Medicaid Innovation Accelerator Program Webinar Transcript

Data Analytics National Webinar - Solving Missing Data Problems
October 23, 2018

Welcome/Agenda

JESSIE PARKER: GTL and Analyst on Medicaid IAP Data Analytic Team, Data and Systems Group, CMCS. We'll cover:

- Ways to identify different types of missing data
- How to adjust missing data with imputation
- Example from Ohio Medicaid of putting imputation to use in order to avoid biased results due to missingness problems
- Joint community session – questions to chat box

Speakers Today

- Frank Yoon, Senior Statistician, IBM Watson Health. He will review the different types of missing data we learned in our previous webinar on missing data before moving on to analytic solutions. We will discuss complete case analysis, single imputation and multiple imputation. We will discuss pros and cons of each approach as well as which methodology is appropriate for your research question. Frank will also discuss practical considerations for your analysis involving missing data.
- Jonathan Barley, Chief, Bureau of Health Research and Quality, Ohio Medicaid
- Tim Sahr, Director of Research, Government Resource Center, Ohio Colleges of Medicine

Jonathan and Tim will be discussing a 2016 analysis they completed on Ohio's expansion population. They designed included survey data with missingness, so they used multiple imputation to avoid bias introduced by the missing data. This is a real world example of when the methodologies may be put to use.

Introduction

This webinar is produced through the Medicaid IAP. Our primary goal for today's webinar is for you to learn not just how to identify types and patterns of missing data but also how to approach missing data in your own analyses. Part 1 of this webinar series was meant to instill an appreciation for the seriousness of missing data problems, and now in Part 2 we hope to begin to empower you with the tools to address these problems in a methodologically appropriate way. Frank's presentation will cover these questions from a more academic perspective whereas Ohio Medicaid will demonstrate a real world use case of how these tools can be put into action.

Solving Missing Data Problems

Dr. Frank Yoon will discuss how to solve missing data problems.

FRANK YOON: As indicated, there was a previous webinar discussing an introduction to missing data problems where we talked about various types of missing data, thinking about how to assess patterns of missing data and why those patterns might be useful in determining the right approach for your given analysis. The three main types of missing data we'll talk about today will be:

- Completely at random (MCAR)
- At random (MAR)

- Not at random (MNAR)

As a high level overview, in this presentation we will be talking about:

- How to assess patterns in missing data
- What the pattern tells us in terms of the right approach for our analysis
- Various technical approaches and particular implementation in statistical software to adjust for missing values

For example, here is a simple illustration of thinking about diagnosing missing data. In this task, let's suppose we're trying to predict annual healthcare costs or utilization. That adjusts also for items in administrative claims data, as well as clinical intake forms. In this case, let's suppose that the key intake items to analyze the outcomes would be sociodemographic factors, for example, race, ethnicity, clinical factors. Let's say we're talking about a situation where behavioral health indicators might influence healthcare cost utilization in this population. Finally, let's suppose we have an instrument that collects primary data from beneficiaries. In this case, suppose we have a screening form for depression risk through the PHQ-9.

As a first suggestion, we would assess the data set itself, look for the indicators of missingness, and try to feel for why certain fields and values may be missing on the basis of the outcome as well as the covariants. Here we are just illustrating a simple flat file layout where we have outcomes and costs at the left-hand column and all the covariants thereafter, along with the missing indicators with the red question marks.

Generally speaking, we would expect to be looking at the raw data set to assess patterns but here I just wanted to give an illustration of where we were going to go with the analytics in the following slides.

Back to the three types of missingness, and MCAR, whereby the missingness of the data doesn't have a relationship to the outcomes or the covariants, and MAR, whereby we can assess patterns of missing data on the basis of information we have, for example knowing that certain types of individuals respond at different rates in our populations. Finally, MNAR, which means the data are missing on the basis of important information we don't get to observe, for example in outcomes.

The first poll here for the audience is which situation here would present one where we believe that the missing data mechanism is MAR? That is, we have a variable we can use to adjust for missing values simply by understanding the reasons for missing values.

a) Analyst spills coffee on client intake forms

b) Men are less likely to complete a PHQ-9

c) High utilizers of health services do not report race and ethnicity

Looks like most people chose that MAR was the situation where individuals are not reporting their race or ethnicity. In fact, we're looking at a situation where the sex of the beneficiary would indicate the missingness probabilities. For example, we're looking specifically at men who are not choosing to respond to the intake form for behavioral health. Knowing that the data are MAR, we can use the fact that men are less likely to complete this form to adjust for those missing values simply by, for example, taking the average response of men for the available observations and using that to impute the missing values for the PHQ-9 form.

Jumping to our analytic approaches here, standard approaches to missing data would include complete case analysis, whereby you only take the observations that have complete and available values for all the covariants and outcomes of interest in your analysis here. Think about in which particular situation would a complete case analysis actually be useful and valid. The flipside question would be in which situation would we definitely want to avoid using this complete case analysis.

If we're not going with the complete case analysis here, we want to try and leverage all the observations in our data set, in particular those with missing values for covariants and absence of interest. Generally speaking, we like to talk about imputation methods, that is, trying to leverage the associations with the variables in our data set to impute or fill in the missing values for the covariants of interest.

The two main types of approaches we'll talk about here today will be an imputation approach. For example, in a situation where we know men are less likely to complete the PHQ-9 intake form, we can take the average values of that form for men versus women. For the missing values in either sex group we can use the average response to impute any values that are missing.

Extending that out, we can also look at regression methods that also incorporate other covariants, say a multiple regression approach, again to impute the missing values through a prediction-based approach.

That said, imputation has some limitations. To address those limitations, we also want to talk about multiple imputation methods, whereby we want to account for uncertainty in the missing values or the imputed values. For example, taking a prediction from a regression equation does not incorporate the uncertainty surrounding the values.

To further these ideas and to start illustrating methods for missing data analysis through imputation methods, we're borrowing an example from the National Health and Nutrition Examination Survey or NHNES. In this case, we're looking at four covariants—age, BMI levels, hypertension status, and cholesterol levels—for the respondents in this population. For a quick summary, we have a data set of 25 observations, again with four covariants. And a nice check, in R, for example, to assess initially the patterns of missing data is to do just a quick summary of the contents of the data set. Important outputs of this quick diagnostic would the bottom rows shown for each of the covariants that generally tell you the frequency and missing values in that data set.

A refresher here. For types of missing data, we're looking at MCAR. Takeaways are, generally speaking, we have an unbiased example that is representative of the population of interest. That is, we can ignore any observations that have any missing values, simply by assuming that those missingness patterns do not depend on any of the covariants or outcomes in our data set.

MAR takes that a little step further, thinking about the reasons for missingness, specifically knowing that the covariant information can explain reasons why, for example, a respondent did not respond to a particular item on the survey, let's say.

Finally, MNAR. In this situation, we have a pattern of missingness that depends on very important outcome information. For example, high utilizers of healthcare may not be responding to various instruments or items in our data set, in which case we obviously cannot ignore them and in some way need to account for the patterns of their missingness.

Jumping back to the NHNES example. So in this diagnostic, we have five different types of missingness. The first case is no missingness. On the left-hand column we have 13 observations that have complete observations across age, hypertension, BMI, and cholesterol, whereas in the bottom row we have seven observations that are complete with age but missing the other covariants. On the right-hand side we're looking at the number of variables missing for each case. Again for the first case here, zero variables are missing in that first missingness pattern, and in the bottom row we have three of the four covariants missing information here, with standard observations in our data set with that pattern.

So what I'm doing in this example is using the package VIM and using a margin plot to assess the missingness patterns, as well as the patterns of actually observed values in our data set here. We're looking here in the upper right quadrant essentially with the blue dots. We're looking at the association of the observed values for the complete hairs or cases in our data set between BMI and cholesterol levels.

The red dots indicate the distribution of the incomplete pairs. For example, on the vertical strip, we're looking at the distribution of BMI for those observations that are also missing cholesterol levels. Likewise, on the horizontal strip where we see those two pink dots, we're looking at the distribution of cholesterol levels for those observations that are missing BMI.

A poll question: Keep in mind the pattern is missing data is as assessed through this plot. In the case of MCAR, how should these boxplots look, in particular the distribution of the pink boxplots, which tell you the distribution of the incomplete cases for each of the covariants there?

a) Alike

b) Different

c) Not sure; need more information

In fact, under MCAR here, you can think about this through say a random sampling sort of framework, whereby the observed sample should be representative or similar to the target population. So in this case, borrowing that intuition here, the distribution of the incomplete and complete cases generally should look alike if we're thinking about these data MCAR.

A quick recap, for a complete case analysis we would assume quite strongly that the data are MCAR and that the observed values, that is the complete cases are representative of our target population. That is, we can ignore any observations with any missing values.

Simple imputation is an approach whereby we can take different types of units in our study sample, look at the probabilities or frequencies of missing values according to the types of those individuals, and then use imputation methods to fill in the blank for missing values based on those particular characteristics. Again, thinking about if men are less likely to complete behavioral health intake forms, we can take average responses for the men to assume missing values for that particular group.

For a prompt in thinking about why multiple imputation might be a preferred approach, let's start thinking about what could go wrong with single imputation methods. So for the two types of elementary solutions, again in a complete case analysis we assume that the complete cases reflect incomplete cases. That is, we can ignore the reasons why the incomplete cases might have missing values.

For single imputation methods, we're thinking about we can parse out our data set and use the covariant information to impute mean values for different types of beneficiaries in our analysis. In this case, we will be assuming MAR. Given that, I'm going to prompt the audience here to think about in what other situations would MAR be appropriate. Back to looking at BMI versus cholesterol, again we're looking at the complete cases in a blue scatter plot as well as incomplete hairs represented by the red dot and boxplots.

Let's say in this case we want to assess or fill in the missing BMI values for these cases here indicated by the red arrow. The idea here generally is that assuming MAR, assuming that we can leverage the information from cholesterol values to fill in the BMI values, single imputation might take observations with similar cholesterol levels. Then, using those observations, get an average for those individuals with that particular cholesterol level and use that value to fill in the missing BMI values for this observation. You can extend this approach through the Russian-based framework whereby you could leverage more than one covariant to missing values for particular covariants.

For another prompt here, the question is what could go wrong with that approach, whereby we take a single predicted value to fill in the missing values in our data set here between BMI and cholesterol? Look at the plot and we'll jump to the poll on what could go wrong:

a) Nothing, it's perfect

b) There's just one imputed data point

c) We have to assume MAR and not MNAR

I should have told you actually there was more than one correct response in that multiple choice question. However, there is one more important response, and it seems the audience identified the correct one. In this case here we're just dealing with one single imputed data point, in which case we're not accounting for the uncertainty estimating the average BMI for those observations at that given cholesterol level, so that is correct. However, our results also can be much more sensitive to the distinction between MAR versus MNAR in our data set. That is, we have to assume that we can use the cholesterol levels to accurately predict the missing values of BMI in our data set here.

Just a simple heuristic here, thinking about our intuition about missing data. It's always going to be important to assess missing patterns, and based on those patterns, to essentially justify your assumptions across the three types of missing data cases: MCAR, MAR, and MNAR.

Addressing inherent limitations of single imputation here, we want to think about applying multiple imputation approaches, whereby we're essentially trying to leverage the underlying statistical or data-generating model to generate values or distributions of imputations such that we can account for the uncertainty in the adjustments that we're making.

Multiple imputation essentially has three sets. The first step, we fill in the missing values multiple times. Think of this like bootstrapping methods whereby we're trying to leverage the data distribution to account for uncertainty. In essence, we're predicting missing values.

The second step is to analyze the multiple data sets that contain multiple imputed values.

Finally, we want to pool our analyses across those imputed data sets, again to incorporate the uncertainty surrounding the prediction or the estimation of missing values on the basis of those covariants.

So in R there is a nice package called "multivariate imputations by chained equations" (MICE). This is a well-established and well-developed package that allows you to do multiple imputation pretty successfully and pretty quickly. I won't get in much detail on the functionalities of the package, but just wanted to give a quick overview of what it actually is doing. It will help you inspect the missing data pattern. It will conduct multiple imputations, provide diagnostics, and run the analysis, for example through standard methods such as regression.

I'm going to walk quickly through some of the code here just to give you a sense of how easy it is to do multiple imputation here in R. So here we're taking our NHNES data set. We're specifying five iterations of the multiple imputation, then we're going to diagnose the multiple imputation quality as well as analyze our data. A default method for multiple imputation in the MICE package is predictive mean matching, which I'll talk about later.

So when we have our imputed values, we initially want to assess the quality of those imputations, specifically by making sure things like for imputing age or BMI, that the imputed values through the estimated model are essentially in the same range as the observed values. We also want to think about things like logical sort of inconsistencies. For example, we want to make sure that our imputations for a non-negative result such as BMI doesn't contain negative values.

The quick way to do this would be to produce what's called a strip plot. Again this is coming from the MICE package whereby we want to look at our observed values in the blue and then compare them to the actual imputed values, in this case the five multiple imputations here. In the strip plot we essentially just want to make sure that the imputed values in the pink look similar to the observed or actionable values in the blue.

Another poll question: With respect to the diagnostic here, given our understanding of multiple imputation as well as the underlying assumptions, what do you think we should be seeing here if the multiple imputations are working correctly?

a) Imputed look similar to observed values

b) There is variation in the imputations

c) Imputed values appear stable over iterations

d) All of the above

Our results: Wonderful. Looks like we're trying to get a sense of what our imputations should look like when the quality of our imputations are good and providing robust and accurate results for subsequent analysis here. So in fact, yes, looking at the imputed versus observed values, we want to make sure they look similar, that the distributions are in essence overlapping; that we have variation accounting for the statistical uncertainty of the estimation methods used to predict missing values; and the imputed values over those five iterations appears stable.

Once we have our imputations here we can apply the MICE package to conduct analysis, for example, using regression methods here. In this case, my analysis goal is to look at the association of cholesterol as regressed on age and BMI in the respondent on the intake survey. We can use the MICE package to incorporate multiple imputations here. The first upper half, we're looking at a situation before where we had the five multiple imputations. That is, we run the imputation methods five times, collect those five sets of results and pool them in a linear model here through the MICE package.

If you're concerned that five imputations might not be enough, given the nice accessibility of the MICE package in R, we can increase the number of imputations here. In this case I'm increasing it to 50 imputations, whereby we're trying to get a sense for the stability of the results on the basis of the number of imputations that we're conducting here, what we could do to do a quick comparison of the point estimates for the regressions as well as the standard errors. They don't look too different but general rule of thumb, certainly apply more imputations here if you're in doubt whether you have enough imputations. This simple problem, of course, runs very quickly, but of course for your problems you would do well to consider the computational intensity required to conduct many more iterations of the multiple imputation.

For our final polling question, thinking about the analysis on the basis of multiple imputed data whereby we have again multiple data sets of filled-in values which we're going to pool to get things like regression coefficients and other sort of estimates in our analysis here, thinking about what we would need to do differently with regression modeling with multiple imputed data. So thinking about a situation where you have a regression, a single regression versus one with multiple imputations. Let's think about the multiple steps we would need to take to do the right analysis here. So again, single regression versus one with multiple imputations, what would we need to do here?

a) We need to double check our coefficient estimates to make sure they agree over multiple imputations

b) We need to throw out results that don't look right

c) We need to combine results over multiple imputations in order to calculate standard errors of estimated coefficients in the regression model

So importantly with multiple imputation methods we want to pool the results, combine the results across the iterations, and again with software such as the MICE package in R we can do that in essentially one fell swoop through some simple coding statements here.

I'm going to spend time on the specifics of the underlying methods of multiple imputation in the MICE package here. As I mentioned, it conducts predictive mean matching. Essentially it's going to generate predictions of missing values and match them—that's where the matching term comes from—with the observed values and essentially conduct random draws of those matched values to fill in the missing values for that particular observation that presented the missing information here. I'm not going to get into more detail but the documentation and the MICE package have a nice summary of these methods as well as others which you can find in online references.

To wrap up the technical specifics here, for multiple imputation methods we want to think about modeling choices, think about various methodologies to do that predicting in predictive mean matching. Initially in the previous example, we were using simple regression models to predict missing values. You can consider, however, other more sophisticated approaches, such as classification and regression trees, Bayesian methods. You can specify your model to have interaction terms and whatnot. Again, all of this sort of depends on your understanding and your assumptions about the missing data and how you can leverage the available covariant information to predict those values in your analysis here.

If there's anything to take away from this presentation, you want to think about how you can stand by your analysis, specifically by justifying your assumptions as to whether or not the missing data are MCAR, MAR or MNAR, and make appropriate adjustments and stand by your assumptions about why those adjustments are appropriate.

Do you want to give a quick indication of other software platforms whereby you can predict with multiple imputation? For example, in SAS there is a well-established macro called IVEware, as well as the data which has the multiple imputations in the MI library. The caveat is to check all settings to make sure the default setting is the right one for your approach and if not to explore the other options as you implement your multiple imputations in software.

I'll wrap up. The next speaker will give an illustration about an actual interpretation of multiple imputations in analysis.

Ohio Medicaid's Experience

JONATHAN BARLEY, Ohio Department of Medicaid: I’m here with Tim Sahr. We partnered to do a study on our expansion population. A little background to put things in context. Ohio Medicaid expanded through the ACA Medicaid in 2014, very controversial. The Governor of Ohio pushed it through the Legislature. Many in the Legislature were very reluctant to go along with it, so they mandated that the department conduct a study of the expansion and the impact of the expansion on those who were covered.

So we set up a population comparison study. This was conducted a couple years after, so in 2016. The G-VIII it refers to the name for the expansion population. We set up a straight up 2-population comparison to look at the pre-expansion population in Medicaid compared to the post expansion to those actually covered by the expansion, looking at their health and socioeconomic statuses.

Our general design—we worked closely with a local partner at Government Resource Center located at Ohio State University, since it was such a politically hot topic, we really needed sound methodology to make sure the results could not be argued with. We did the following, looking at this slide. We looked at multiple modes or sources of data to find out how well different measures from these different modes could come up with common findings. You can look through the list. All the different modes of data are listed there. As you can see, looking through them all, it’s not just a survey and administrative data, which is probably a pretty classic way to do the survey. We added the biometrics measures. We looked at medical records, extractions. We looked also at the various interviews with enrollees and stakeholders.

We’re looking for common findings across all these data collection methods for this survey. So we needed a bullet-proof result, whatever the result was, that couldn’t be shot down. It was very important to get it right.

A little background again looking over the different modes of data collection and the time period. Again, we had two populations we were looking at, the pre-Medicaid enrollment and post-Medicaid enrollment. You can see how many records or folks were surveyed for each one of the modes of data collection and what time period that covered.

I’m going to go to Tim Sahr of the Government Resources Center at Ohio State University, who was the lead in charge of the survey.

TIM SAHR: Actually, methodology-wise for things like this, we use multiple methodologists. RTI International was the vendor that collected data and worked with our methodologists. We worked with methodologists at OSU. We worked with methodologists at the University of Cincinnati. We worked actually with the State and Local Surveys Group out of National Cancer Institute and a variety of other people to refine the method.

A couple things before we go onto the slides. The collection in 2016, as well as the one we just finished in 2018, had very little missing data. We were very surprised by that, be it the survey or Medicaid data. For the most part everybody in the G-VIII study—the name G-VIII came out of the clause in the Social Security Act that allows the expansion of Medicaid so we thought it would be nice so none of us would get confused—we basically had very little missing data. We basically imputed for two purposes. One, to make more complete specific or key variables we wanted to look at and also in order to assist with weighting of the survey. We didn't want to have missing data for the weighting calculations.

In one sense, then, we only imputed for this project on race, marital status, chronic condition status, Hispanic origin, education, smoking status. The variables imputed were weighting calculations, except for smoking, and the weighting calculations used the inverse probability of selection for each of 400 strata.

You may be asking how does a small thing like this have 400 strata? You have those enrolled prior to ACA and we followed them. You had those enrolled after ACA. We always, with work related to the State of Ohio agencies, particularly the Department of Medicaid, look at different types of counties, and in fact classifications for counties.

Ohio is only really like Pennsylvania and New York in one sense. That is, if we have metropolitan areas we have very robust suburban areas, rural farm areas, and then we have Appalachia and Appalachian counties. The blind mix of those counties gets to be a problem so we set up a strata and then for the weighting, again which imputation plays into, we do it per strata and then we rake and blend to get a common weight. That's kind of important actually and we'll get into that with imputation here in a minute.

The Department of Medicaid and state agencies actually make data available for academics and others to use in a very transparent way. That being the case, the ease of use of data matters and having the unified weights and things like that matter. With this, we did use it for weighting. This is the weighting adjustment. We worked off the sample plan of eligibility criteria for each strata. We looked at unequal weighting effects, and the weighting effects were actually all within tolerance, which is a good thing, and all analyses of survey data and by the way of the biometric data were weighted analyses. The biometrics and the survey data, by the way, were weighted independently of each other given the radical difference of collection mode.

Imputation was conducted on survey variables. We did not do imputation on the biometric screening data or any of the other data sets, and again, we were less than 5% missing on everything. In fact, we were 3% on missing race, which is not uncommon. We were around 1%, 1.5% missing on any other variable or less, so it wasn't much of a problem.

This was a summary slide. What we ended up really doing is using the weighting. We used stochastic imputation and the reason is the literature tells us one, it's good for categorical variables without modeling assumptions. We didn't have truthfully enough missing data to worry about modeling assumptions. On other projects we do for the state, Medicaid in particular, that's not the case. Also the literature suggests that multiple imputation and single imputation, the difference is minimal when the missing rate is very, very low. So we went with stochastic.

What we ended up having was four key weighting variables and key variables of interest, say like behavior risk, we imputed. We put the weighting to the strata. We unified the weight and hence the imputation method provided more complete cases, and at the same time provided more accurate weighting, which resulted in less UWE. That's about it.

JESSIE PARKER: Thank all our speakers. Those are great examples of how we put imputation into action and analysis. We also liked to hear how states are addressing data analytics problems in the real world. We'll end today with a Q&A session. Submit questions to the chat box and I'll turn it over to Tracy to facilitate.

Q&A

TRACY: We do have questions. The first is for Frank. This question is asking about the pros and cons of alternative methods for missing data, such as hot deck imputation or weighting, compared to multiple imputations. The audience would like to know why one would use one method over the other.

FRANK YOON: It's more a longstanding method, let's say, such as a hot deck imputation. It might be more akin to making edits in your data set to fill in the missing values, and not necessarily leveraging the associations of the covariants and also in our case the outcomes to assess and analyze patterns firstly, and to use those patterns to fill in essentially the missing values on the basis of models we use to predict them.

In that respect, I do think the methods that rely more on modeling-based techniques, specifically multiple imputation as well as weighting, might bring more value to your analysis there.

The distinction being I think multiple imputation versus weighting is more a matter of personal taste, whether you think you can rely on a survey sort of weighting approach to use the available and observed values to essentially fill in the missing observations using data sets or whether you want to use a more predictive approach through imputation, whereby you're taking specific outcomes, say a prediction model, to fill in in a sense or estimate the missing values in your data set there. There's a fine line, but I think it's a matter more of personal taste and how you think the models are performing with respect to the underlying assumptions as well as your analytic goals.

TRACY: The next question is related: How many iterations are really sufficient for multiple imputations? You presented two versions in the example. One was an M of 5 and one was an M of 50.

FRANK YOON: Generally, the rule of thumb out there is that maybe 5-10 might be sufficient, but of course in this day and age, with the computational power as well as our ability to access these methods through statistical software such as SAS and R, I would say in some respects to look at as many as you think is computationally feasible with your analysis there.

A similar question, for example, arises in bootstrapping methods. A lot of people ask should I use 500 versus 1,000 versus 5,000 iterations. It just depends on what is your analytic goal and what are your computational resources, so try a bunch.

TRACY: Next for John and Tim, for your assessment, the questions asks, "Would I measure beyond Medicaid administrative data to determine health benefit levels for the newly enrolled in the Medicaid expansion?"

TIM SAHR: If I am interpreting the question properly, the Medicaid data gives you event data. Usually that event data is at a stage of say disease progression or a condition or manifestation of behavioral risk or whatever. We use all the data actually. The only data that proved a little difficult honestly was the medical records and that had to do with coding of the medical records by physicians and professionals apart from Medicaid administrative data, survey data, biometric data, and even the qualitative interviews. There was congruence among all data sets except for the medical records data, which we do think the billing or how people bill was a little bit difference in how people record.

So we were able to determine benefits by tracking over time the Medicaid administrative data on the Medicaid record to the individual. We interviewed the individual about behavioral risk, different types of prevention issues, exercise, do you smoke, what is your socioeconomic status currently, how do you participate in employment, a whole bunch of things, even some stuff on housing security.

Then we backed that up as an almost internal verification, sort of like some of Carmine and Zeller's work, if you're familiar with that, after the Sage publication, on basically looking to see as iterater and then using in some cases software that basically coded against the main findings of Medicaid administrative data in the survey. They had great congruence. We were expecting more variation between our sources than we got. So when we met a point of congruence and met basically 80%, 75% or more congruence between sets, we thought we had what we needed.

JONATHAN BARLEY: If I could add this being a highly sensitive topic politically, the optics of it, it looks very sound even if you're not an expert in study design or what have you. It's nice to have the solid backing of all these data sets coming together and showing the same thing. It just makes it harder and harder to argument with using all these data collection methods.

TIM SAHR: To answer one other thing in the question, everything was nested out of a sample drawn from Medicaid data. It was not a population-based survey. It was a directive survey. Beyond the Medicaid data and random selection, we had a series of nesting, so that the survey people came from Medicaid data, the biometric people came from the survey data, the medical records came from the combination of the biometric and survey data. We tried to use a nested approach with Medicaid records being the starting point and in all honesty the gold standard, understanding the limits that people do things that in fact get them a Medicaid billing.

TRACY: Related to that, how did you decide to use imputation for your analysis compared to other methods?

TIM SAHR: We have methodologists, including myself, who favor key questions if it's doable using imputation to have as a full a data as possible for modeling so we don't have to worry about perilous-wise deletion, one. Two, we're very cautious about how you weigh data, particularly when it's nested and you know you have response issues as far as respondents and who's out, who's in, how random is somebody not showing, how not random. We basically decided to use imputation to make the data sets as full as possible for the analytic we need.

TRACY: Last question to Frank: If I have a categorical variable that has some missing data, is it still a good idea to use MICE? I usually have been using mean in case the missing data is small, but I've never worried about the variance change. If I have to fill in a categorical variable for missing data, which step would be preferable, MICE or mean?

FRANK YOON: That's an interesting question. I haven't worked specifically with imputation on categorical data. But I imagine that one important consideration is when using the underlying predictive to fill in the missing categories or missing values for that particular variable, I think maybe you have the thresholding that is used, for example, in the simple case of a binary variable. (There) the imputation method might be something like a linear probability model or statistic regression, which is going to produce a continuous—specifically the probability of that binary variable, I imagine that you'd want to think about how the underlying methodology is thresholding that probability to generate the actual categorical values. That's one thing I would consider.

With respect to MICE versus a mean approach, it just depends on your underlying assumptions and how you believe your methodologies or approaches might be sensitive to those specific assumptions.

Key Takeaways from Today's Webinar

JESSIE PARKER: Some takeaways from today's webinar:

- To solve the missing data problem, we first need to consider:
  - What does the pattern tell us about our data?
  - When is it appropriate to adjust for missing values?
- Analytic solutions may include complete case analysis, mean or regression imputation, or multiple imputation, depending on the pattern of missing data. The strengths of multiple imputation in particular were demonstrated today.
- Imputed values and analytic findings may be checked for content validity and interpretation using qualitative interviews and variation post analyses.

Thank You/Webinar Ends

Thank you to you our speakers and attendees. Please take our post webinar survey. Slides and a recording of this session will be posted on our data analytics website and we will email all participants with the link. For more questions on the IAP program or data analytics team, reach us at: medicaidIAP@cms.hhs.gov.