health status of Medicaid and CHIP children.

The National Health and Nutrition Examination Survey (NHANES) provides national estimates from in-home interviews and physical examinations. NHANES

biomarkers are relied upon to establish population reference ranges, track exposure trends, and prioritize research needs. The NHANES questionnaire incorporates detailed information about study participants' health insurance, including self-reported Medicaid/CHIP enrollment status. Previous research, however, which has compared Medicaid status reported in surveys with administrative records, has shown that Medicaid enrollment is often underreported on health surveys (Davern 2007, Davern et al. 2009). This phenomenon is referred to as the "Medicaid Undercount". One report using NHANES data which have been linked to the Centers for Medicare and Medicaid Services' Medicaid Analytic eXtract files (CMS MAX) indicates that among 1999-2004 NHANES participants under the age of 18, approximately 74% of those enrolled in Medicaid actually reported being enrolled (unweighted percentage, Mirel et. al. 2014).

Due to the Medicaid Undercount, using linked files to determine Medicaid and CHIP status may lead to estimates that are more accurate. Within the linked dataset, the administrative data provide information regarding monthly enrollment status, eligibility group, and use and costs of services during the year, while survey data capture sociodemographic characteristics, health history (addressed and unaddressed by doctors), dietary habits, health-related behaviors, access to health care, laboratory measures, and physical examination components.

A disadvantage of linked data is that not all survey participants can be linked to administrative files. NHANES participants who do not provide sufficient personal identifiers, such as their social security number or their health insurance claim number are ineligible for linkage. One way to analyze incompletely linked data is to limit analyses to the linkage eligible individuals. However, survey respondents with sufficient personal identification for linkage are self-selected. If the linkage eligible subset differs systematically from those who are not eligible, then eliminating the linkage ineligibles without adjustments could lead to biased estimates.

In a previous project we compared three methods for determining Medicaid/CHIP status in health analyses of the NHANES-CMS Medicaid linked data: one that used multiple imputation (MI) (Rubin 1987) to impute the administrative Medicaid status of those who are ineligible for linkage, a second that used the linked data restricted to linkage eligible participants with a basic weight adjustment to account for the non-response among linkage ineligibles (Judson, Parker, and Larson 2013), and a third that used self-reported Medicaid/CHIP status from the survey data. We found that when using the NHANES CMS-MAX linked data, both the MI approach and the weight adjustment approach were appropriate and effective ways to address the biases that result from some survey participants being ineligible for linkage. The survey data alone produced statistically unreliable estimates, which were presumably biased based on the Medicaid undercount.

An advantage of the MI approach is that all survey participants can be included in the analysis. A disadvantage of the MI approach is that it requires researchers to access several restricted-use variables in the Research Data Center (RDC). These variables could vary depending on the analysis, but often include the state, month, and year of the NHANES interview, true variance units (as opposed to the publicly released masked variance units), and the number of days the linkage-eligible study participant was enrolled in Medicaid during the month of the NHANES interview.

It is of interest, therefore, to consider the utility of a general use imputation model. A general use imputation model provides a multi-purpose dataset with "complete" cases of administrative Medicaid enrollment for both linkage-eligible study participants and linkage-ineligible study participants, which in this case would be used to analyze health measures within the Medicaid population or to examine associations between Medicaid enrollment and health status. A multi-purpose user dataset makes the MI analysis method accessible to researchers who may not have experience performing multiple imputation themselves and makes comparisons across analyses easier since the same imputed dataset is used for multiple analyses. Identifying the best general use imputation model is a challenge. While it is well known that including all analysis variables (outcomes and covariates) in the imputation model is advantageous, it is impossible to know in advance all of the analyses that might be performed with the multi-purpose dataset.

The objective of this presentation was to compare subject specific imputation models to general use imputation models for a variety of health outcomes among children in order to assess the utility of a multi-purpose user dataset with "complete" cases of administrative Medicaid enrollment. Two general use models were considered: one that did not include any potential health outcome variables (apart from those also considered predictors of Medicaid enrollment) and a second that included 10 variables commonly analyzed as health outcomes. Both models included demographic variables, survey design variables, and predictors of Medicaid enrollment. The motivation behind comparing two general use models was that the first model would provide a valid assessment of whether the multi-purpose dataset would work for analyzing health outcomes that were not considered when the imputation model was being built, while the second model would provide an example of a more informed imputation model. Comparisons were drawn across the three imputation methods (subject-specific, general use without health outcome variables, general use with 10 health outcome variables) through the analyses of four different health outcomes: obesity, untreated dental caries, attention deficit hyperactivity disorder (ADHD), and serum cotinine.

Material and Methods

National Health and Nutrition Examination Survey (NHANES) data:

NHANES is a nationally representative survey of the resident, civilian, noninstitutionalized United States population. It is designed to monitor the country's health and nutritional status and includes an interview in the home followed by a standardized physical examination at a specially designed mobile examination center (MEC). Survey participants are selected using a complex, multistage probability sampling design, details of which have been described elsewhere (Curtin et al. 2012). Sample weights account for oversampling, survey non-response, and post-stratification. During NHANES 1999-2004, oversampled groups included: Mexican-Americans, black persons, low-income persons (at or below 130% of the federal poverty level), and adolescents aged 12-19 years. The oversampling of low income individuals and adolescents increased the sample size of Medicaid/CHIP beneficiaries over what it would have otherwise been had these

populations not been oversampled. A proxy provided information for survey participants who were less than 16 years of age and for individuals who could not answer the questions themselves.

The NHANES survey question from 1999-2004 read, "Is the study participant covered by Medicaid/CHIP?" It did not allow for a distinction between the two or for the exclusion of CHIP beneficiaries from analyses. In efforts to be consistent with the survey question, both Medicaid and CHIP were treated as one category in our analyses.

Centers for Medicare and Medicaid Services' Medicaid Analytic eXtract (CMS MAX) files

Since 1999, Medicaid data have been collected by states and provided to CMS through the Medicaid Statistical Information System (MSIS). These data include enrollee eligibility information, service utilization, and Medicaid claims paid in each quarter of the federal fiscal year. The MAX files are research extracts of MSIS which provide person-level information on demographics, monthly enrollment status, eligibility group, and use and costs of services during the year.

In addition to Medicaid records, the MAX files also contain records from the CHIP. CHIP provides health coverage to low-income, uninsured children and pregnant women in families with incomes too high to qualify for state Medicaid programs. It is administered by states according to federal requirements and is funded jointly by the state and federal governments. States may choose whether to provide Medicaid expansion CHIP programs (M-CHIP), which provide the standard Medicaid benefit package to these children, or separate CHIP programs (S-CHIP), which provide coverage that is actuarially equivalent to other health insurance programs, such as those offered to federal and state employees. Since each state handles S-CHIP differently, S-CHIP is inconsistently reported to MSIS. It is not clear for which states S-CHIP data are included in the CMS MAX files and for which states they are not. The CMS MAX files include all children enrolled in Medicaid, all children enrolled in M-CHIP, and some children enrolled in S-CHIP.

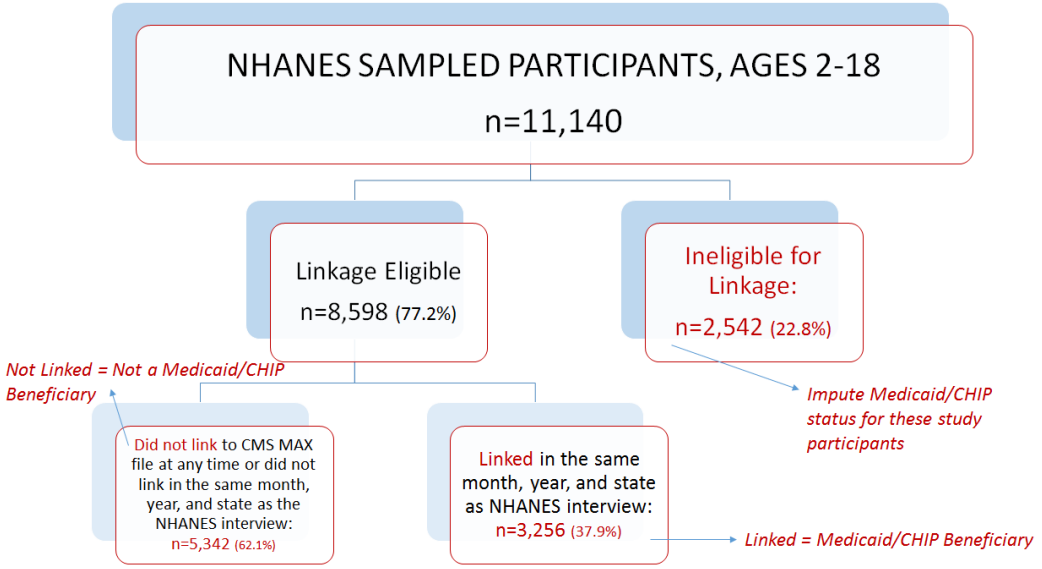
Data Linkage

Data linkage between NHANES and the CMS MAX files is performed regularly by the National Center for Health Statistics' Data Linkage program. For NHANES 1999-2004, survey participants are linkage-eligible if they supply sufficient personally identifiable information (such as social security number and health insurance claim number) and if their SSN is verified by the Social Security Administration's Enumeration Verification System (Golden, et al. 2015). Survey participants are ineligible for linkage if personally identifiable information is not provided. Linkage eligible survey participants matched with the CMS MAX files are considered "linked". The linkage between NHANES data and the CMS MAX files is complete for continuous NHANES 1999-2004. Linked enrollment and claims data for NHANES 2005 through 2012 are expected to be available by the end of 2017. The currently linked data correspond to all Medicaid/CHIP claims files between 1999 and 2007.

Analytic Sample

Using NHANES 1999-2004, this study included children ages 2-18 years who participated in the MEC examination. Figure 1 indicates how many NHANES 1999-2004 participants aged 2-18 years were linkage eligible, how many of the linkage eligible were linked versus not linked, and how many were ineligible for linkage. For this study, children were identified as linked if their linked files included enrollment in Medicaid, S-CHIP, or M-CHIP within the same state, month, and year as the survey. For all analyses using the linked data, children who were linkage eligible and were linked to the administrative records were classified as Medicaid/CHIP beneficiaries (n=3,256), those who were linkage eligible and not linked were classified as non-Medicaid/CHIP beneficiaries (n=5,342), and those who were ineligible for linkage had unknown Medicaid/CHIP status (n=2,542). MI was used to impute enrollment status for children who were ineligible for linkage/had unknown Medicaid/CHIP status. MI was simultaneously used to impute missing information for all other covariates used in the imputation model.

Figure 1:



The sample sizes varied depending on which outcome variable was being analyzed. This was because of differences in data collection across the different components of NHANES. Without placing restrictions on the sample used to analyze each outcome, data included in the imputation model could be systematically missing for certain age groups or survey cycles; for example, the early childhood questionnaire, which provided many covariates for the ADHD imputation, was only administered to children under the age of 15 years. Systematic missingness is not appropriate for the traditional MI model. Figure 2 indicates what restrictions were placed on each of the four outcomes and the final sample sizes corresponding to each outcome specific analytic sample.

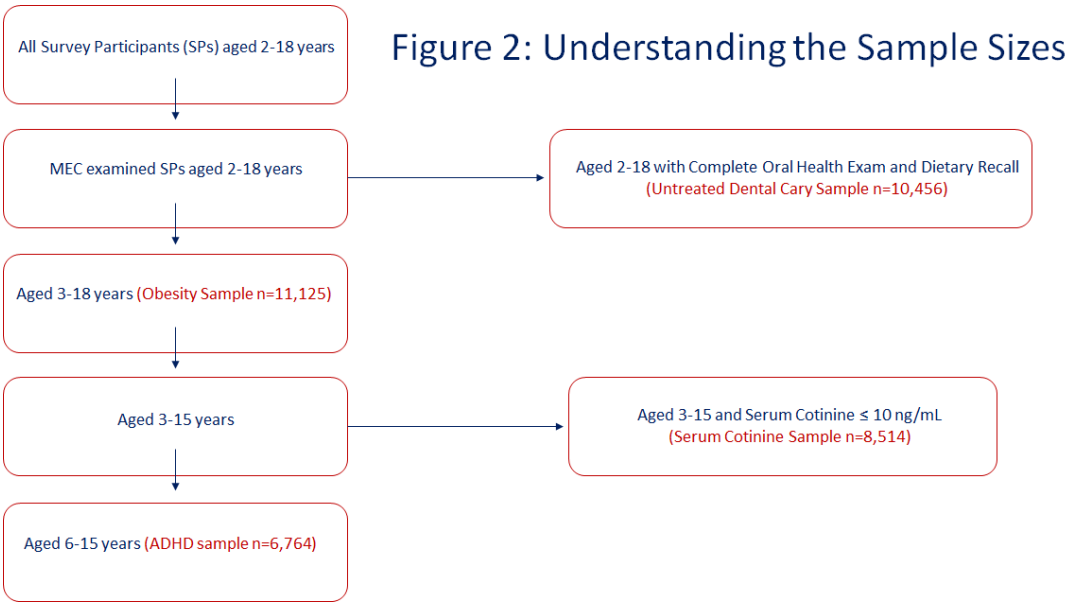


Figure 2: Understanding the Sample Sizes

Multiple Imputation

MI was conducted using SAS version 9.3 PROC MI (FCS option) with 100 imputations. Data were assumed to be missing at random (MAR). Six imputation models were developed: four subject specific imputation models and two general use models. The first general use model included demographic variables, survey design variables, and survey variables related to Medicaid/CHIP enrollment. The second general use model included all of the aforementioned variables, as well as 10 commonly analyzed health outcome variables. The subject specific models included all of the variables used in the first general use model, as well as the outcome variable of interest and predictors related to the outcome variable of interest: obesity, untreated dental caries, ADHD, and serum cotinine. Tables 1 and 2 list the variables included in each of the imputation models.

For all imputations, survey design variables represented the primary sampling units (counties), strata, and sample weights. For technical efficiency, a continuous variable which represents the percentage of Medicaid/CHIP beneficiaries within each PSU based on the linked NHANES CMS MAX data was created to replace the original PSU variable, which had 87 categories. Using PSU level characteristics, rather than PSU indicators, is a technique that was previously implemented when imputing income for the National Health Interview Survey (Schenker et al. 2006). In addition to the survey and design variables, a final explanatory variable was created by combining self-reported family income with Kaiser's 2004 reports of Medicaid and S-CHIP state income thresholds (http://kff.org/state-category/medicaid-chip/). This variable classifies children as Medicaid eligible, S-CHIP eligible, or neither.

Table 1: Variables used in the Two General Use Models

National Health and Nutrition Examination Survey linked to Centers for Medicare and Medicaid Services' Medicaid data: 1999-2004

General use model without health measures and all other imputations	General use model with health measures
	All variables used in the general use model without health measures, plus…
Gender	
Age	
Race/Ethnicity	Untreated dental caries (binary)
Poverty Income Ratio (Socioeconomic Status)	ADHD (binary)
Nativity	Asthma (binary)
Citizenship status	BMI category (underweight, normal weight, overweight, obese)
Age of household reference person	
Education of household reference person	
Nativity of household reference person	Serum cotinine
Self-reported general health status	Blood lead
Average # of health care visits each year	Hemoglobin
Does SP have a routine place for health care	Total cholesterol
Home ownership (binary)	C-Reactive Protein
What type of home SP lives in	Vitamin B12
Census Region	
Medicaid enrollment based on self-report	
Medicaid eligibility status based on self-reported income	
Average administrative Medicaid enrollment across the primary sampling unit (PSU)	
Strata	
Exam weights	
Interview weights	
Dietary Interview weights	

Linear regression was used to impute continuous variables, logistic regression for binary and ordinal variables, and the discriminant function for all other categorical variables. Citizenship status was an exception; though it was only two categories (citizen, not a citizen) the discriminant function was used. Imputation for all missing variables was performed jointly to fully incorporate the relationship among these variables as well as with aforementioned predictors (Collins, Schafer, and Kam 2001).

Table 2: Variables used in the ADHD, Obesity, and Untreated Dental Caries Subject Specific Model
National Health and Nutrition Examination Survey linked to Centers for Medicare and Medicaid Services' Medicaid data: 1999-2004

Obesity	Untreated Dental Caries	Attention Deficit Hyperactivity Disorder (ADHD)	Serum Cotinine
All variables used in the general use model without health measures, plus…	*All variables used in the general use model without health measures, plus…*	*All variables used in the general use model without health measures, plus…*	*All variables used in the general use model without health measures, plus…*
Body weight	WIC status (Special Supplemental Nutrition Program for Women, Infants, and Children)	Seen a mental health professional in the last 12 months	Mother smoked while pregnant
Standing height	Dental sealants	Mother smoked while pregnant	# of cigarettes/day smoked in the home
Waist circumference	Time since last dental visit	Smoker in the home now	
Triceps skin fold	Reason for last dental visit	Currently taking medications that are typically prescribed for ADHD	
Sub scapular skin fold	Categorical BMI	Birthweight	
Hemoglobin	Total number of carbs eaten yesterday (gm)	Blood lead levels	
Total cholesterol	Total plain water drank yesterday (gm)	Maternal age at birth	
C-Reactive Protein (CRP)	Candy eaten yesterday (# of times)	Categorical BMI	
# times/week eat restaurant food	Soda drank yesterday (# of times)	# of hours/day spend time watching TV, playing video games, or on the computer	
# hours/day spend time watching TV, playing video games, or on the computer			

Within each analytic sample, sample sizes also varied across imputation methods. All imputation models imputed missing values for all variables that were included in the imputation model. However, variables that were included in the analysis model, but not the imputation model were subject to item non-response and survey participants with item non-response were excluded from analyses. The general use model without health outcome variables included/imputed the administrative Medicaid enrollment variable and demographic covariates, but did not include/impute the outcome variable or covariates that were specific to the outcome variable. Thus study participants with item non-response associated with either the outcome variable or outcome specific covariates were excluded from analyses based on the general use imputation without health outcomes. The general use model with health outcome variables included/imputed the administrative Medicaid enrollment variable, the outcome variable, and demographic covariates, but covariates that were specific to the outcome variable were not included in the imputation model. Thus study participants with item non-response associated with outcome specific covariates were excluded from analyses based on the general use imputation with health outcomes. The subject specific imputation model included/imputed the administrative Medicaid enrollment variable, the outcome variable, and all covariates used in the final analysis. Thus no variables had item non-response for analyses using datasets from the subject specific imputations. The differences in sample size are displayed in the results section (Table 3).

Analysis

Logistic regression models were fit to examine the association between obesity and Medicaid/CHIP enrollment, ADHD and Medicaid/CHIP enrollment, and untreated dental caries and Medicaid/CHIP enrollment. A log linear model was fit to examine the association between serum cotinine levels and Medicaid/CHIP enrollment. Medicaid/CHIP enrollment was defined as a binary variable: enrolled or not enrolled.

All models controlled for the following sociodemographic characteristics: gender (male/female), race/Hispanic origin (Mexican American, non-Hispanic white, non-Hispanic black, all other races and ethnicities including multi-racial), age at the time of the

mobile examination (varied across models: sometimes categorized as 1-5, 6-11, 12-18 and sometimes included as a continuous variable), and poverty-income ratio (ordinal: ≤1, 1.01-2, 2.01-3, 3.01-4, >4). With the exception of gender, all of these variables have been previously shown to be associated with Medicaid/CHIP enrollment (Dubay and Kenney, 1996; Kincheloe, Frates, and Brown 2007; Simon et. al. 2013). All models with the exception of the ADHD model controlled for education of the household reference person (≤ High school graduate/GED, some college/associates degree/college graduate or higher). This variable was considered in the ADHD model, but it was not statistically significant and its inclusion/exclusion did not impact the estimates associated with other covariates, so it was ultimately excluded. In addition, the ADHD model controlled for self-reported health status (excellent, very good, good, fair, poor), the untreated dental caries model controlled for time since the last dental visit (never, <6 months, 6-12 months and >12 months), and the serum cotinine model controlled for whether or not someone in the home smokes (yes/no).

The household reference person is the first household member, 18 years of age or older who is listed on the screener questionnaire household member roster who owns or rents the residence where members of the household reside. The education variable for the household reference person is the highest grade or level of education completed by him/her with response categories corresponding to less than 9^{th} grade education, $9\text{-}11^{th}$ grade education (includes 12^{th} grade and no diploma), High school graduate/GED, some college or associates (AA) degree, and college graduate or higher. The poverty-income ratio variable is an index for the ratio of self-reported family income and a federal poverty guideline specific to family size, year, and state provided by the Department of Health and Human Services' (HHS) poverty guidelines.

All analyses were performed with SAS-callable SUDAAN, version 9.3 PROC REGRESS, and accounted for the complex survey design. Variance estimates were calculated using the Taylor linearization with replacement method and Student's t-tests were conducted to test the null hypothesis that β coefficients were equal to zero by using a significance level of $p<0.05$.

Results

Table 3 presents the results and Figure 3 shows the beta coefficient and 95% confidence interval corresponding to Medicaid enrollment within each regression. As can be seen by comparing the unadjusted regressions to the full models, adjusting for covariates substantially affected the Medicaid/CHIP coefficients across all four outcome variables. Differences across imputation methods were small. Across all three imputation methods there was a statistically significant association between ADHD and Medicaid and between serum cotinine and Medicaid. Similarly, across all three imputation methods there were no statistically significant associations between obesity and Medicaid or untreated dental caries and Medicaid.

The beta coefficient corresponding to Medicaid enrollment within the ADHD analysis was 0.73 (SE=0.19) using the subject specific imputation, 0.61 (SE=0.18) using

the general use imputation without health outcomes, and 0.65 (SE=0.19) using the general use imputation with health outcomes. The relative standard errors (RSE=[standard error/estimate]*100) associated with these estimates were 26.0%, 29.5%, and 29.5%, respectively. The corresponding odds ratios were 2.08 [95% CI: (1.43, 3.05)], 1.83 [95% CI: (1.26, 2.66)], and 1.91 [95% CI: (1.31, 2.78)], respectively. All three imputation methods indicate that after controlling for gender, age, race, poverty-income ratio, and self-reported health status, the odds of children enrolled in Medicaid having ADHD was approximately two times that of children who are not enrolled.

Table 3 : Coefficients Associated with Binary Medicaid Enrollment Status

National Health and Nutrition Examination Survey linked to Centers for Medicare and Medicaid Services' Medicaid data: 1999-2004

Outcome	Subject Specific MI			General Use MI[6] No health outcomes			General Use MI[7] 10 health outcomes		
	β (SE)	*p*	*n*	*β (SE)*	*p*	*n*	*β (SE)*	*p*	*n*
Adjusted[1] ADHD Model[2]	0.73 (0.19)	0.0003	6,764	0.61 (0.18)	0.002	6,750	0.65 (0.19)	0.0012	6,764
Unadjusted	*0.58 (0.16)*	*0.0008*	*6,764*	*0.51 (0.16)*	*0.003*	*6,750*	*0.53 (0.16)*	*0.002*	*6,764*
Adjusted[1] Cotinine Model[3]	0.34 (0.08)	0.0001	8,514	0.30 (0.08)	0.0005	6,740	0.32 (0.08)	0.0001	8,392
Unadjusted	*1.18 (0.10)*	*<0.0001*	*8,514*	*1.18 (0.09)*	*<0.0001*	*6,807*	*1.15 (0.08)*	*<0.0001*	*8,496*
Adjusted[1] Obesity Model[4]	0.13 (0.10)	0.2	11,125	0.16 (0.10)	0.11	10,900	0.16 (0.10)	0.11	11,125
Unadjusted	*0.25 (0.08)*	*0.003*	*11,125*	*0.27 (0.08)*	*0.002*	*10,900*	*0.27 (0.08)*	*0.001*	*11,125*
Adjusted[1] Untreated Dental Caries Model[5]	0.10 (0.13)	0.44	10, 456	0.13 (0.13)	0.33	9,818	0.11 (0.13)	0.38	9,818
Unadjusted	*0.64 (0.11)*	*<0.0001*	*10,456*	*0.64 (0.11)*	*<0.0001*	*10,456*	*0.64 (0.11)*	*<0.0001*	*10,456*

[1] All adjusted models control for gender, race/Hispanic origin, age, and poverty-income ratio

[2] Children aged 6-15 years at the time of the MEC examination; additional covariates included self-reported health status

[3] Children aged 3-15 years at the time of the MEC examination with serum cotinine levels ≤10 ng/mL; additional covariates included education of household reference person and whether or not there is a smoker in the home

[4] Children aged 3-18 years at the time of the MEC examination; additional covariates included education of household reference person

[5] Children aged 2-18 years at the time of the MEC examination with a complete oral health exam and complete 24-hr dietary recall; additional covariates included education of household reference person and time since the last dental visit

[6] Children with item-nonresponse for the outcome variable or for outcome specific covariates not included in the analyses

[7] Children with item non-response for outcome specific covariates not included in the analyses

The beta coefficient corresponding to Medicaid enrollment within the serum cotinine analysis was 0.34 (SE=0.08) using the subject specific imputation, 0.30 (SE=0.08) using the general use imputation without health outcomes, and 0.32 (SE=0.08) using the general use imputation with health outcomes. The RSEs associated with these estimates were 23.5%, 26.7%, and 25.0%, respectively. All three imputation methods indicate that after controlling for gender, race, age, poverty-income ratio, the education of the household reference person, and whether or not there is a smoker in the home, the average serum cotinine levels of Medicaid/CHIP beneficiaries was about 35-40% higher than that of non-Medicaid/CHIP beneficiaries (after exponentiation).

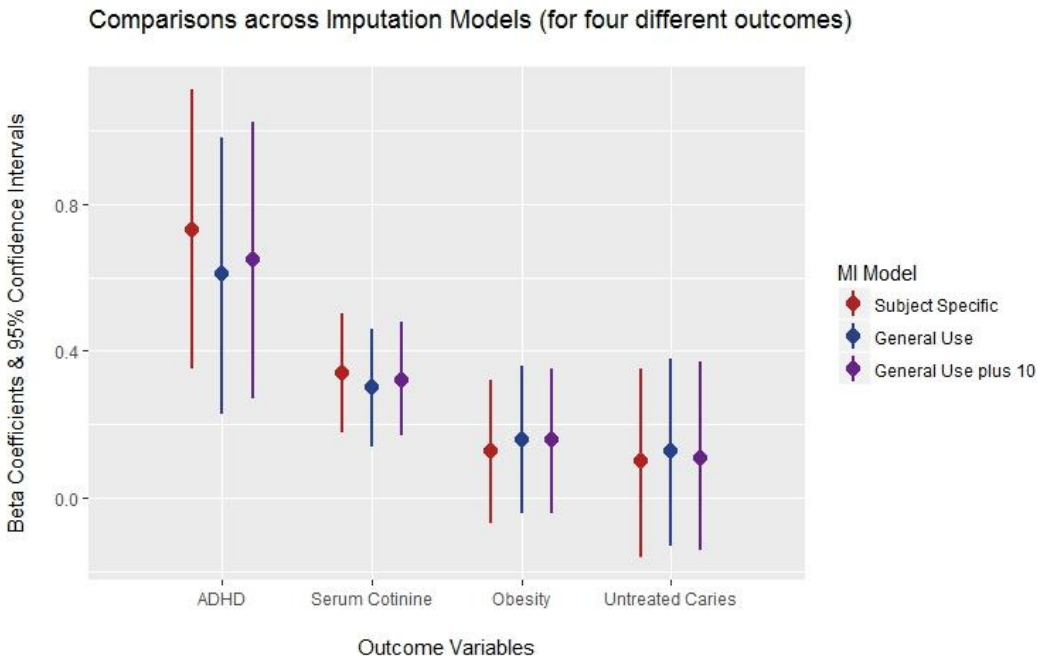
The beta coefficient corresponding to Medicaid enrollment within the obesity analyses was 0.13 (SE=0.10) using the subject specific imputation, 0.16 (SE=0.10) using the general use imputation without health outcomes, and 0.16 (SE=0.10) using the general use imputation with health outcomes. The RSEs associated with these estimates were 76.9 %, 62.5%, and 62.5%, respectively. The corresponding odds ratios were 1.13 [95% CI: (0.93, 1.38)], 1.17 [95% CI: (0.96, 1.43)], 1.17 [95% CI: (0.96, 1.43)], respectively.

The beta coefficient corresponding to Medicaid enrollment within the untreated dental caries model was 0.10 (SE=0.13) using the subject specific imputation model, 0.13 (SE=0.13) using the general use imputation without health outcomes, and 0.11 (SE=0.13) using the general use imputation with health outcomes. The RSEs associated with these estimates were 130%, 100%, and 118%, respectively. The corresponding odds ratios were 1.10 [95% CI: (0.86, 1.42)], 1.13 [95% CI: (0.88, 1.46)], and 1.12 [95% CI: (0.87, 1.44)], respectively.

These results demonstrate that for both the ADHD and serum cotinine analyses the subject specific imputation produced the most precise estimates and that as compared to the subject specific imputation, the two general use imputations both produced results that were unbiased. For the ADHD analysis, the Medicaid enrollment beta coefficient corresponding to the general use imputation without health outcomes was within 17% of the beta coefficient produced using the subject specific imputation, while the Medicaid enrollment beta coefficient corresponding to the general use imputation with health outcomes was within 11% of the beta coefficient produced using the subject specific imputation. For the serum cotinine analysis, the Medicaid enrollment beta coefficient corresponding to the general use imputation without health outcomes was within 12% of the beta coefficient produced using the subject specific imputation, while the Medicaid enrollment beta coefficient corresponding to the general use imputation with health outcomes was within 6% of the beta coefficient produced using the subject specific imputation.

The results associated with obesity and untreated dental caries indicate slightly higher precision among the general use imputation models. Differences across beta coefficients ranged from 10% to 30%. For the obesity analyses, the Medicaid enrollment beta coefficients corresponding to the general use imputations were within 23% of that produced using the subject specific imputation. For the untreated dental caries analyses, the Medicaid enrollment coefficient corresponding to the general use imputation without health outcomes was within 30% of that produced when using the subject specific imputation and the Medicaid enrollment coefficient corresponding to the general use imputation with health outcomes was within 10% of that produced using the subject specific imputation.

In short, the three imputation methods were comparable. Across all three methods, differences between RSEs were small and estimated beta coefficients were similar.



General Use model: demographic variables, survey design variables, and survey variables related to Medicaid/Chip enrollment.
General Use plus 10 model: General Use model +10 commonly analyzed health variables.
Subject Specific model: General Use model plus the outcome of interest and predictors related to the outcome of interest.

Discussion

The results indicate that a multi-purpose user dataset with "complete" cases of administrative Medicaid enrollment provides reliable estimates that closely match those produced using a standard subject specific imputation model when analyzing health outcomes within the Medicaid population or examining associations between Medicaid enrollment and health status. Across all four outcomes presented, the general use imputation methods produced estimates that were similar to the subject specific imputation method in terms of both precision and magnitude of effect. This is encouraging since there are two practical advantages of a multi-purpose user dataset. First, it makes an administrative Medicaid status variable more available to researchers and second, it allows for consistent comparisons across analyses using the MI method to account for the potential bias due to linkage ineligibles.

More exploration is needed to determine an optimal general use imputation model. Two candidate models are provided here, but specific criteria must be established to determine exactly what variables should be included. Furthermore, it may be that different datasets should be produced for different types of analysis. Based on these examples, we make two observations. First, the effectiveness of the general use model with health outcomes demonstrates that including commonly analyzed health outcome variables informs the imputation and produces estimates that most closely match the estimates expected from subject specific imputation models. Second, the effectiveness of the general use model without health outcome variables demonstrates that regardless of what type of analysis is performed, and specifically whether or not the health outcome was included in

the original imputation, a multi-purpose user dataset can produce reliable estimates associated with Medicaid enrollment, similar to those expected from subject specific imputation models.

In some cases, however, a subject specific imputation model may still be preferred. Using a multi-purpose dataset may exclude children who have item non-response for any variables not originally included in the general use imputation model. In these cases, a subject specific imputation would maximize the number of study participants included in the final analyses.

Conclusion

This study illustrates the ability of a general use imputation model to produce a multi-purpose user dataset with "complete" cases of administrative Medicaid enrollment for analysis of health measures within the Medicaid population or to examine associations between Medicaid enrollment and health status. The best general use imputation model for this task is unknown.

References

Collins, Linda M. Schafer, Joseph L. Kam Chi-Ming. Dec. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4): 330-351.

Curtin LR, Mohadjer L, Dohrmann S, et al. 2012. The National Health and Nutrition Examination Survey: Sample design, 1999–2006. National Center for Health Statistics. Vital Health Stat 2(155).

Davern, Michael. 2007. Phase 1 Research Results: Overview of the National Medicare and Medicaid Files. Research Project to Understand the Medicaid Undercount: The University of Minnesota's State Health Access Data Assistance Center, the Centers for Medicare and Medicaid Services, the Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation, and the U.S. Census Bureau. [accessed on September 1, 2017]. Available at: https://www.census.gov/did/www/snacc/docs/SNACC_Phase_I_Full_Report.pdf

Davern, Michael. Klerman, Jacob Alex. Baugh, David K. Call, Kathleen Thiede. Greenberg, George D. 2009. An Examination of the Medicaid Undercount in the Current Population Survey: Preliminary Results from Record Linking. Health Services Research 44(23):965-87. doi: 10.1111/j.1475-6773.2008.00941.x

Department of Health and Human Services. 2015. "2015 Actuarial Report on the Financial Outlook for Medicaid." [accessed on September 1, 2017]. Available at: https://www.medicaid.gov/medicaid/financing-and-reimbursement/downloads/medicaid-actuarial-report-2015.pdf

Dubay, Lisa C and Kenney, Genevieve M. 1996. Revisiting the Issues: The Effects of Medicaid

Expansions on Insurance Coverage of Children. The Future of Children SPECIAL EDUCATION FOR STUDENTS WITH DISABILITIES 6(1):152-161. doi: 10.2307/1602499.

Golden C, Driscoll AK, Simon AE, Judson DH, Miller EA, Parker JD. 2015. Linkage of NCHS population health surveys to administrative records from Social Security Administration and Centers for Medicare & Medicaid Services. National Center for Health Statistics. Vital Health Stat 1(58).

Judson DH, Parker JD, Larsen MD. May 2013. "Adjusting sample weights for linkage-eligibility using SUDAAN." National Center for Health Statistics, Hyattsville Maryland. [accessed on September 1, 2017]. Available at: http://www.cdc.gov/nchs/data/datalinkage/adjusting_sample_weights_for_linkage_eligibility_using_sudaan.pdf

The Kaiser Commission on Medicaid and the Uninsured. 2013. "Medicaid A Primer: Key Information on the Nation's Health Coverage Program for Low-Income People." Report No. 7332-05. [accessed on September 1, 2017]. Available at: https://kaiserfamilyfoundation.files.wordpress.com/2010/06/7334-05.pdf

Kincheloe J, Frates J, Brown ER. April 2007. Determinants of children's participation in California's Medicaid and SCHIP programs. Health Services Research. 42(2):847-66.

Mirel, Lisa B. Simon, Alan E. Golden, Cordell. Duran, Catherine R. Schoendorf, Kenneth C. 2014 Jan 6. Concordance Between Survey Report of Medicaid Enrollment and Linked Medicaid Administrative Records in Two National Studies. National Health Statistics Reports. (72):1-9.

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

Schenker, Nathaniel. Raghunathan, Trivellore E. Chiu Pei-Lu. Makuc, Diane M. Zhang, Guangyu. Cohen, Alan J. Sept 2006. Multiple Imputation of Missing Income Data in the National Health Interview Survey. Journal of the American Statistical Association. 101(475): 924-933. doi 10.1198/016214505000001375.

Simon, Alan E. Driscoll, Anne. Gorina, Yelena. Parker, Jennifer D. Schoendorf, Kenneth C. 2013. A Longitudinal View of Child Enrollment in Medicaid. Pediatrics 132 (656). doi 10.1542/peds.2013-1544.