IAP | Medicaid Innovation Accelerator Program

WEBINAR SUPPLEMENT: SOLVING MISSING DATA PROBLEMS BY MULTIPLE IMPUTATION

The Medicaid Innovation Accelerator Program (IAP) provides targeted technical assistance and tools to Medicaid agencies in data analytics, one of the four Medicaid IAP functional areas for Medicaid delivery system reform. On June 7, 2018, Medicaid IAP hosted an introductory webinar to address problems with missing data. Medicaid IAP held a subsequent webinar on this topic on October 23, 2018, titled "Solving Missing Data Problems." This webinar shared solutions and information on statistical methods for imputation, a method by which to substitute missing data with estimated values. Slides, audio recordings, and transcripts are available for both webinars on the Medicaid IAP Data Analytics web page under National Dissemination.[1]

This webinar supplement provides an overview of missing data problems and standard techniques to address them in Medicaid policy and program analysis. Specifically, we highlight **multiple imputation** as a standard approach for handling missing data and provide guidance on how programmers or analysts could apply this method in a study. Multiple imputation is discussed alongside **complete-case analysis** and **single imputation**; these are alternative methods that rely on strong assumptions about missing data patterns.

The first section of this supplement provides an overview of missing data, including how to diagnose and assess problems related to missing data. In the second section, we present standard techniques for complete-case analysis, single imputation, and multiple imputation, along with their assumptions, benefits, and risks for robust data analytics that incorporate missing data. Finally, we provide an example illustration that shows how to implement multiple imputation. Statistical software resources are listed in the Appendix, and example code is available on the Medicaid IAP Data Analytics web page under National Dissemination.[2]

I. Diagnosing the Missing Data Problem

Missing data refers to data elements within an observation or observations within a sample that are incomplete. The term does not refer to observations with no data, nor does it include values that would otherwise come from unobserved or unobservable variables (e.g., an enrollee's motivation to seek necessary preventive care is not measured in claims data). If handled improperly, missing data can pose problems for data analysis. For example, a disparities analysis of preventive care use by Medicaid enrollees would be potentially biased if a large portion of the enrollment data were missing race and ethnicity values. Before conducting an analysis that includes missing data, the analyst needs to investigate whether the missing data pattern provides information about relationships in the observed data.

Methods for conducting missing data analysis require assumptions to address sources of bias and statistical uncertainty when making inferences. To diagnose the missing data problem, the analyst should consider whether missing values for variables in the dataset follow a pattern (i.e., the corresponding dummy variables for missingness) and how that pattern may alter the study's conclusions on the basis of

[1] Centers for Medicare & Medicaid Services. Data Analytics. https://www.medicaid.gov/state-resource-center/innovation-accelerator-program/iap-functional-areas/data-analytics/index.html
[2] Ibid.

an analysis of the observed data. Importantly, assumptions about the pattern of missing data will guide the selection of statistical adjustments to address the missing values. There are three types of missing data patterns; for each, Table 1 provides definitions, examples, and proposed approaches.[3]

Table 1. Definitions, Examples, Diagnosis, and Analysis for Three Types of Missing Data Problems

Pattern	Definition and Example	Diagnostic and Analytic Methods
Missing completely at random (MCAR)	The pattern of missing values in the dataset is unrelated to the study variables. Complete cases can be characterized as a random sample of the full dataset. **Example:** From a survey sample, a respondent's questionnaire was lost in the mail.	**Diagnosis:** Little's MCAR test (continuous data), chi-squared test (categorical data), and independence test between missingness dummy variable and observed variables are recommended.[a] **Analysis:** Complete-case analysis, listwise deletion; pairwise deletion
Missing at random (MAR)	The pattern of missing data **does not depend on the missing values** but could depend on observed values of other variables. Other variables can be used to model and predict missing values. **Example:** Some racial and ethnic groups are less likely to report income on surveys.	**Diagnosis:** Correlations of missingness dummy variable with observed variables is recommended. MAR is generally untestable, so sensitivity analyses are also recommended to assess how study results and conclusions may change under different missing data models. **Analysis:** Regression adjustments, weighting, multiple imputation
Missing not at random (MNAR)	The pattern of missing data **depends on the missing values**. The missing data mechanism cannot be ignored and must be modeled. **Example:** Patients with the highest severity of illness are more likely to drop out of a clinical study.	**Diagnosis:** MNAR is generally untestable, so sensitivity analyses are recommended to assess how study results and conclusions may change under different missing data models. **Analysis:** Multiple imputation; sensitivity analysis

[a] See the section on Statistical Testing for details on these methods.

Exploratory Visualization

The analyst should investigate the pattern of missing data using visualization methods. The appropriate solution for missingness (discussed in the next section) often can be determined by identifying the right missing data pattern. A test of differences in rates of missing data between groups (e.g., men may be less likely than women to provide information on a mental health questionnaire) also can be used as a diagnostic tool. Both visualization and more formal procedures help prepare the data for imputation. Some techniques are illustrated in the next section.

Statistical Testing

Missing Completely at Random (MCAR) Versus Missing at Random (MAR)

Per Table 1, the analyst might approach a missing data problem by first assessing the missing completely at random (MCAR) assumption. For MCAR, the missing data pattern has no relationship to any observed

[3] Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581-592.

or missing values for all study variables. Under this assumption, the analyst can approach the complete cases as if they were a random sample from the study sample.

For each missing data pattern, we consider the collection of cases with that pattern and the observed means. For example, if there are three variables (Variables 1, 2, and 3), then there are seven possible patterns of missing data (Table 2, Patterns 2 through 8).

Table 2. Possible Missing Data Patterns—M=Missing and "+"=Observed

Pattern	Variable 1	Variable 2	Variable 3	No. of Missing Values
1	+	+	+	Zero
2	M	+	+	One
3	+	M	+	
4	+	+	M	
5	M	M	+	Two
6	M	+	M	
7	+	M	M	
8	M	M	M	Three

Across the missing data patterns, the MCAR assumption would yield observed means that are similar; in Table 2, the observed means for Variable 1 would be similar across the Patterns 1 through 8. Tests of independence generally are based on the chi-squared test statistic. If the sample (i.e., observed) means vary by missing data patterns, then independence is rejected, which implies that the pattern is not MCAR. Little's test of MCAR can be applied to continuous variables; for categorical data, Fuchs' test statistic can be calculated.[4,5] Similar tests can be applied that assess the association between the missingness pattern and (observed) study variables, for example, by analyzing the rates of a variable's missingness against the observed values. The individual-level missing data can be represented by a dummy variable, which equals one (=1) if the value is missing and zero (=0) otherwise. The correlations between the missingness dummy variable and observed covariates can then provide a basis to develop imputation models under the missing at random (MAR) assumption.

MAR Versus Missing not at Random (MNAR)

An inherent assumption of MAR analysis is that missingness does not depend on information that is unobserved. In contrast, an MNAR pattern depends on the missing values. There is no empirically based approach for testing MAR or MNAR. Ideally, the analyst could collect more data (e.g., conduct follow-up telephone calls for non-respondents to a health care survey) and use them to assess the missing data pattern; however, additional data collection is often infeasible.[6]

---

[4] Li C. Little's test of missing completely at random. *The Stata Journal.* 2013;13(4):795-809.

[5] Fuchs C. Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association.* 1982;77(378):270-278.

[6] For an alternative framework, the analyst could design and implement a sensitivity analysis that would compare results across different models for the missing data pattern. For example, values generated by imputation could be transformed, such as by multiplication or a constant shift: if the results differ across the transformations, this may

II. Analytic Solutions and Techniques

What Technique Should Be Applied?

Complete-Case Analysis

In situations of MCAR, cases for which all data elements are observed or measured—called *complete cases*—are essentially a random sample of the study sample. For example, consider a hypothetical situation in which all selected respondents for a mail-in survey completed and mailed their surveys, forming a complete sample, but 5 percent of those surveys were randomly lost in the mail. In this case, akin to analyzing a sample mean by a t-test, inferences about the larger target population can be made using the complete cases. Complete-case analysis is appropriate when missing data are MCAR because results are generalizable to the larger population. However, MCAR is a very strong assumption and usually unrealistic within real world data. It is important to note that estimates will be less precise than if the analysis was based on the larger sample (when all data are available), because a smaller number of complete cases are used to produce the estimates.[7]

Imputation

To leverage the most study data from cases originally identified, some analysts choose to "fill in" or impute the missing values with statistical estimates from a predictive model.

1. In single imputation, the analyst imputes missing values with a single value, such as the mean of the cases from the observed data or a single prediction from a regression model that uses data from the observed cases to predict missing values.
2. In multiple imputation, the analyst draws multiple predictions *m* times from a distribution rather than a single instance. The analyst will have *m* completed datasets—each dataset is a replicate of the original dataset with different imputed values. Each dataset is analyzed for a total of *m* analyses, and their results are pooled. The parameter of interest (e.g., a regression coefficient) is estimated by the mean and variance of the pooled results.[8,9]

Caveats of Each Method

It is important to be mindful of the drawbacks of these techniques. Complete-case analysis and single imputation depend on the MCAR pattern. They assume, respectively, that the missing data pattern is independent from the observed values or that missing values can be predicted without statistical uncertainty. These untenable assumptions introduce possible error in the analysis. For instance, single imputation yields just one prediction for each missing value; therefore, this method decreases the variance of the distribution because it ignores the uncertainty of the imputed value, which is a statistical estimate. Figure 1 illustrates this phenomenon with a normally distributed variable with five and twenty percent missing values and single mean imputation applied; note how the distribution of the dataset with single

---

indicate a violation of the MAR assumption; it may further imply MNAR, if the missing data pattern depends on the values of the dependent variable that are missing. See Yang Y, *Sensitivity Analysis in Multiple Imputation for Missing Data*. SAS270-2014. SAS Institute Inc.; 2014. http://support.sas.com/resources/papers/proceedings14/SAS270-2014.pdf

[7] Pigott TD. A review of methods for missing data. *Educational Research and Evaluation.* 2001;7(4):353-383.

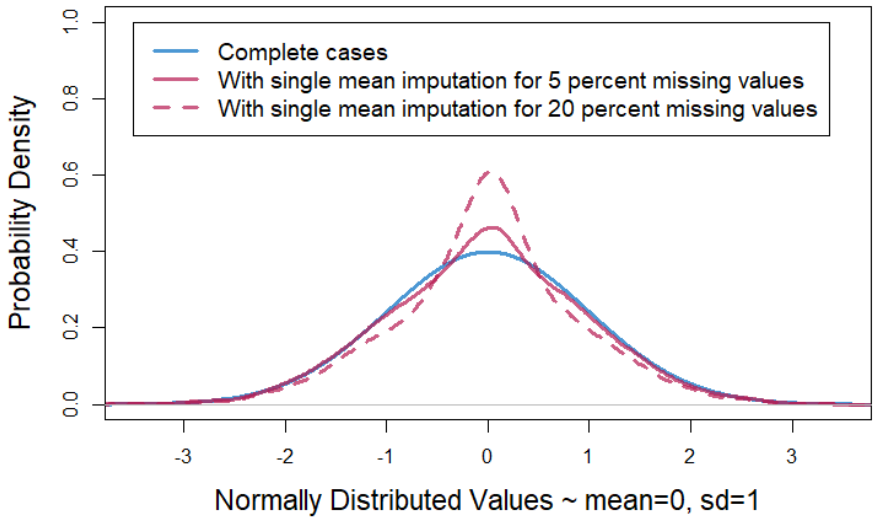
[8] Yuan YC. *Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0).* SAS Institute Inc.; 2010. https://support.sas.com/rnd/app/stat/papers/multipleimputation.pdf

[9] van Buuren S. *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press; 2012.

imputation for 20 percent missing values has a tighter spread and therefore smaller variance than the true distribution of the complete cases.

Multiple imputation addresses this limitation of single imputation by incorporating the distribution of imputed values, thereby reflecting the statistical uncertainty in imputation. Although multiple imputation is flexible to data that are MAR, MNAR, and MCAR, it also requires advanced computation. However, sophisticated statistical packages in SAS®, STATA®, and R allow for quicker, more confident implementation of multiple imputation.

Figure 1. How Single Imputation Can Reduce the Variance of the Distribution for a Variable With Missing Values



Abbreviation: sd, standard deviation.

III. Example: Solving Missing Data Problems by Multiple Imputation

To illustrate multiple imputation, we present a simple study based on the 2018 Behavioral Risk Factor Surveillance System (BRFSS) survey.[10] This example shows how to leverage data relationships to build a multiple imputation model. For the Medicaid program, at the national population level, an outcome of interest is the **percentage of adult enrollees who had a flu vaccination during the past 12 months**. For consideration in this study, factors that may influence flu vaccinations include socioeconomic characteristics (e.g., age, race/ethnicity, and income), utilization of preventive care (e.g., access to a personal doctor and routine checkups), and measures of health status (e.g., number of poor physical or mental health days in the past month). Table 3 summarizes the distribution of survey variables, including the number of observations with missing values for each variable.

[10] "The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world." Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/index.html

Table 3. Study Data—Summary Statistics of Adult Medicaid Enrollees in the 2018 Behavioral Risk Factor Surveillance System (BRFSS) Survey (n=3,464)

Category	Mean (SD)	Sample Size, n	No. of Missing Values, #
Age, years [a]	42.3 (13.6)	3,464	0
Income, $ [b]	20,198 (15,034)	2,887	577
Number of days in poor physical or mental health in the past month	9.1 (11.2)	2,461	1,003
Category	**Percentage, % [c]**	**Sample Size, n**	**No. of Missing Values, #**
Race and ethnicity [c]			
White, non-Hispanic	55.0	1,906	0
Black, non-Hispanic	17.8	615	
Asian, non-Hispanic	0.8	28	
American Indian/Alaska Native, non-Hispanic	6.3	217	
Hispanic	16.5	570	
Other race, non-Hispanic	3.7	128	
How long since last routine checkup by doctor			
Within past year	78.2	2,710	44
Within past 2 years	10.2	354	
Within past 5 years	5.5	189	
5 or more years ago	4.1	142	
Never	0.7	25	
At least one personal doctor			
Yes	22.9	2,654	17
No	76.6	793	
Flu vaccination (shot or nasal spray) in past 12 months			
Yes	24.1	834	180
No	70.7	2,450	

Abbreviation: SD, standard deviation.

[a] Continuous values (in years) was simulated from BRFSS categories.

[b] Continuous values (in dollars) was simulated from BRFSS categories.

[c] Percentages account for missing values, so they do not necessarily sum to 100 percent.

[d] Race/ethnicity values were imputed and provided by BRFSS.

Source: 2018 Behavioral Risk Factor Surveillance System Survey. Centers for Disease Control and Prevention.

Step 1: Explore Data Distribution

Summary statistics help us understand the data distribution and missing data patterns. In Table 3, we tabulate the percentage and frequency of each level for each categorical variable (race/ethnicity, routine checkups, personal doctor, and flu vaccination); we note the mean and standard deviation for each continuous variable (age, income, and poor health days). We show the frequencies of missing values for both sets of variables.

Step 2: Examine Missing Data Pattern

The pattern of missing data can be represented by a table. Table 4 indicates which variables have missing values and the pattern of missing data across variables; in other words, the table shows the structure of missing data. In Table 4, we see that 1,992 cases have complete data. Multiple imputation uses all the observed values to predict missing values in the imputation model; specifically, multiple imputation leverages the observed relationships between these variables in the complete cases to calculate imputed values for incomplete cases.

Table 4. Missing Data Pattern—M=Missing and "+"=Observed

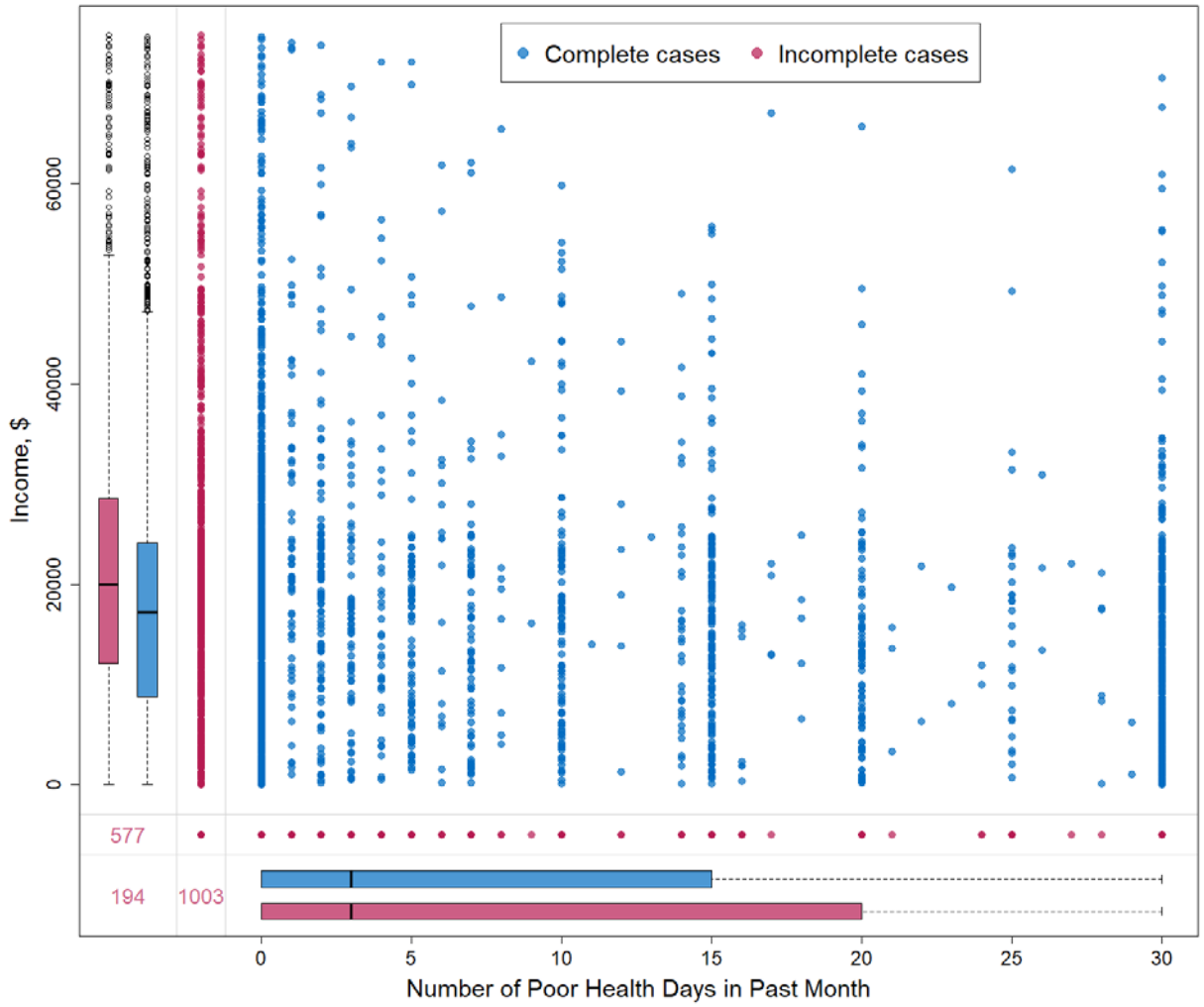
No. of Cases	Age	Income	Poor Health Days	Race/ Ethnicity	Routine Checkup	Personal Doctor	Flu Shot	No. of Missing Values Within Case
1,992	+	+	+	+	+	+	+	0
757	+	+	M	+	+	+	+	1
320	+	M	+	+	+	+	+	1
154	+	M	M	+	+	+	+	2
64	+	+	+	+	+	+	M	1
33	+	+	M	+	+	+	M	2
53	+	M	+	+	+	+	M	2
30	+	M	M	+	+	+	M	3
13	+	+	+	+	M	+	+	1
13	+	+	M	+	M	+	+	2
8	+	M	+	+	M	+	+	2
10	+	M	M	+	M	+	+	3
9	+	+	+	+	+	M	+	1
6	+	+	M	+	+	M	+	2
2	+	M	+	+	+	M	+	2

Source: 2018 Behavioral Risk Factor Surveillance System Survey. Centers for Disease Control and Prevention.

Building on the structure of missing data, we can analyze associations between the variables in complete and incomplete cases. The association can be shown by a margin plot (Figure 2), which is a version of a scatterplot. Complete cases represent those with observed values for both income and poor health days. Incomplete pairs are records where either value is missing. The scatterplot of complete cases shows the relationship between observed values, while observed values of the incomplete pairs are illustrated by vertical and horizontal dot plots.

We can assess the type of missing data pattern through the margin plot in Figure 2. Under MCAR, the distribution of each variable (shown by the boxplots) should be similar between complete and incomplete pairs. For example, there are 1,003 cases missing a value for the number of poor health days; of these,

Figure 2. Margin Plot for Pairwise Data



In the largest panel (upper right), there are n=2,078 cases with observed values for both income and poor health days. There are n=577 cases missing income values and n=1,003 cases missing values for poor health days. The missing cases present these patterns: in the smallest panel (lower left), n=194 cases are missing both values; in the thinnest vertical panel (left side), n=809 cases have income values but not poor health days values; and in the thinnest horizontal panel (bottom side), n=383 cases have values for poor health days but not income. Boxplots show univariate distributions of each variable, with outliers (leftmost panel for income) as clear dots.

No. of Cases	Income	Poor Health Days
2,078	Observed	Observed
809	Observed	Missing
383	Missing	Observed
194	Missing	Missing

Source: 2018 Behavioral Risk Factor Surveillance System.
Centers for Disease Control and Prevention.

809 records include observed values for income, which are represented by the dotplot and the corresponding leftmost boxplot. Using the median (solid line in the middle of the boxes), we see that these 809 cases have income values roughly centered around 20,000 dollars ($), whereas the complete cases have values centered around 17,000 dollars ($). The horizontal boxplots in the bottommost panel show this and other slight discrepancies in the distribution of observed number of poor health days

between the incomplete pairs and complete pairs; specifically, the number of poor health days for those cases with missing income data have a higher third quartile (75th percentile), as shown by the rightmost end of the boxplot. These discrepancies suggest that the missing values may not be MCAR. Thus, complete-case analysis is not appropriate for our study, and multiple imputation is a more fitting solution.

Step 3: Impute Missing Values

Through multiple imputation, we aim to "fill in" missing values for incomplete cases. These imputed values are estimated from a predictive model that captures the underlying association of the observed variables. A common approach is predictive mean matching (PMM) which is outlined below.

1. For each variable to be imputed, regress its observed values on the other variables. For example, to impute missing values for number of poor health days, estimate a regression model of number of days on age, race/ethnicity, income, time since last routine checkup, and having a personal doctor.
2. With the estimated regression model, predict new values for the given variable (which has missing values) for **all** observations. For example, using the regression model of poor health days on the other variables, predict this variable for all n=3,464 records in the dataset.
3. For each incomplete case with a missing value, note its **predicted** value and then identify nearby complete cases with **predicted** values that are close to it. Heuristically, this is the analytic basis of PMM, whereby incomplete cases are **matched** to complete cases.
4. Randomly choose one complete case among those matched to the incomplete case and apply its **observed** value to the missing value of the incomplete case. The random selection induces variation in the imputation procedure, which propagates through the final study analysis.
5. Repeat the previous step *m* times and store the imputed values for each incomplete case. These values will be used to generate study results over the *m* imputations, which then are pooled for the final analysis to include the variation of those multiple imputations.

In software, standard routines for multiple imputation detect which variables have missing values (e.g., in our study, the software will detect that five of the seven study variables have missing values; see the Appendix for this software example and additional resources). The routine then estimates a predictive model for each variable and performs PMM. In other words, the software routine will create *m* datasets that contain imputed values for the incomplete cases.

Step 4: Diagnose Imputed Values

Imputed values should "look like" (i.e., have similar distributions to) the original data. In other words, imputed values should be plausible. In Table 5, imputed values for number of poor health days appear to be consistent with the observed data distribution; for example, there are no negative values, and the center and spread of the distribution is similar to that of the observed data. Therefore, our imputed values resemble the original data.

Table 5. Distribution of Imputed Values for Number of Poor Health Days (*m*=5 imputations)

Summary Statistic	Observed Values	Imputed Values				
		First Set	Second Set	Third Set	Fourth Set	Fifth Set
Maximum	30	30	30	30	30	30

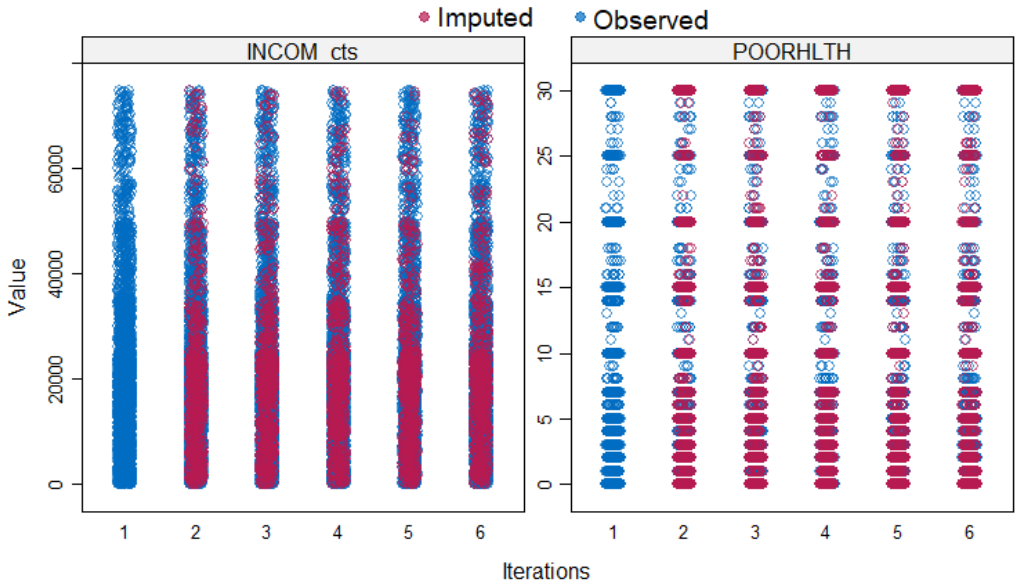
Summary Statistic	Observed Values	Imputed Values				
		First Set	Second Set	Third Set	Fourth Set	Fifth Set
75th percentile	15	15	15	15	15	15
Mean	9.12	8.81	8.87	8.93	8.82	8.76
Median	3	3	3	3	3	3
25th percentile	0	0	0	0	0	0
Minimum	0	0	0	0	0	0
No. in sample	2,461	1,003	1,003	1,003	1,003	1,003

Note: The observed number of poor health days takes on integer values in the Behavioral Risk Factor Surveillance System (BRFSS) survey; as such, imputation by predictive mean matching (PMM) will retain the integer scale for each case's imputed value, which is reflected by the percentiles of imputed values. The consistency of the percentile values across the imputations, as compared to the observed percentiles, indicates that multiple imputation by PMM will generate plausible values. The mean imputed values, however, will fluctuate, since it is an aggregate summary of many integer values.

Source: 2018 Behavioral Risk Factor Surveillance System Survey. Centers for Disease Control and Prevention.

A strip plot (Figure 3) indicates that imputed values have variation similar to that of the observed values, and the imputed values appear stable (i.e., consistent) across imputations. A rule of thumb is to implement multiple imputation at least five times (yielding m=5 "complete" datasets); with modern software, increasing this number should be relatively feasible. Increasing the number of imputations can assure that the imputed values are in fact stable and do not render the study conclusions sensitive to the statistical uncertainty that was induced by estimating the predictive model and choosing the values for imputation.

Figure 3. Strip Plot to Compare Imputed and Observed Values



Source: 2018 Behavioral Risk Factor Surveillance System Survey. Centers for Disease Control and Prevention.

Step 5: Pool Analyses Over Multiple Imputations

Finally, with the *m* "complete" datasets in hand, we analyze each one using a statistical model. For example, in our study to analyze the relationship between flu vaccination rates and sociodemographic factors, access to care, and health status, we applied logistic regression. Table 6 shows the results of our analysis, setting *m* equal to 5 and 50 for the rounds of multiple imputation; note that the conclusions of our study did not change (i.e., statistical tests on the regression terms from the two models do not contradict each other). Additionally, the increased number of imputations appears to have induced more variance in the regression estimates, appropriately reflecting uncertainty from statistical modeling.

By pooling analyses over the datasets from multiple imputation, we have successfully incorporated statistical uncertainty of the imputation process. As cursory checks on the inferences from multiple imputation, we can assess the regression estimates to make sure that they agree over multiple imputations; however, the initial review of the stability and consistency of the imputed values in Step 4 should provide a strong indication of whether the regression analysis should be sensitive to potential instability generated by the multiple imputation.

IV. Further Considerations

Application of any statistical method requires consideration of the underlying assumptions to ensure its appropriate use. The analysis of missing data must use context (i.e., subject matter expertise), observed data distributions, and missing data patterns to ensure that the chosen methodology can support accurate analyses and conclusions.

- The missing data pattern provides valuable information to diagnose the strength and weaknesses of the observed study data. In particular, the pattern can inform decisions about whether data are MCAR or MAR so that the analyst can appropriately adjust for missing data by using the observed values.
- In a sense, multiple imputation is a statistical method for prediction; as such, it is tantamount to diagnose the imputed values (including their variation and underlying uncertainty) to support their role in the data analysis. The analyst must assess potential limitations engendered by imputation and whether the imputed values introduce inadvertent bias, as compared to using the observed study alone. Context and subject-matter expertise can help guide analytic decisions for the study, such as whether to limit analyses to specific subgroups that have little or no missing data. Additionally, subject-matter expertise can help assess the plausibility of imputed values.
- Generally, it is important to consider that imputation is a statistical estimation procedure and thus generates uncertainty in its estimates (i.e., imputed values). Multiple imputation addresses this issue by incorporating the uncertainty throughout the estimation process without requiring significantly more computational and analytic effort.

In supporting Medicaid program design and operations, multiple imputation leverages the most information in data analytics by providing an appropriate statistical framework to use records with incomplete data. Through this framework, the analyst can weigh the analytic advantages and limitations of using imputation methods, as compared to a rote complete-case analysis. In doing so, state agencies can ensure that information is used efficiently and that potential limitations are acknowledged for more robust decision-making.

Table 6. Logistic Regression Analysis Following Multiple Imputation

Regression Term	Estimate	Standard Error
m=5 imputations		
Intercept	2.362	0.201
Age, years	-0.021	0.003
Income, $	0.000	0.000
Poor health days	0.004	0.004
Race/ethnicity		
White, non-Hispanic [a]	--	--
Black, non-Hispanic	0.022	0.114
Asian, non-Hispanic	0.374	0.465
American Indian/Alaska Native	0.522	0.167
Hispanic	0.399	0.115
Other race, non-Hispanic	-0.212	0.242
Last routine checkup		
Within past year [b]	--	--
Within past 2 years	-0.513	0.156
Within past 5 years	-0.747	0.236
5 or more years ago	-1.807	0.506
Never	-9.254	219.5
Personal doctor		
Yes [c]	--	--
No	0.528	0.127
m=50 imputations		
Intercept	2.385	0.200
Age, years	-0.020	0.003
Income, $	0.000	0.000
Poor health days	0.006	0.004
Race/ethnicity		
White, non-Hispanic [a]	--	--
Black, non-Hispanic	0.035	0.112
Asian, non-Hispanic	0.474	0.462
American Indian/Alaska Native	0.514	0.168
Hispanic	0.401	0.116
Other race, non-Hispanic	-0.168	0.244
Last routine checkup		
Within past year [b]	--	--
Within past 2 years	-0.508	0.159
Within past 5 years	-0.784	0.241
5 or more years ago	-1.865	0.425
Never	-12.596	259.4
Personal doctor		
Yes [c]	--	--
No	0.539	0.122

[a] Reference category.

[b] Reference category.

[c] Reference category.

Source: 2018 Behavioral Risk Factor Surveillance System Survey. Centers for Disease Control and Prevention.

APPENDIX: EXAMPLE PROGRAMMING CODE

R

The example study in this webinar supplement was implemented by the "mice" package in R.

Annotated code and the analytic dataset are included on the Medicaid IAP Data Analytics page under National Dissemination

SAS

Annotated examples of programming used to handle missing data in SAS can be found on the University of California, Los Angeles, Institute for Digital Research and Education website

- Multiple imputation

STATA

Annotated examples of programming used to handle missing data in STATA can be found on the University of California, Los Angeles, Institute for Digital Research and Education website:

- Multiple imputation